Hybrid AI Systems in Automated Content Moderation and Analysis

Dissertation

zur Erlangung des Grades eines Doktors der Naturwissenschaften
(Dr. rer. nat.)

am Fachbereich Mathematik und Informatik der
Freie Universität Berlin

vorgelegt von **Veronika Solopova**

Berlin 2024

Erstgutachter:	Prof. Dr. Christoph Benzmüller
Zweitgutachter:	Prof. Dr.-Ing. Dorothea Kolossa
Tag der Disputation:	09.07.2024

Declaration of authorship

Name: Solopova
First name: Veronika
I declare to the Freie Universität Berlin that I have completed the submitted dissertation independently and without the use of sources and aids other than those indicated. The present thesis is free of plagiarism. I have marked as such all statements that are taken literally or in content from other writings. This dissertation has not been submitted in the same or similar form in any previous doctoral procedure.

I agree to have my thesis examined by a plagiarism examination software.

Date: 06/02/2024 Signature: _______________

Contents

1 Introduction 1

1.1 Automated Content Moderation 4

1.2 Hybrid Artificial Intelligence 12

1.3 Goals and Contributions 13

1.4 Outline and Synopsis 15

2 Automated Multilingual Detection of Pro-Kremlin Propaganda in Newspapers and Telegram Posts 16

2.1 Preface 17

2.2 Contributions 18

2.3 Introduction 19

2.4 Related Work 21

2.5 Methods 23

2.6 Data 24

2.7 Results 28

2.8 Additional Analysis 29

2.9 Discussion 33
2.10 Conclusion 36

3 The Evolution of Pro-Kremlin Propaganda from a Machine Learning and Linguistics Perspective 37

3.1 Preface 38
3.2 Contributions 39
3.3 Abstract 40
3.4 Introduction and Related Work 40
3.5 Methods 42
3.6 Results 44
3.7 Error analysis 44
3.8 Comparative Analyses 46
3.9 Discussion and Conclusion 49

4 Check News in One Click: NLP-Empowered Pro-Kremlin Propaganda Detection 52

4.1 Preface 53
4.2 Contributions 55
4.3 Abstract 55
4.4 Introduction 56
4.5 Related Work 57
4.6 Methods 58
4.7 Results 66
4.8 Discussion 70
4.9 Conclusion 72

5 **Dissemination Mechanisms of Harmful Speech Online: Anatomy of Two Shitstorms** 75

5.1 Preface 76

5.2 Contributions 77

5.3 Abstract 78

5.4 Introduction 78

5.5 Empirical Basis: Two Example Storms 79

5.6 Groups: Supporters and opponents in the shitstorm 86

5.7 Chronological analysis 89

5.8 Automatic linguistic differentiation 91

5.9 Discussion and Conclusion 92

6 **Implications of the New Regulation Proposed by the European Commission on Automatic Content Moderation** 94

6.1 Preface 95

6.2 Contributions 96

6.3 Abstract 97

6.4 Introduction 97

6.5 Terminology and Legal Background 99

6.6 Risk assessment of countermeasures for fake news and hate speech detection . . . 103

6.7 Discussion 108

6.8 Conclusion 109

7 **A German Corpus of Reflective Sentences** 111

7.1 Preface 112

7.2 Contributions 113

7.3 Abstract 114

7.4 Introduction . . . 115

7.5 Related Work . . . 116

7.6 Data Collection . . . 117

7.7 Annotation Guidelines . . . 118

7.8 Annotation Process . . . 120

7.9 Analysis . . . 121

7.10 Conclusion . . . 123

7.11 Future work . . . 124

7.12 Acknowledgements . . . 124

8 PapagAI: Automated Feedback for Reflective Essays 125

8.1 Preface . . . 126

8.2 Contributions . . . 127

8.3 Abstract . . . 130

8.4 Introduction . . . 130

8.5 Related work . . . 131

8.6 Methods, Components and Performances . . . 132

8.7 System Architecture . . . 134

8.8 Comparison with GPT-3 . . . 136

8.9 Discussion and Conclusion . . . 138

9 Automated Feedback Can Foster Deeper Reflections 140

9.1 Preface . . . 141

9.2 Contributions . . . 141

9.3 Abstract . . . 142

9.4 Introduction and Related Work . . . 143

9.5 Methods and Study Designs 144
9.6 Results 149
9.7 Discussion and Conclusion 156

10 Automated Content Moderation Using Transparent Solutions and Linguistic Expertise 158
10.1 Preface 159
10.2 Contributions 160
10.3 Abstract 160
10.4 Introduction 160
10.5 Methods 161
10.6 Conclusion 163

11 Discussion 165
11.1 Contributions 166
11.2 Limitations and Future Prospects 167

12 Conclusion 173

A Appendices 176
A.1 German summary 177
A.2 Chapter 2 178
A.3 Chapter 3 179
A.4 Chapter 4 182
A.5 Chapter 5 185
A.6 Chapter 7 187
A.7 Chapter 8 191
A.8 Additional tools 194

References **224**

Listing of figures

2.1 Genre-related differences between newspapers and Telegram propaganda subsets. . 27

2.2 Interpreting SVM predictions 30

2.3 BERT versus SVM with linguistic features: advantages and disadvantages 35

3.1 The linguistic feature comparison between the 2022 and 2023 propaganda subsets 48

4.1 Check News in One Click. System's flow mock-up 61

4.2 The distribution of the learnt features according to the stance. German and Italian models SVM models. 63

4.3 Check News in One Click. User study demographics. 64

4.4 Check News in One Click. User study questionnaire results. 65

4.5 Check News in One Click. Most frequent linguistic indicators shown to users during the user study. 67

4.6 Check News in One Click. Most frequent keywords explained to users during the user study. 68

4.7 Check News in One Click. Differences in the text length between the models' training sub-corpora and the user inputs 72

5.1 Tweets from Elon Musk about the war in Ukraine. 79
5.2 Timeline of the Elon Musk shitstorm on Twitter and Telegram. 80
5.3 Use of emojis in Elon Musk's shitstorm, Twitter versus Telegram. 82
5.4 Distribution of toxic language on Telegram in Elon Musk's shitstorm 82
5.5 Distribution of toxic language on Twitter in Elon Musk's shitstorm. 83
5.6 Timeline of the eSports shitstorm on Twitter. 84
5.7 Use of emojis in eSports shitstorm, Twitter vs. Reddit. 84
5.8 Distribution of toxic language on Reddit and Twitter in the eSports shitstorm. . . 85
5.9 Distribution of the different groups over time in the Elon Musk shitstorm on Twitter. 85
5.10 Distribution of the different groups over time in the Elon Musk shitstorm on Telegram. 87
5.11 Distribution of different groups over time in the eSports shitstorm on Twitter and Reddit 88
5.12 Time course of the cross-platform shitstorm. 92

7.1 A German corpus of reflective setences: the main components of our approach. . . 116

8.1 PapagAI: system's diagram 135
8.2 PapagAI: example output 137

9.1 PapagAI user study II. The eight feedback rating questions results. 151
9.2 PapagAI user study II. The distribution of the most statistically significant linguistic features for Group A and Group B. 153
9.3 PapagAI user study II. The distributions of linguistic features with positive trends 155

A.1 Normalized averages from the comparative analysis. Linguistic features. 179
A.2 Normalized averages from the comparative analysis. Keywords. 180

A.3 Error analysis. Normalized averages of linguistic features for the groups of errors. . 181

A.4 Error analysis. Normalized averages of keyword occurrences for the groups of errors. 181

A.5 Elon Musk shitstorm: time distribution of keywords on Twitter and Telegram. . . 185

A.6 eSports shitstorm: temporal distribution of keywords on Twitter. 186

A.7 eSports shitstorm: temporal distribution of keywords on Reddit 187

List of Tables

2.1 Propaganda corpus: statistics and sources 24

2.2 Comparative results of propaganda detection models 27

3.1 Propaganda detection models' performance on 2022 and 2023 subsets. 42

4.1 The results of the German and Italian BERT and SVM models training experiments 62

9.1 PapagAI user study I. Adapted pedagogic feedback constituents. 145

9.2 PapagAI user study II. Survey questions. 149

9.3 PapagAI user study I. Affective indicators analysis. 151

9.4 PapagAI user study II. Statistically significant differences between Group A and Group B before and after feedback. 152

9.5 PapagAI user study I. User questionnaire results 152

9.6 PapagAI user study II. Statistically significant differences between Group A and Group B before and after feedback. 154

A.1 A German corpus of reflective sentences: linguistic features. 190

A.2 Metrics used to evaluate PapagAI's components. 191

A.3 Topics clusters from Bertopic model used in PapagAI 192

A.4 PapagAI emotion detection. Labels. 192

A.5 PapagAI: labels definitions for the reflective level model. 193

Acronyms

ACM Automated Content Moderation. 4, 14, 99, 101, 103, 106–109, 159, 168

AGB Allgemeine Geschäftsbedingungen. 148

AI Artificial Intelligence. 2–4, 6, 10–15, 95, 97–99, 101–110, 112, 118, 120, 130–132, 135, 139, 141–144, 146, 147, 149–152, 156, 166, 169, 170, 172

BERT Bidirectional Encoder Representations from Transformers. 6, 23, 27, 29, 34, 35, 39–42, 44–46, 49, 50, 54, 59, 60, 62, 69, 70, 72, 73, 127, 131, 133, 134, 168, 178, 184

BFGS Broyden–Fletcher–Goldfarb–Shanno. 178

CPU Central Processing Unit. 126

DNR Donetsk People's Republic. 33, 46, 47, 68

EC European Commission. 97–99, 103, 104

EU European Union. 22, 59, 95, 97–101, 103, 109

FDR False Discovery Rate. 148

FN False Negative. 27–29, 38, 43–45

FP False Positive. 27–29, 38, 43–45, 49

GPT Generative Pre-trained Transformer. 6, 9, 15, 126, 127, 136, 138, 139

GPU Graphics Processing Unit. 126, 127

HLEG High-Level Expert Group on AI. 99, 104

LLM Large Language Model. 9, 10, 73, 112, 126, 127, 130, 138, 139, 168, 170

LNR Luhansk People's Republic. 33, 46, 47, 68

ML Machine Learning. 101, 104, 131

MLP MultiLayer Perceptron. 27

NetzDG Netzwerkdurchsetzungsgesetz. 100

NLP Natural Language processing. 19, 34, 36, 55, 57, 58, 112, 130, 159

QWK Quadratic Weighted Kappa. 134, 191

RBF Radial Basis Function. 62

RLHF Reinforcement Learning from Human feedback. 10

RoBERTa A Robustly Optimized BERT Pretraining Approach. 127, 133, 136

RW Reflection Writing. 147, 151, 152

SLM Small Language Model. 127

SRL Self-Regulated Learning. 143

SVM Support Vector Machine. 18, 23, 27–30, 33–35, 38–40, 42, 44, 49, 54, 55, 59, 60, 62, 72

UMAP Uniform Manifold Approximation and Projection for Dimension Reduction. 134

URLs Uniform Resource Locator. 58

To all of the fallen soldiers, who stood for the dream of a free Ukraine.

Acknowledgments

I would like to thank my dear first supervisor, Christoph Benzmüller, for being there for me whenever I needed (even when I asked to give comments two days before the submission deadline), letting me be myself and make my own mistakes. I would like to equally appreciate my second supervisor, Tim Landgraf, who followed me in my day-to-day struggles, was my conscience whenever I was lazy to strive for perfection, taught me how to tell a story, and let me be part of a great family that his lab is. Thank you both for supporting my research on pro-Kremlin propaganda while many were afraid.

I would also love to acknowledge Tatiana Scheffler, who introduced me to social media analysis. This defined me as a scientist.

I wish to express gratitude to my partner, Lev Petrov, my source of inspiration and support, who criticised every plot of this thesis, gave multiple pieces of advice on statistical testing, and many other tips which were not always welcomed with gratitude they deserved. Also, many thanks to my family: my mom and grandmom for unconditional love and support, dad whose views influenced my choice of research topic and granddad, who inspired me to go the academic path.

Finally, thank you, nameless soldiers who protected the sky above my home, and all of the Ukrainian army, who have been shielding Europe while I wrote this thesis.

Hybrid AI Systems in Automated Content Moderation and Analysis

Abstract

Automated content moderation in a modern sense starts taking many forms: moderating social media, debates, therapy diaries and student learning processes such as essay writing. To tackle these tasks one can apply different AI techniques, such as classifications, information retrieval, chatbots, symbolic logical reasoners, and sometimes all of the above, combining them into so-called hybrid AI systems. Combining and running multiple AI components with different characteristics in a connected manner, or employing one model to elucidate another, emerges as a viable alternative to end-to-end systems. This is primarily because of their manageable and transparent nature, offering a potential improvement over end-to-end systems. Additionally, they may provide a more accurate representation of the various elements found in human cognition, blending resilient learning with rapid pattern recognition alongside reasoning facilitated by logical operations. In this thesis, two instances of hybrid AI systems are developed in combination with two content moderation use cases. "Check News in One Click" is a web application designed for streamlined news verification. It incorporates a fusion of statistical linguistic, transformer-based, and rule-based components that I developed and integrated into a productive system with a user-friendly interface. Specifically, this application specializes in verifying content from both conventional news sources and social media news channels, with a focus on identifying manipulative language and the presence of pro-Kremlin propaganda, which became a major problem in light of the Russian invasion of Ukraine. PapagAI is an online platform for higher

education students, where I created, combined and implemented an AI module for automated moderation of reflective essays using supervised models, a clustering, a linguistic processing module and a heuristic determiner which mines a prompt database for appropriate questions and amelioration suggestions. Through this application, my objective was to address the German educational system's requirement for improving teacher trainee retention rates at universities and easing the workload of tutors by streamlining the feedback process. In addition to the user tests, to evaluate the developed systems, here I also discuss questions related to the Ethics of AI, the European Union legal framework regarding automated content moderation, as well as the interpretability and sustainability of deep learning models.

Publications

Peer-reviewed Publications

The following publications contain scientific contributions and results produced as part of the doctoral research of this thesis:

1. [P1] Veronika Solopova, Oana-Iuliana Benzmüller, Christoph Benzmüller, and Tim Landgraf (2023). Automated Multilingual Detection of pro-Kremlin Propaganda in Newspapers and Telegram Posts. Datenbank Spektrum 23, 5–14.

   https://doi.org/10.1007/s13222-023-00437-2

2. [P2] Veronika Solopova, Christoph Benzmüller, and Tim Landgraf (2023). The Evolution of Pro-Kremlin Propaganda From a Machine Learning and Linguistics Perspective. EACL 2023. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics.

   https://doi.org/10.18653/v1/2023.unlp-1.5

3. [P3] Vera Schmitt, **Veronika Solopova**, Vinicius Woloszyn, Jessica de Jesus de Pinho Pinhal (2021). Implications of the New Regulation Proposed by the European Commission on Automatic Content Moderation. Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication, 47-51, doi: 10.21437/SPSC.2021-10.

   https://doi.org/10.21437/SPSC.2021-10

4. [P4] Veronika Solopova, Oana-Iuliana Popescu, Margarita Chikobava, Ralf Romeike, Tim Landgraf, and Christoph Benzmüller (2021). A German Corpus of Reflective Sentences. In Proceedings of the 18th International Conference on Natural Language Processing (ICON), pages 593–600, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

   https://aclanthology.org/2021.icon-main.72/

5. [P5] Veronika Solopova, Eiad Rostom, Fritz Cremer, Adrian Gruszczynski, Sascha Witte, Chengming Zhang, Fernando Ramos López, Lea Plößl, Florian Hofmann, Ralf Romeike, Michaela Gläser-Zikuda, Christoph Benzmüller & Tim Landgraf (2023). PapagAI: Automated Feedback for Reflective Essays. In: Seipel, D., Steen, A. (eds) KI 2023: Advances in Artificial Intelligence. KI 2023. Lecture Notes in Computer Science(), vol 14236. Springer, Cham, pp 198–206.

   https://doi.org/10.1007/978-3-031-42608-7_16

6. [P6] Veronika Solopova (2023). Automated content moderation using transparent solutions and linguistic expertise. In E. Elkind (Ed.), Proceedings of the Thirty-Second Interna-

tional Joint Conference on Artificial Intelligence, IJCAI-23 (pp. 7097–7098).: International Joint Conferences on Artificial Intelligence Organization. Doctoral Consortium.
© 2023 International Joint Conferences on Artificial Intelligence, Inc. (IJCAI). All rights reserved. This paper was originally published in the proceedings of IJCAI 2023. The final version is available at: https://doi.org/10.24963/ijcai.2023/823.

7. [P7] Tatiana Scheffler, **Veronika Solopova**, Mihaela Popa-Wyatt (2022). Verbreitungsmechanismen schädigender Sprache im Netz: Anatomie zweier Shitstorms. Rupert Gaderer, Vanessa Grömmke (Hg.): Hass teilen. Tribunale und Affekte virtueller Streitwelten. Bielefeld: transcript 2024 (Virtuelle Lebenswelten 3).
Accepted and set to be published in 2024.
Pre-print available in arXiv: http://arxiv.org/abs/2312.07194

8. [P8] Veronika Solopova, Viktoriia Herman, Christoph Benzmüller, and Tim Landgraf (2024). Check News in One Click: NLP-Empowered Pro-Kremlin Propaganda Detection. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 44–51, St. Julians, Malta. https://aclanthology.org/2024.eacl-demo.6/
This article is licensed under a Creative Commons Attribution 4.0 license.

Manuscripts

The following unpublished works, which are to be submitted to a journal or a conference, contain contributions and results produced as part of the doctoral research of this thesis:

1. [M1] Cheng Ming Zhang, Fernando Ramos López, **Veronika Solopova**, Florian Hofmann, Tim Landgraf, Michaela Gläser-Zikuda (2023). Automated feedback can foster deeper reflections.

Related Work not Part of This Thesis

My following publications contain work that is prior or related to the work in this thesis, but not considered part of this thesis:

1. [Rw1] Berfin Aktaş, **Veronika Solopova**, Annalena Kohnert, and Manfred Stede. 2020. Adapting Coreference Resolution to Twitter Conversations. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2454–2460, Online. Association for Computational Linguistics.

   https://aclanthology.org/2020.findings-emnlp.222/

2. [Rw2] Scheffler, Tatjana, **Veronika Solopova**, and Mihaela Popa-Wyatt. 2021. "The Telegram Chronicles of Online Harm". Journal of Open Humanities Data 7 (0), p. 8. DOI:

   https://doi.org/10.5334/johd.31

3. [Rw3] **Veronika Solopova**, Tatjana Scheffler, and Mihaela Popa-Wyatt. 2021. "A Telegram Corpus for Hate Speech, Offensive Language, and Online Harm". Journal of Open Humanities Data 7 (0), p. 9. DOI:

   https://doi.org/10.5334/johd.32

4. [Rw4] Scheffler, Tatjana Scheffler, **Veronika Solopova**, Zolotarenko, Olha, Razno, Maria. (2022). Automated Identification of Discourse Connectives in Ukrainian. In: Ermo-

layev, V., et al. Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2021. Communications in Computer and Information Science, vol 1698. Springer, Cham.

https://doi.org/10.1007/978-3-031-20834-8_5

1

Introduction

In the modern world, there exists a multitude of traditional tasks that humans may not want to or cannot handle on their own anymore. The sheer volume of user-generated content and an overwhelming number of requests produced by the digital age outstrips the capacity for traditional, human-led oversight (Moniz et al., 2021). This situation has led to the rising popularity and necessity of automated content moderation systems, offering a scalable and efficient solution to a problem that is rapidly outgrowing human capabilities alone. Automated content moderation has evolved to encompass various applications, including moderating social media, debates, therapy diaries, and student learning processes like essay writing.

Social media moderation gained special attention since social networks emerged as a key battleground for election interference and foreign influence campaigns, utilizing AI-driven personalized targeting and bot-generated comments to amplify perceived public sentiment (Bhatt & Rios, 2021; Yang & Menczer, 2023; Kruikemeier et al., 2022, 2016). This orchestrated approach aims to sway internet users towards a specific message, as witnessed in notable events like Brexit and the 2016 American elections (Chan, 2019; Narayanan, 2017). The dangers of the Kremlin propaganda war became especially apparent with the Russian full-scale invasion of Ukraine, supported by a massive international informational campaign (Anderson, 2018; Blank, 2022; Bokša, 2019). Several Western countries recognised that the problem was long overlooked and underrated, resulting in the striking unpreparedness to defend their informational eco-sphere (Lamberty & Frühwirth, 2023; Katerynchuk, 2017). In this sense, the need for an automated propaganda detector tailored to this particular type of propaganda became urgent, and to the best of our knowledge, no tool was available.

Our research was motivated by whether this source of propaganda can be not only efficiently but also transparently detected. We think that the knowledge one can learn from the propaganda patterns the AI systems identify may also teach us how to detect it better as humans and thus weaponise an average citizen against its influence. That's where linguistic knowledge, long overlooked by the NLP community, becomes central. Horne et al. (2020); Bozarth & Budak (2020) show that in times of major events, trained models start struggling quickly on the new data. Thus, it is a challenging task to train transparent systems to be robust and generalised enough and also stay highly performant over time as the events of war evolve.

The implementation of such a tool invites an interdisciplinary outlook. For instance, several questions related to the field of User Studies become relevant here, as it is unclear in which form (e.g. browser extension or a website) such system should be built to be the most useful to a lay user and how these assumptions may be tested. With automated detectors' decisions often used to enforce community rules, not only Freedom of Speech but also someone's income can be at stake when misclassifications happen. Since the current and the soon-to-be-implemented European legal frameworks regulate such systems, the questions of automated content moderation in social media are also strongly inter-related with the field of Ethics of AI and Law.

Another fruitful field of application of automated content moderation is education. With the growing population (Sadigov, 2022) and increased study-related migration (Weber & Van Mol, 2023), optimisation of educational processes, for instance, in higher education, also becomes a matter of high priority. Since we as a society are committed to providing quality education to a broad demographic, leveraging AI to enhance the efficiency of staff time investment becomes imperative.

One of the important tasks to automatize remains essay evaluation, which is a central task of both secondary and higher education and a popular method for formal assessment of student educational achievements. While argumentative essay scoring received considerable scientific attention Ramesh & Sanampudi (2022), there is a notable gap in understanding the other genres, such as reflective essays, that have become popular in teaching education, medical training, pharmaceuticals, and computer science is largely under-researched in comparison Olex et al. (2020); Liu et al. (2019b); Chen et al. (2019); Wulff et al. (2020). The challenges addressed in this research include determining the linguistic and other features of the essay required to capture the quality of reflection, training the models for the optimal automatic extraction of the identified features, and identifying and constructing the most optimal system architecture. The main purpose of this investigation was to determine the effectiveness of automated systems in helping students learn to reflect deeper, compared to the usefulness of human feedback in the same setting.

The motivation behind this thesis is multi-faceted, with both abstract and technical dimensions. At a higher conceptual level, the solutions I have developed align with the public demand for enhanced automated content moderation techniques in social media and education. Concurrently, the scope of this thesis encompasses several additional technical contributions. These include the optimization and refinement of various transformer models for text classification, exploration of post-hoc interpretability and error analysis techniques, transferability of the developed features to another task and the investigation of ways to incorporate linguistic knowledge into the hybrid AI productive systems and models, aiming for increased transparency without sacrificing robustness. To gain a comprehensive understanding, let us now delve into the background of automated content moderation and review relevant work, particularly in the context of social media and education.

1.1 Automated Content Moderation

1.1.1 Social media

Automated Content Moderation (ACM) is often understood as the moderation of social media content. As social media grew in popularity, so did the amount of user-generated content, which since its early days includes a wide range of malicious activity and anti-social behaviour (Dibbell, 1994): hate speech, online harassment, cyberbullying, misinformation, trolling, spamming, conspiracy theories and many more.

Minorities and historically oppressed groups often become the main targets. For example, the analysis of a multilingual dataset of extreme speech by Udupa et al. (2023) reveals that the predominant forms of hate speech vary across countries: anti-immigrant and Islamophobic messages in Germany, attacks on religious minorities in India, targeting of women and sexual minorities in Brazil, and predominantly ethnic group-directed extreme speech in Kenya.

Spread. It has been shown that Fake news spreads faster than true news on Twitter, and the main spreaders are real people and not bots (Vosoughi et al., 2018). In our joint work (Scheffler et al., 2021) we showed how hate speech also may be "contagious". In our dataset, the posts written in response to other posts containing harmful language had a significantly higher probability of containing harmful language themselves. In this thesis, we also investigate how the internet scandal migrates and propagates into different platforms with the help of community activists (Chapter 5). Several other types of research confirm the idea of rule-breaking behaviour spreading from comment to comment (Popa-Wyatt, 2023; Cheng et al., 2017; Kim et al., 2021), while some results have not shown significance (Han & Brazeal, 2015; Rösner et al., 2016). Such online harms as fake news and hate speech are also said to provoke offline harms, such as inciting real-life violence or convincing the users not to get vaccinated during a deadly pandemic (Horne, 2023).

Countermeasures. A high number of dangerous content attracted legislative and community attention, calling for the platforms to eliminate harmful content. Many moderation approaches are used, ranging from harsh to soft: banning repeated rules-infringers, quarantine communities (like subreddits), demonetizing content, removing, demoting a post, making it challenging to discover on a platform, such as shadow-banning and deprioritizing (Horne, 2023), blurring and flagging, which placing a warning label on the post about the potential lack of credibility or accuracy of the content (discussed further in Chapter 6). In reality, small policy teams manage a large number of underpaid human moderators (Ye et al., 2023). In 2017-2018, most content moderators were contracted by firms in India and the Philippines, with a full-time salary ranging between $300 and $500 a month (Chen, 2017; Roberts, 2017). Being on a better payroll in the US, the moderators usually do not receive corporate benefits nor the mental health support provided. Meanwhile, evidence of social media moderators suffering from psychological trauma leading to insomnia and anxiety because of exposure to toxic content is accumulating (Stone, 2010; Steiger et al., 2021; Chen, 2017).

Automated Solutions and their Limitations. The scale of the problem quickly attracted the industry's attention towards automated solutions, claiming it should solve the problem entirely or at least partially alleviate the burden that human moderators overtook (Gillespie, 2020). Automated moderation mechanisms developed by the big platforms from the very beginning **lacked transparency** and attracted a lot of criticism. Different moderation tools tend to be marketed as "AI": not only machine learning techniques but also often simple pattern matching using a blacklist which produces a high amount of both false positives and false negatives (Gorwa et al., 2020). There are, however, many successful works on hate speech (Davidson et al., 2017; Kiela et al., 2020; Basile et al., 2019), trolling (Kumar et al., 2014; Shafiei & Dadlani, 2022; Áine MacDermott et al., 2022), bot activity (Alsmadi & O'Brien, 2020; Hostiadi & Ahmad, 2022; Hostiadi et al., 2020), cyberbullying (Rahat Ibn Rafiq et al., 2015; Raj et al., 2022; Reynolds et al., 2011) and online harassment detection (Stoop & Kunneman, 2019; Abarna et al., 2022; Ahirwar et al., 2022; Marwa et al., 2018).

From a technical point of view, the methods evolved from methods using Bag of Words, Term Frequency-Inverse Document Frequency (tf-idf) and n-grams to such embedded language representations as (Schmidt & Wiegand, 2017), Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and its siblings like Electra (Clark et al., 2020) and XML-Roberta (Conneau et al., 2019). Since 2023, content moderation using GPT-4 (OpenAI, 2023) emerged as a trend (Bernard, 2023).

Adaptability is a major concern, as user content is inventive and ever-changing, so the policies should be able to evolve with them (Sinnreich, 2018). Machine detection is easy to trick using misspellings, changed syntax, and codes (Burnap & Williams, 2015; Gröndahl et al., 2018; Scheffler et al., 2021). Automated tools are also limited in their capacity to account for unseen contexts and differences in dialects (Ribeiro et al., 2018; Sap et al., 2019). They are unable to detect terrorist propaganda cited in a journalistic context (Llansó, 2019). They also find challenging such linguistic phenomena as irony, sarcasm, and understatements and are insensitive to the use of the same content in/out-group settings, by thus, marginalizing and disproportionately censoring groups that already face discrimination (Duarte et al., 2017; Buolamwini & Gebru, 2018).

This lack of adaptability is also hard to capture, as the models are tested on the data of the same distribution as the training samples (Bengio et al., 2021), as we also see in Chapter 4, reported scores for performance can be **misleading**. Hence, according to Horne et al. (2020); Bozarth & Budak (2020), models trained on U.S. news stumble on British articles, those trained on data biased towards one political leaning may get confused by articles from another, while models trained on one period of time, or during major events decrease in performance on unseen years and new events, which we also investigate in Chapter 3. The dataset construction is also complicated due to raw data from real distribution being highly imbalanced, where only ≥ 30% of samples include the phenomena we seek to moderate (Zampieri et al., 2019b). In the joint work on hate speech in Telegram channels of President Trump supporters (Scheffler et al., 2021) only 17% of annotated samples were harmful.

Facebook's claim to successfully flag 65.4% harmful content before users' reports sparked a lot of controversy. Gillespie (2020) claimed that this is "deliberately misleading, implying that machine learning techniques are accurately spotting new instances of abhorrent content, not just variants of old ones." In contrast, according to Meta's analysis Ribeiro et al. (2023), their comment deletion policy effectively decreases subsequent rule-breaking behaviour, while the minimizing effect of rule-breaking stayed longer than its effect on reducing commenting in general. Hence, users would not just stop commenting at all but would start adhering to community guidelines in their comments. On the other hand, hiding content, as a moderation method, had statistically insignificant effects. Removing content and providing explanations for removal also showed to reduce rule-breaking behaviour on Reddit (Srinivasan et al., 2019; Jhaver et al., 2019). Chat modes on Twitch used as proactive moderation, proved effective at discouraging spamming, while it was concluded that reactive bans prevented a wider variety of subsequent rule-breaking (Seering et al., 2017).

The Consequences of Mismoderation. An endless struggle between freedom of speech and community safety unfolds in the vastness of social media platforms and the Internet as a whole. While being the human moderator has been shown to lead to negative psychological effects, being the moderated without transparency often feels unjust and misunderstood (West, 2018), insignificant, and failed at being seen as an individual by the mega machines (Hill, 2019). Especially, as some norms are universal while others can be unique to each specific platform or even forum (Chandrasekharan et al., 2018).

Meanwhile, the decision to block an account can have strong consequences, and blocking appeal mechanisms are not always effective. Back in 2016-2017, it took a lawsuit for the H3H3 channel for YouTube to revise its Fair Use policies concerning the usage of clips of other YouTube channel videos in a commentary genre (Chan, 2016). It took 613 days for a famous Twitch streamer, Tyler1, to have his case reviewed by Riot Games and his ban lifted 613 in League of Legends.

The case of the Russian war against Ukraine, which became a background to many chapters of this thesis, showed how crucial it is for a social media platform to adapt to the evolving community standards, taking historical context and current events into account when covering sensitive content. According to Yuskiv (2023), after the retreat of Russian troops from Kyiv and the uncovering of massive graves, Instagram restricted numerous hashtags for not fitting the community rules, including #Bucha, #Irpin, #BuchaMassacre, #Azov, #RussiaIsATerroristState, and #StandWithUkraine, #russiaisaterrorisstate. According to Meta, this was caused by massive complaints and not special sanctions (Yuzefyk, 2022). On the 23rd of February, Twitter, currently X's representative Yoel Roth, confirmed that Twitter had mistakenly blocked several OSINT-reporters accounts, which were posting data about the build-up of Russian army forces near the Ukrainian border accounts. [1] In September 2021, fa-

[1] https://twitter.com/yoyoel/status/1496544199478583297.Lastaccess:09.11.2023

mous Ukrainian illustrator Olena Pavlova reported that Facebook blocked her illustration of classical Ukrainian literature with the caption "Read classics, like kitties".[2] Many Ukrainian bloggers have reported being shadow-banned, meaning their content was hidden and never appeared in the recommendations (Фіялка, 2022).

Nonetheless, user criticism provokes specific adaptations. The Ministry of Digital Transformation of Ukraine appealed to Meta, which, in response, revised its content moderation policy for war coverage. Meta's Oversight Board overturned the decision to remove a post comparing Russians to Nazis and admitted that the body of the victim of the Bucha massacre did not violate the requirements of the Meta Policy on the depiction of graphic content (Yuskiv, 2023). According to Digital Transformation of Ukraine, Meta also pledged not to block content about the Azov regiment.[3]

Anglo-centric Approach. It is also important to highlight that the moderation efforts and underlying funding are unevenly and unequally distributed. Hence, non-Western languages are more prone to be under and miss-moderated (Udupa et al., 2023; Nicholas & Bhatia, 2023). For instance, 87% of Facebook's global budget, which is spent on misinformation, is set for the United States, while only 13 per cent is designated for the rest of the world. At the same time, users from North America account for only 10 per cent of daily users of this platform (Frenkel & Alba, 2021). Recently introduced Large Language Model (LLM), such as the aforementioned GPT models, claimed to approach the answer to linguistic imbalance. These models are trained on sizeable multilingual text corpora. Although the exact language proportions are not publicly disclosed, in GPT-4 System report OpenAI (2023) it is mentioned that English texts occupy a substantial part of the training set. However, generative models were shown to understand lower-resource languages surprisingly well by learning their connections to higher-resource ones (Artetxe et al., 2020; Nicholas & Bhatia, 2023).

[2] https://twitter.com/Nympha_Blin/status/1439152593566044161

[3] https://www.kmu.gov.ua/en/news/meta-poobitsiala-ne-blokuvaty-kontent-pro-azov-rezultaty-domovlenosti-z-mintsyfry

The main critics of this approach indicate that this way LLMs import English-language assumptions and viewpoints, namely the Anglocentric frame. Nicholas & Bhatia (2023) illustrate the disadvantages of this import for automated content moderation with the word “dove”. Being a symbol of peace in many Western languages and cultures, in Basque, it is often used in derogatory and homophobic contexts.

According to Murgia (2023), ChatGPT testers reported that it refused to generate recruitment propaganda for terrorist groups when promoted in English but did not have any scruples to generate it in Farsi. Another consequence of the Reinforcement Learning from Human feedback (RLHF) technique applied to minimize the presentation of offensive content to users may lead to unexpected phenomena, namely, that the model struggles to recognize hate speech, irony, and offensive language (Zhang et al., 2023a).

With all this being said, corporate lack of transparency for how these methods are applied is intensified by the newest algorithms being less and less transparent. Despite efforts to apply explainable AI methods to transformer algorithms, mainly through attribution methods like Integrated Gradients (Janizek et al., 2021; Sundararajan et al., 2017a), the results have not yet proven to be convincing. As automated algorithms are slower to adapt to the context of the ever-changing concepts of community good, the research has shifted towards the idea of AI as an enhancing tool for a moderator.

1.1.2 Education

Although Social Media is one of the most popular areas for automated content moderation, moderating a diary or therapeutic conversation was there at the dawn of AI tools. ELIZA is one of the earliest chatbot examples created for psychological purposes (Weizenbaum, 1966), while many new modern applications moderating an internal dialogue appear (Kapoor et al., 2021; He et al., 2022).

Meanwhile, AI is also successfully learning to moderate debates. To the best of my knowledge, IBM's Debator (Slonim et al., 2021) is the most prominent automated debating system and debate moderator, which is based on argumentation mining.

Chatbots. Another fruitful area has been Education, namely technology-mediated learning (Winkler & Söllner, 2018). Here, AI tools manifest themselves usually in the form of chatbots and learning platforms, enhancing the educational experience with personalized and adaptive components (Chassignol et al., 2018; Valderrama et al., 2011).

At this point, more than 36 educational chatbots were proposed in the literature, with educational areas mainly including computer science, language, and education (Kuhail et al., 2023). Famous examples include Replika (Pentina et al., 2023), an AI chatbot companion for students, which unites educational and psychological help; Piazza (Ruthotto et al., 2020; Wang et al., 2020b), is a collaborative educational discussion facilitator and moderator; Ada tutoring chatbot, answering questions and providing feedback (Kabiljo et al., 2020; Konecki et al., 2023).

Beneficial Effects. According to Labadze et al. (2023) 67-study review, in terms of AI tools, the students benefit the most from homework and study assistance, personalized learning experiences, and the development of various skills. At the same time, educators appreciate time-saving and improved pedagogy. In Chaiprasurt et al. (2022), the study showed that students agreed that chatbot facilitated their learning. Most studies also indicate that chatbots improve student motivation (Okonkwo & Ade-Ibijola, 2021a; Wollny et al., 2021; Neji et al., 2023; Okonkwo & Ade-Ibijola, 2021b).

Limitations. While AI feedback teaches students to be more autonomous, human teachers' feedback is still perceived as more relatable (Chiu et al., 2023). In our study in Chapter 9, the results do not show a significant difference between student acceptance of AI feedback versus real Tutor feedback, and neither does it show better learning outcomes following either source of feedback. Therefore, AI feedback systems may be advantageous when individual support is difficult to organise, such as large-scale lectures and online courses where hundreds of students participate (Winkler & Söllner, 2018).

Meanwhile, AI tools also raise many concerns for accuracy, reliability, data privacy, and ethical issues (Labadze et al., 2023). Consequently, as in the case of social media moderation, most researchers conclude that AI tools should not and cannot be a full replacement for human emotional support and interactive learning and may only complement them (Annuš, 2023). This is where the idea of Hybrid AI systems comes into play, where the term "hybrid" calls for further clarification.

1.2 Hybrid Artificial Intelligence

The term "Hybrid Artificial intelligence systems" may be used by the AI community in two ways, both of which are relevant to this thesis. In case (1), we talk about hybrid human AI intelligent systems, or human-in-the-loop approach, where AI support human decision-making and humans and smart systems collaboratively accomplish goals (Bredeweg & Kragten, 2022), which is relevant for autonomous driving and smart industrial equipment (Ostheimer et al., 2021) and also both social media (Link et al., 2016) and education (Baker, 2016), which are the use cases of this thesis. The second understanding of this term (2) refers to symbolic-statistical intelligent systems that may combine multiple symbolic (reasoning), sub-symbolic (deep learning), knowledge-bases (ontologies) and cognitive system elements (van Bekkum et al., 2021). In the case of Check News in 1 Click application (Chapter 4), we explore what van Bekkum et al. (2021); Sarker et al. (2017) defines as "Explainable learning systems through rational reconstruction". According to their scheme, the result of a prediction of a trained machine learning (ML) system (A) is passed on to a symbolic reasoning system (B) which uses background knowledge to produce a "rational reconstruction" of a reason to justify the input/output pair of the learning system. That involves a post-hoc justification, which may not necessarily reflect the actual statistical computation of model (A).

In the case of the PapagAI app system (presented in Chapter 8) it is more similar to "Learning an intermediate abstraction", where a module containing multiple models (A) collects knowledge, which is processed by a rule-based reasoner (B) to make the final decision about the feedback. The architectural choice in both cases is called to bring more control and interpretability to remedy the "black-box" problem (Weld & Bansal, 2018). The hybrid, knowledge-driven, reasoning-based approach is also what Marcus (2020) defines as a path to robust AI in contrast to the current data-centred- deep learning approach, which saw building human knowledge into ML systems as cheating and yet was not able to solve that problem despite the resources invested. In response, the typical criticism around Hybrid AI mentions scalability, negative past evidence (these systems were outperformed in the past) and the fact that there is no established neuro-scientific evidence that our brain works in symbols. According to Marcus (2020), while scalability is indeed an open problem, some of the most commercially successful AI examples such as Google Search, Open AI Rubik (OpenAI et al., 2019), and ChatGPT are hybrid systems with symbolic algorithm and deep reinforcement learning. As for the biological plausibility, psychological evidence supports the idea that symbol-manipulation happens in the brain, with "the ability of infants to extend novel abstract patterns to new items" being a primary example (Iris Berent, 2007; Gallistel & King, 2010). In this work, the scalability and performance issues are mostly confronted.

1.3 Goals and Contributions

The work on transparent Pro-Kremlin propaganda detection, presented in this thesis as a case of social media moderation, is a pioneering work, the first of its kind, as the topic seemed to be either underrated or even taboo in the research community before, with most scientists skirting the issue and focusing on other propaganda sources instead. The data analytical part, including chronological analysis of the evolution of propaganda, gives an important insight into the functioning of modern informational warfare. The final product of this research is a web application, Check News in 1 Click, which was built to help raise personal awareness and responsibility for the content internet users consume. The user study we conducted with this application sheds light on user preferences and points of interest for content moderation. The collected data, models, code and website styling are all open-source and are already used by the research community. The application of linguistic feature extraction script on another content moderation task, namely shit-storm detection, shows potential for transferring knowledge across content moderation tasks. Analysis of the AI regulations proposed by the European Union about automated content moderation fulfils the ethical and legal framework of this part of the thesis. It contributes to the community's efforts to ensure the most optimal laws are adopted. The PapagAI, created as a tool for reflection moderation for higher education students, is also an ambitious and continuing research. It was envisioned as part of the answer to the immense teacher shortage problem in Germany as well as disproportionate dropping out rates from teacher education in German universities (Becker, 2021; Klemm & Zorn, 2019). Reflection writing, as one of the compulsory and central tasks in Teacher Education, requires timely tutor feedback, which usually takes months to get. That is why quick, efficient, automated moderation can be beneficial. The created system involved training of multiples language understanding models, with all the annotated data set, known as German Corpus of Reflective sentences Solopova et al. (2021) presented in Chapter 7, as well as the models, were released and available to the community. The user tests, although lacking in quantity of participants, show optimistic results: AI can be used in higher education, and young teacher trainees guided by AI feedback, despite generally low AI acceptance, show on-par results to those receiving human feedback, and improve their reflection skills. The objectives of this thesis can be outlined as follows:

1. In developing and interpreting both deep learning and linguistic features-based models, investigate the trade-off between transparency of automated decisions and accuracy in ACM;

2. Develop, deploy, user-test and provide to the public ethical and user-friendly interfaced solutions to content moderation in social media and education using suitable hybrid AI architec-

ture;

3. Contribute to the research community with the collected and manually labelled training data, trained models and code available open-source, facilitating future studies and investigations in the field.

1.4 Outline and Synopsis

This thesis is structured in two thematic parts: social media moderation and educational reflective practice moderation. In Chapter 2, we present our novel approaches for propaganda detection. We verify how these models generalise on the data collected one year after the data used in the training set in Chapter 3. In Chapter 4, we demonstrate the resulting hybrid system behind the GUI application and the user study results. Then, we reapply the same feature extraction to the chronological internet scandal modelling task in Chapter 5 and analyse the patterns of the life cycle of the shit storms and the induced online hate. This part is finalized with Chapter 6, where we look into the legal and ethical framework of automated content moderation in light of the new AI regulation proposed by the European Commission. Chapter 7 opens the reflection moderation part, describing the annotated corpus and the linguistic features we developed for the task, as well as statistical analysis of the corpus. The models and system description of the PapagAI app are discussed in Chapter 8, with the proposed models evaluated against GPT-3.5 model. Concluding the essay moderation part with Chapter 9, we assess different PapagAI feedback formats and their effectiveness with students. Finally, Chapter 10 focuses on the various linguistically inspired contributions of this thesis and my research as a whole, including the hate speech detection topic, while the overall outlook, limitations and future work are discussed in Chapter 12.

We shall fight on the beaches, we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills; we shall never surrender.

Winston Churchill

2

Automated Multilingual Detection of Pro-Kremlin Propaganda in Newspapers and Telegram Posts

Joint work with **Oana-Iuliana Popescu**, German Aerospace Center, Jena, Germany.

Supervised by **Christoph Benzmüller** and **Tim Landgraf**.

2.1 Preface

This Chapter was previously published as: Solopova et al. (2023b). Automated Multilingual Detection of pro-Kremlin Propaganda in Newspapers and Telegram Posts. Datenbank Spektrum 23, 5–14 (2023). https://doi.org/10.1007/s13222-023-00437-2

In this Chapter, we investigate if propaganda can be detected through statistical linguistic features as well as using deep learning models to understand the potential trade-off between accuracy and transparency in automated content moderation. While several previous studies focused on linguistic indicators of propaganda, the pro-Kremlin state-sponsored propaganda that we target was and still is overwhelmingly under-researched in comparison to those originating in the United States and China, including such instances as green-washing (Ende et al., 2023; Kim et al., 2023) and agenda-driven publications sponsored by pharmaceutical companies (Fabbri et al., 2018). The absence of appropriate tools to counteract the Kremlin's informational campaign became especially evident with the Russian full-scale invasion of Ukraine. The paper presents a hybrid approach to detecting and analysing manipulative language consistent with pro-Kremlin narratives in an explainable and robust fashion, allowing for a comparative study of how this kind of propaganda of a common origin manifests itself in different languages and media types: classic news outlets and telegram news channels. In addition, to the best of our knowledge, "Automated Multilingual Detection of pro-Kremlin Propaganda" is the first application of transformers technology to train a propaganda detection model. It also offers an extensive comparison and discusses the trade-off between the robustness and inherent transparency of classical algorithms empowered by linguistic features. The study also contributes to the general knowledge of multilingual models based on transformer architecture. Here, we investigated what is the minimal amount of data from different genres and languages required for the model to be optimised enough to recognise them efficiently. The models from this paper were used in the following Chapter 3 and 4, where they are at the core of the algorithm behind the Check News in One Click application. The script for linguistic feature extraction created for this study is also used in Chapter 5, where it is shown to be helpful for the analysis of other social media phenomena.

2.2 Contributions

Conceptualization: [Veronika Solopova];
Methodology: [Veronika Solopova (70%), Oana-Iuliana Popescu (30%)]. Formal analysis and investigation: [Veronika Solopova(70%), Oana-Iuliana Popescu(30%)]. Data curation: [Veronika Solopova]; Visualization: [Veronika Solopova (60%); Oana-Iuliana Popescu (40%)] Writing - original draft preparation: [Veronika Solopova];
Writing - review and editing: [Veronika Solopova, Tim Landgraf, Oana-Iuliana Popescu];
Supervision: [Tim Landgraf, Christoph Benzmüller];
Project administration: [Veronika Solopova].

Oana-Iuliana Popescu contributed to the collection of data and verified features and dictionaries for the Romanian language; she also proposed SVM co-efficient analysis, and its visualisation. Oana-Iuliana Popescu and Tim Landgraf contributed strongly to the review and editing of the final manuscript.

I contributed to the design and development of the feature extraction scripts and dictionaries for each language, as well as data collection for all of the languages, except Romanian, and further dataset curation for all languages. I also conceptualised experimental design and trained and evaluated the models in focus, analysed the data using the extracted linguistic features and interpreted the feature importance of the SVM, also creating the corresponding visualizations. Finally, I completed writing the entire first draft of the paper and eventually became the corresponding author, fully responsible for the submission processes, communication with the editors, the implementation of the reviewers' comments and the camera-ready version.

Abstract

The full-scale conflict between the Russian Federation and Ukraine generated an unprecedented amount of news articles and social media data reflecting opposing ideologies and narratives. These polarized campaigns have led to mutual accusations of misinformation and fake news, shaping an atmosphere of confusion and mistrust for readers worldwide. This study analyses how the media affected and mirrored public opinion during the first month of the war using news articles and Telegram news channels in Ukrainian, Russian, Romanian and English. We propose and compare two multilingual automated pro-Kremlin propaganda identification methods based on Transformers and linguistic features. We analyse the advantages and disadvantages of both methods, their adaptability to new genres and languages, and ethical considerations of their usage for content moderation. With this work, we aim to lay the foundation for further development of moderation tools tailored to the current conflict.
Keywords: Propaganda, Fake news, NLP, Kremlin, Ukraine, Automated Content Moderation.

2.3 Introduction

Propaganda influences an audience to support a political agenda (Smith, 2022; n. OED Online. Oxford University Press, 2022). Propaganda has been shown to play a vital role in the Russian invasion of Ukraine, shaping the war approval rate (Khvostunova, 2022) by, e.g. fabricating explanations for war crimes (Roth, 2022). As a result, fake news also spreads through Ukrainian, Central European and Western media (Heritage, 2014), seeding mistrust and confusion Paul (2022).

With every day of the war having a large amount of potentially false information produced, human quality control thereof is limited. Especially during a war, the journalistic virtue of fact-checking may be substantially obstructed. This poses the question of whether statistical analysis can provide us with a reliable prediction of the intent behind a piece of news. Given a sufficiently high separability, automatic moderation tools could process the news and warn readers about the potential disinformation instead of entirely placing the responsibility on human moderators (Steiger et al., 2021). This study aims to detect war propaganda produced in the context of the 2022 Russian invasion of Ukraine in a transparent, explainable way, as such a tool can be used for content moderation in social media.

While Russian propaganda creates a 'cacophony' of fabricated opinions and sources in its media ecosystem (NATO Strategic Communications Center of Excellence, 2016), it also has several uniform strategies and approaches which are recurrently mentioned in research throughout the whole of the Russian-Ukrainian war by international bodies and independent researchers (Carroll, 2017; Meister, 2016; Fortuin, 2022; U.S. Department of State, 2020). Hence, we hypothesize and aim to prove that propaganda can be successfully detected using certain stylistic and syntactical features behind these strategies, independent of keywords. Naturally, keywords change depending on the course of events, while the tactics of the propaganda stay similar. Traditional algorithms empowered by such features are inherently "interpretable" and may perform on par with non-transparent neural networks. Here, we propose a linguistics-based approach for detecting war propaganda in the context of the ongoing war between Ukraine and Russia since the start of the full-scale invasion in 2022, using models trained on news from media outlets from Ukraine, Romania, Great Britain and the USA. We extracted news from fact-checked outlets identified as credible sources and outlets recognized as spreading fake news. We train and evaluate classifiers using a set of linguistic features and keywords and a multilingual Transformer to classify articles as containing pro-Kremlin and pro-Western narrative indicators.

With this work, we provide an open-source tool for the identification of such fake news and propaganda in Russian and Ukrainian, which, to the best of our knowledge, is the first of its kind. We demonstrate that discriminating propaganda from neutral news is not entirely possible in the current situation, as news from both sides may contain war propaganda, and Western media is heavily dependent on Ukrainian official reports.

In Section 2.4 we present previous work related to our research. In Section 2.5, we introduce our training setup for each experiment, describing the data and model configurations. In Section 2.6 we first describe the sources of our data and its collection process (3.1-3.2), then, we expand upon the linguistic features (3.3) and the keywords that we extract (3.4). In Section 2.7, we present the results for each setting, while in Section 2.8 we provide additional analyses, looking into feature importance coefficients of some models (6.1) and distributional exploratory analysis of the classes (6.2), exploring chronological, language and narrative-specific differences. In Section 2.9, we consider our work's main findings and limitations. We lay the ground for future work opportunities and delve into ethical dangers in terms of its potential usage for automated content moderation. In Section 2.10, we summarize the main contributions of our study.

2.4 Related Work

To address the issue of human quality control limitations and minimize the number of snippets a human moderator has to check, the automated fact-checking research investigates many potential solutions, such as identifying claims worth fact-checking, detecting relevant fact-checked claims, retrieving relevant evidence, and verifying a claim (Nakov et al., 2021).

Our work is motivated by the fact that despite the assumed involvement of Russia in Brexit (Narayanan, 2017) and the 2016 US presidential elections (Cosentino, 020a), there is still only a small number of peer-reviewed research publications investigating Russian state-sponsored fake news (Elswah & Howard, 2020; Beskow & Carley, 2020). Furthermore, existing publications are not always neutral, with some using accusations and emotional lexicon (Rosulek, 2019), while others accuse Western media of propaganda used to discredit Russia (Chudinov et al., 2019).

Wilbur (2021) examines claims of Russian propaganda targeting the US population by analysing weblogs and media sources in terms of their attitude towards Russia and finding a positive correlation between the Russian media outlet Sputnik and several US media sources.

Timothy Snyder in his books (Snyder, 2018; Sly, 2017) analyses the Kremlin's informational campaign against the sovereignty of the United States and the integrity of the European Union.

Several studies investigated Russian bots in social media. Alsmadi & O'Brien (2020) used a decision tree classifier on features extracted from the tweets, concluding that bot accounts tend to sound more formal or structured, whereas real user accounts tend to be more informal, containing slang, slurs, and cursing; Beskow & Carley (2020) analyzed the geography and history of the accounts, as well as their market share using Botometer, pointing to a 'sophisticated and well-resourced campaign by Russia's Internet Research Agency'. Narayanan (2017) performed a basic descriptive analysis to discern how bots were being used to amplify political communication during the Brexit campaign. There is a considerable amount of research focusing on fake news detection. Asr & Taboada (2019) show-cased that significant performance can be achieved even with n-grams; Antoun et al. (2020); Mahyoob et al. (2020); Rashkin et al. (2017) implemented different linguistic feature-based solutions, while Li et al. (2021) demonstrated the application of Transformers. While fake news makes up a big part of the propaganda toolkit, propaganda also relies on specific wording, appealing to emotions or stereotypes, flag-waving and distraction techniques such as red herrings and whataboutisms (Yu et al., 2021). Research on propaganda detection is less frequent. Although existing works proposed using both feature-based and contextual embedding approaches (Tundis et al., 2020; Yu et al., 2021; Oliinyk et al., 2020; Dadu et al., 2020), these studies focused mostly on the English language. To the best of our knowledge, there are no benchmark corpora and no open-source multilingual tools available. To address this research gap, we identified key questions we aim to answer with our study:

- Which machine learning approach is best suited for the task of propaganda detection and what is the trade-off between the accuracy and transparency of these models?

- Can propaganda be successfully detected only with morpho-syntactic features?
- Do linguistics features differ significantly among languages and sides of a conflict?

2.5 Methods

We implement a binary classification using the following models for input vectors consisting of 41 handcrafted linguistic features and 116 keywords (normalized by the length of the text in tokens): decision tree, linear regression, Support Vector Machine (SVM) and neural networks, using stratified 5-fold cross-validation (10% for test and 90% for training). For comparison with learned features, we extract embeddings using a multilingual BERT model (Devlin et al., 2019) and train a linear model using them.

We performed 3 sets of experiments contrasting the handcrafted and learned features:

Experiment 1. Training models on Russian, Ukrainian, Romanian and English newspaper articles, and evaluating them on the test sets of these languages (1.1) and on French newspaper articles (1.2). We add the French newspapers to benchmark the multilingualism of our models. We choose French because it is in the same language family as Romanian.

Experiment 2. Training models on Russian, Ukrainian, Romanian, English and French newspaper articles, and validating them on the test set (2.1). Additionally, we use this model to test the Russian and Ukrainian Telegram data (2.2.). Here the goal is to investigate whether this model will perform well out-of-the-box for the Telegram articles, which are 10 to 20 times shorter. See an example of the genre-related difference in distributions in Figure 2.1.

Experiment 3. Training models on the combined newspaper and Telegram data and applying them to the test set. Here we verify whether adding the Telegram data to the training set can improve generalization power, although data distributions differ.

Table 2.1: Corpus statistics, including the sources per language and stance.

Language	Source	Amount of texts
pro-Western newspapers		
Ukrainian	‘Europeiska Pravda’,‘Ukrainska Pravda’, ‘Espresso’, ‘5.ua’, ‘Hhromadske’, ‘Liga.net’	3298
Romanian	‘digi24’, ‘mediafax’, ‘g4media’	
English	‘The Guardian’, ‘BBC’, ‘The New York Times’, ‘Reuters’)	
French	‘Tv5monde’, ‘Le Monde’ and ‘Le Figaro’	458
Russian	‘Raintv’	7
Pro-Kemlin newspapers		
Ukrainian	‘Newsua’, ‘Strana.ua’, ‘Vesti.ua’, ‘Ukranews’, ‘Zik’	3579 (474 in Ukrainian and 3105 in Russian)
Romanian	‘Antena3’, ‘Stiripesurse’, ‘Romaniatv.net’, ‘Cyd.ro’, ‘Activenews’ and ‘Dcnews’.	3007
French	‘RT’ French edition	123
Russian	‘Ria news’, ‘Russia Today’, ‘Interfax’, ‘Lenta.ru’ and ‘Ukraine.ru’.	2648
Telegram posts		
Ukrainian	‘Goncharenko’, ‘InformNapalm’,‘Brati po zbroi’, ‘Spravdi’,‘Operativni ZSU’	7263 (1568 in Ukrainian and 1568 in Russian)
Russian	‘Rybar’,‘Siloviki’,‘Vysokigovorit’	61595

2.6 Data

2.6.1 Newspapers

We automatically scraped articles from online news outlets using the newspaper[1] framework. Our data collection spans the period from the 23rd of February 2022, on the eve of the Russian full-scale attack on Ukraine, until the fourth of April, and we sample at eight time points during that period.

Our choice of media outlets and languages is based on the geopolitical zones which might have been affected by propaganda. We collected news from Ukrainian and Russian media outlets, choosing sources supporting pro-Kremlin narratives in Ukraine confirmed by journalistic investigations to directly copy news from Russian news outlets (UkraineWorld.org, 2022). We included American and British English-speaking outlets as a positive control of widely recognised high-quality news, as well as French news sources. We also added Romanian news as representative of the Central European block, which is one of the secondary targets of propaganda campaigns (Rosulek, 2019), and used websites that have been categorized by Rubrika.ro[2] as containing fake news. Except for English, all languages have two subsets, one supporting the Russian side of the conflict, and one supporting the Ukrainian

[1] https://newspaper.readthedocs.io/en/latest/

[2] https://rubrika.ro/extensie-browser

one. In total, we collected 18,229 texts: 8872 texts featuring pro-Western narratives and 9357 reflecting the official position of the Russian government. The sources are listed in Table 2.1.

Note that the ground-truth labels were assigned only according to the news source without manual labelling.

2.6.2 Telegram posts

Since the start of the war, many Telegram channels became widely used in Ukraine and Russia for quicker updates on the war and for posting uncensored reports (Bergengruen, 2022; O'Brien, 2022; Valeriya Safronova, 2022). However, it is a source without moderation and fact-checking, hence fake news and emotional lexicon, including profanity and propaganda, are not unusual (Sweney, 2022). Therefore, we included both Russian and Ukrainian Telegram news in our data collection.

2.6.3 Linguistic Feature Selection

We start processing the collected texts by extracting per-article linguistic features. The first set of features have been used previously in (Mahyoob et al., 2020) to detect fake news: a number of negations, adverbs, average sentence length, proper nouns, passive voice, quotes, conjunctions (we also count the frequency of the conjunction 'but' separately to capture contrasting), comparative and superlative adjectives, state verbs, personal pronouns, modal verbs, interrogatives.

Since fake news and propaganda can be associated with 'overly emotional' language (Asr & Taboada, 2019), we generate word counts for each basic emotion category: anger, fear, anticipation, trust, surprise, sadness, joy, disgust, and identify two sentiment classes, negative and positive, using the NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013). We translated each list of this lexicon from English to the other 4 languages, using automated translation and manual correction procedures. The translations are available on our GitHub[3]. We count the presence of each entry in lemmatized or tokenized texts.

Following Rashkin et al. (2017), we also extract the number of adjectives, the overall number of verbs and action verbs, as well as abstract nouns (e.g. 'truth', 'freedom'), money symbols, assertive words, and second person pronouns ('you'), and the first person singular ('I'). Inspired by the journalistic rules of conduct for neutrality[4], we count the number of occurrences of words from several dictionaries: survey, reporting words, discourse markers[5], reflecting the surface coherence of the article, words denoting claims(e.g. 'reckon', 'assume'), high modality words (e.g. 'obviously', 'certainly'). The dictionaries were created by synonym search and can be found in the form of lists in the feature extraction script.

By counting conjunctions, as syntactic features, we measure the number of subordinate clauses of concession (e.g. 'although', 'even if'), reason (e.g. 'because', 'as'), purpose (e.g. 'in order to'), condition (e.g. 'provided', 'if'), time (e.g. 'when', 'as soon as') and relative clauses, which reflect different ways of justification and argumentation.

All features are automatically extracted in a pipeline separately for each language using simplemma[6] and pymorphy2[7] for part-of-speech extraction in Ukrainian and spacy[8] for Russian, English, Romanian and French. The code is available on our GitHub repository.

[3] https://github.com/anonrep/pro-Kremlin_propaganda

[4] https://www.spj.org/ethicscode.asp

[5] http://connective-lex.info

[6] https://adrien.barbaresi.eu/blog/simple-multilingual-lemmatizer-python.html

[7] https://pymorphy2.readthedocs.io/en/stable/

[8] https://spacy.io

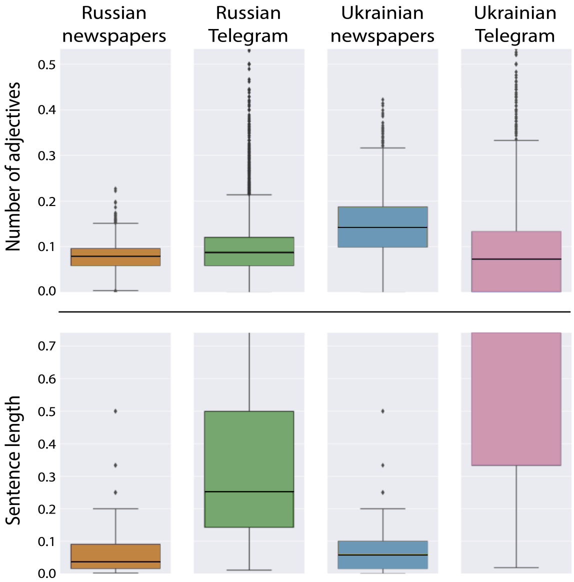


Figure 2.1: Examples of genre-related differences between newspapers and Telegram subsets. The boxplot represents 25% around the median, the whiskers show the first and last quartiles. Ukrainian news has more adjectives than Telegram posts, while this is vice-versa for Russian news. Sentences are longer in Telegram for both languages.

Table 2.2: Comparative results achieved on best folds.

Algorithm	Cohen's κ	F1	False Positive	False Negative
Experiment 1.1 Test on subset (1768 texts)				
Decision tree	0.49	0.73	16	450
Linear logistic model	0.58	0.79	113	265
SVM	0.75	0.87	156	151
MLP	0.64	0.80	103	229
BERT	0.84	0.92	97	42
Experiment 1.2 Test on French (20 texts)				
SVM	0.01	0.50	6	16
BERT	0.05	0.52	19	0
Experiment 2.1 Test on subset (1827 texts)				
SVM	0.75	0.88	120	151
BERT	0.86	0.93	111	12
Experiment 2.2 Test on Telegram (14525 texts)				
SVM	0.25	0.64	2013	3402
BERT	0.17	0.58	5770	212
Experiment 3. Test on subset (8709 texts)				
SVM	0.66	0.88	707	267
BERT	0.81	0.92	136	162

2.6.4 Keywords

As a list of keywords, we use the glossary[9] prepared by the National Security and Defense Council of Ukraine. It contains a list of names, terms and phrases recommended for use by public authorities and diplomatic missions of Ukraine as well as versions of these terms used in Russian political discourse. We translate this glossary to the target languages and add a short list from the military lexicon being avoided by Russian officials (e.g. 'war', 'victims', 'children', 'casualties') (Gessen, 2022).

2.7 Results

We evaluate the performance of our models using Cohen's κ (Cohen, 1960) and F-measure (Powers, 2008). While the F1-score is easy to interpret and most frequently used, subtracting the Expected Accuracy, Cohen's Kappa removes the intrinsic dissimilarities of different data sets, which makes two classification problems comparable, as K can compare the performances of two models on two different cases (Grandini et al., 2020). We also evaluate the number of false positives and negatives, which help build a complete picture of the model's performance. The results for all settings averaged over five models, can be found in Table 2.2. Details about the models and hyperparameters can be found in Appendix and GitHub depository.

Experiment 1. When training on Russian, Ukrainian, Romanian and English newspaper articles, the best result on the handcrafted linguistic features (no keywords) was achieved with an SVM: 0.87 F1-measure and 0.75 κ. The model is almost equally prone to false positives (108) and false negatives (120) across 1768 test samples (FP-rate: 0.06, FN-rate: 0.06). Linear models and a 2-layer neural network performed considerably worse (F1: 0.8). As the SVM performed best, we continued our experiments with this model and added our extracted keywords to the dataset, but found no improvement.

[9] https://www.rnbo.gov.ua/files/2021

The linear model using BERT embeddings achieves higher results than the handcrafted feature models (F1: 0.92, and κ: 0.84). While it produces a similar quantity of false positives as the SVM, the false negative rate decreases considerably.

When testing on 40 French texts (20 pro-Kremlin, 20 pro-Ukrainian), the performance drops considerably for the feature-based approach (F1: 0.5, κ: 0.01) with 14 false negatives and 6 false positives, and for BERT embeddings (F1: 0.52, κ: 0.05) with 19 false positives and only one true negative.

Experiment 2. The addition of French newspaper articles to the training set increased the F1-score by 0.08 for both SVM and embeddings-based models. However, the models do not perform well when tested on Telegram data. Without keywords, the SVM model scored 0.61 F1-measure, with a very low κ of 0.24, 2078 false positives and 3422 false negatives out of 14525 test samples. Adding keywords increases performance (F1: 0.62, κ: 0.25), lowering the false positive and false negative (FP-rate: 0.13, FN-rate: 0.23). The embeddings-based model scores even lower (F1: 0.58, κ: 0.17, FP-rate: 0.39, FN-rate: 0.014)

Experiment 3. Finally, we train on the full dataset with both newspaper articles and Telegram posts. The handcrafted feature-based model increases the F1-score to 0.88, but decreases κ to 0.66, with 707 false positives and 267 false negatives out of 8709 test samples. The embeddings-based model reaches 0.90 F1-measure and 0.81 κ, with 136 false positives and 162 false negatives.

Both models make disproportionately more errors when tasked with the classification of Romanian texts.

2.8 Additional Analysis

2.8.1 Feature Importance

We further analyse our best-performing SVM model to obtain feature importance for both linguistic features and keywords using the feature permutation method (Breiman, 2004). The analysis is illus-

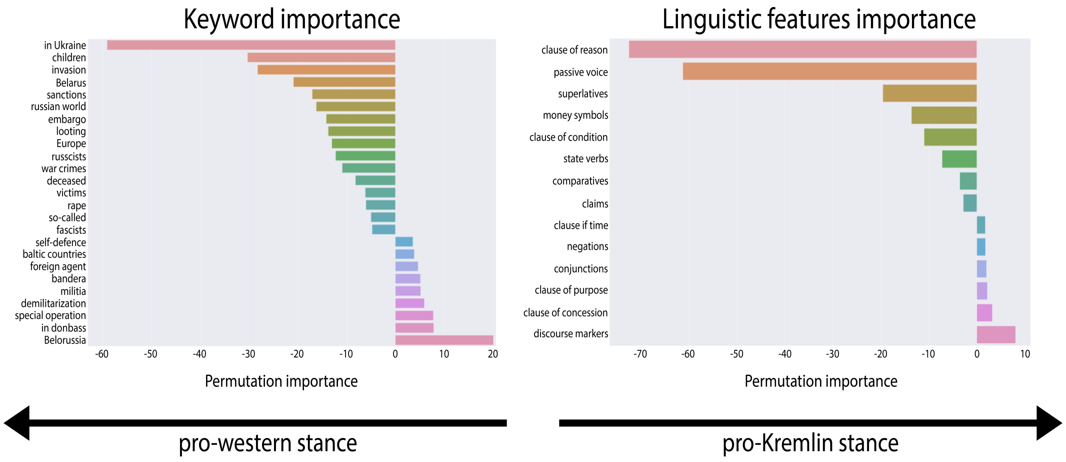


Figure 2.2: Permutation importance (drop of F1 score in %) for an SVM with linear kernel. Keyword importance is on the left side and the importance of linguistic features is on the right. Negative bars indicate features that are important for classifying a data point as pro-Western, while positive bars represent features indicative of pro-Kremlin propaganda.

trated in Figure 2.2. We find that various subordinate clauses prove to be important for the model, with the presence of the clause of the reason being the most indicative of pro-Western narratives, as well as passive voice. To a lesser degree, the following features were also deemed as important: superlatives, money symbols and words, clauses of condition, state verbs, comparatives and words indicating claims. For those features, it can be stated that they are unlikely to be found in pro-Kremlin news. At the same time, discourse markers (e.g. 'however', 'finally') as well as clauses of concession, clauses of purpose, conjunctions, negations and clauses of time separate pro-Kremlin news the best.

We find many keywords coming from the list provided by the Ukrainian Security Council glossary to be important. However, some of them need cultural and historical context to be understood. We find that the formulation 'in Ukraine' is the most reliable marker of pro-Western news, while in Russian news the formulation is 'on Ukraine', which indicates its use as 'on a territory' and not 'in the country'. Interestingly, the use of 'in Donbas' is the second highest indicator for Russian news. While it is a conventional name for the territory shared by two Ukrainian regions, it would preferably be used with the preposition 'on', e.g. 'on the Western Balkans'. The usage of 'in' gives linguistic legitimacy to the idea of the independence of the quasi-republics. The variant of the country's name 'Belarus' is highly indicative of the Western side, while 'Belorussia', the version found in Russian news, presents the neighbouring country's name rather as 'white' Russia, and not as 'Rus', the historical area in Eastern Europe. The formulation 'special operation'[10] is a euphemism for the word 'war' used by the Russian government and the pro-governmental news agencies. It is a strong indicator towards a pro-Kremlin narrative. On the Western side, we observe that the word 'invasion' has a higher frequency. Other words with high importance values for pro-Kremlin narratives are demilitarisation (of Ukraine), 'self-defence', 'militia', 'Bandera' [11], 'Baltic countries', also commonly called 'Pribaltika' in Russian, again presented more as a territory, and finally 'foreign agent'.

Many of the words we assumed would not be used in pro-governmental Russian articles were found to be important markers. Hence, words commonly used by the pro-Ukrainian news describing the disastrous consequences of war for both sides, e.g. 'children', 'looting', 'war crimes', 'deceased', 'victims', 'rape', 'sanctions', 'embargo', are attributed high importance in Western media. Some other curious keywords often occurring in Western media are 'Russian world' and 'russcists', which are mainly used by Ukrainian media as means of referring to the ideology of the Russian military expansionism (Gaufman, 2016).

2.8.2 Distributional exploratory analysis

Chronological analysis. We also carried out an exploratory study of the feature and keyword distributions over 5 data collections: the 23rd of February, the 1st of March, the eighth of March, the 18th of March and the fourth of April. We looked at the contrast between different languages and between pro-Kremlin and pro-Western media within one language, with the aim of explaining frequent model

[10] https://www.un.org/press/en/2022/sc14803.doc.htm

[11] Politician and theorist of the militant wing of the Organization of Ukrainian Nationalists in 20th-century (Marples, 2006).

errors and observing how media reflects the events of the war.

The most noticeable observation for Ukrainian pro-Western media is an increase in many features on the 18th of March, following the bombing of the Mariupol theatre on the 16th (Daniel Boffey & Borger, 2022): abstract nouns, claims, and negative emotional lexicon (namely words of surprise, disgust, sadness, fear and anger). Some indicators, like reporting words, negations, proper nouns, and modal verbs drop in frequency in March and seem to come back to the pre-war level in April. The use of the word war is constantly increasing throughout our data collection.

In contrast, in Russian pro-governmental media, the collection date with the most deviation from the overall average is the 1st of March when we can observe a drastic increase in the number of adjectives, average sentence length, assertive words, clauses of purpose, but also negative emotional lexicon (including words of trust and anger), and positive emotional lexicon. 1st of March corresponds to the end of the first week of the war when it became clear that the war might extend for a longer period (Harding, 2022).

British and American media remained quite stable throughout this time period, although we can observe an increase in superlatives on the fourth of April, which follows closely the discovery of the war crimes in Bucha, where 412 civilians were killed (Horton et al., 2022). pro-Western Romanian data also did not change considerably, with the exception of a slight increase with each collection in clauses of reason, words of surprise and the keyword ‘war’. At the same time, in the Romanian media flagged as fake news, there is a drop in words of anger and an increase in words of disgust, surprise, happiness and expectation, as well as abstract nouns, modal verbs, clauses of purpose when compared to the pre-war collection.

Language and narrative specific feature differences. When comparing media in different languages we observe interesting trends, which, however, did not account much for the decision of the classifiers. For instance, English and Ukrainian pro-Western media have the highest personal pronouns frequency, while newspapers from Russian media have the highest amount of quotations and are the only using keywords such as: ‘coup d’etat’, ‘DNR’, ‘LNR’, ‘Kiev regime’, ‘russophobia’. Romanian articles from trusted sources have the longest sentences, and all articles from Romanian media have the lowest use of the conjunction ‘but’. Furthermore, all articles have the highest occurrence of comparatives, superlatives and state verbs, which we believe is language-specific. This might be the reason for the low performance when applied to Romanian articles since these three features have high importance for the SVM model.

Articles from Ukrainian media generally have a high frequency of adjectives. At the same time, Ukrainian pro-Western news has the highest amount of emotional words (sadness, expectation, disgust, surprise, fear, anger), while pro-Russian Ukrainian articles do not show such a tendency, and thus might not reflect the same emotional atmosphere. It also has the same distribution of clauses of time as Russian pro-governmental news and the equally lowest usage of passive verbs.

Thus, we do not see a clear tendency for Russian propaganda to be homogeneous among the countries we selected. The only example would be the use of the keyword ‘Soros’, which is used uniquely in pro-Kremlin media in Ukraine and Russia, as well as in fake-news-flagged Romanian media. This can be explained by invocations of antisemitic conspiracy theories explored in Timothy Snyder’s ‘Road to Unfreedom’(Snyder, 2018) as manipulation strategies. Otherwise, media in Romania seems to be much more adapted to the local journalistic specifics, while in Ukraine the pro-Kremlin articles have much more in common with their origin.

2.9 Discussion

Our study is the first attempt to quantitatively and qualitatively analyse and automatically detect Russian propaganda, which does not necessarily use an emotional lexicon. Surprisingly, the propaganda style seemingly mimics international codes of conduct for journalism and adapts to each target language and it is country-specific. Many features of the Western news class can be found in the aforementioned related works for fake news detection, while pro-Kremlin features taken out of context could be interpreted as neutral journalistic style. This indicates that morpho-syntax modelling and analysis without semantics may be misleading. Both sides use propaganda. We found that their features differ, which may be explained by different ideologies and cultural specifics, but it may indicate different goals. Russian-backed media justifies the war and downplays adverse effects, while Ukrainian war propaganda focuses on various emotions, from ridiculing Russian forces, instigating hate against the Russian people as a whole or proposing an over-optimistic view of the military situation. In future work, we propose including an additional class representing neutral Ukrainian, Russian, and international news. However, labelling such datasets would require much more time for manual annotation. Since the appearance of state-of-the-art Transformer-based models, the trade-off between transparency and accuracy has been a topical issue in the NLP community. We show that transparent statistical models based on linguistic expert knowledge can still be competitive. Our best embeddings-based model has only around 0.04 F1-score advantage, but it is less explainable. We cannot control if the BERT model learnt the style of the media outlets instead of propaganda itself, while we can be sure that SVM indeed captures Kremlin's manipulative lexicon. While there are methods to interpret decisions (De Cao et al., 2020) of such models, we leave this for future work. As BERT models can capture syntax (Jawahar et al., 2019), we believe that such embeddings might still be less flexible towards changes in themes and topics, and need retraining if major vocabulary changes occur. The BERT-based model has a clear tendency towards false positives and performs slightly worse on the data from different distributions. In the context of automated content moderation, false positives would mean flagging/filtering a post or banning a person, limiting the freedom of speech. False negatives might lead to posts with propaganda reaching more targets.

Keywords were only beneficial for SVM when applied to new data, where the algorithm had to base its decisions on semantics more than morpho-syntax. The overall journalistic style captured by handcrafted features is more reliable, as the model performance does not drastically change for any of

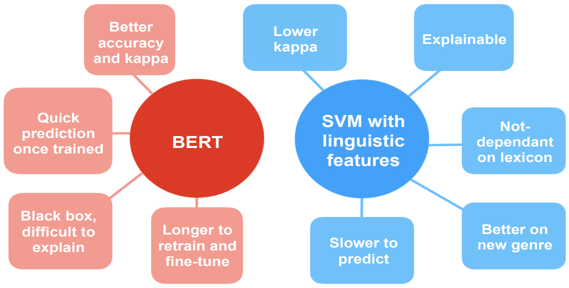

Better accuracy and kappa
Quick prediction once trained
BERT
Black box, difficult to explain
Longer to retrain and fine-tune
Lower kappa
Explainable
SVM with linguistic features
Not-dependant on lexicon
Slower to predict
Better on new genre

Figure 2.3: Advantages and disadvantages of the presented methods.

the languages in focus, even in the face of such events as war. Scalability is, however, a major drawback of feature-based models, as new predictions require first-feature extraction, while models using BERT embeddings can be used out of the box. However, BERT models have important token length limitations, whereas with SVM we pass a stable vector of feature counts normalised by the text length. While it might seem natural to choose these high-performing models in industrial settings, we believe that for the sake of transparency, models using handcrafted features that are competitive can still be used. The summarized comparison can be seen in Figure 2.3.
Both approaches can turn out to be inefficient after a certain period of time, especially in light of the new tendencies towards automatically generated news.

We see our work as a first step towards a browser extension flagging harmful content to raise individual awareness and assist us in filtering the content we consume. However, the classifier can be used to block pro-Western news as well, ensuring the impenetrability of echo chambers, amplifying the effects of propaganda instead of helping to fight it.

2.10 Conclusion

We presented two alternative methods to automatically identify pro-Kremlin propaganda in newspaper articles and Telegram posts. Our analysis indicates that there are strong similarities in terms of rhetoric strategies in the pro-Kremlin media in both Ukraine and Russia. While being relatively neutral according to surface structure, pro-Kremlin sources use artificially modified vocabulary to reshape important geopolitical notions. They also have, to a lesser degree, similarities with the Romanian news flagged as fake news, suggesting that propaganda may be adapted to each country and language in particular. Both Ukrainian and Russian sources lean towards strongly opinionated news, pointing towards the use of war propaganda in order to achieve strategic goals.

Russian, Romanian and Ukrainian languages are under-researched in terms of NLP tools in comparison to English. We hope that our study contributes to social media and individual efforts to moderate harmful content. We share our data and code as open-source tools for the detection of automated fake news or propaganda in order to help local human moderators and common users in those countries.

І жах не в тому, що щось зміниться, — жах у тому, що все може залишитися так само.
And the horror is not that something will change - the horror is that everything may remain the same.

Lina Kostenko

3

The Evolution of Pro-Kremlin Propaganda from a Machine Learning and Linguistics Perspective

Supervised by **Christoph Benzmüller** and **Tim Landgraf**.

3.1 Preface

This Chapter was previously published as: Solopova et al. (2023a). The Evolution of Pro-Kremlin Propaganda From a Machine Learning and Linguistics Perspective. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics. https://aclanthology.org/2023.unlp-1.5/

Language is a living, ever-changing social phenomenon that continues evolving (Markov et al., 2023; Bernabeu & Vogt, 2015). With its abundance of everyday interactions between millions of people, social media only accelerates this process (Almuttalibi, 2023). While the language of conventional news, having its standards and editorial books, is more resistant to change, this does not apply to social media news channels, like those on Telegram and Reddit. In addition, "Force Majeure," such as wars and cataclysms, still can impact the news language in unpredictable ways. From the machine learning perspective, this raises a question of how long the models trained on social media data can accurately predict the desired phenomena as time passes, especially as we speak of the data discussion of such a significant and quickly evolving event as war. Currently, even one of the largest language models on the current market, GPT4, is criticised for performance deterioration over time Chen et al. (2023). In this Chapter, we investigate the resilience of propaganda detection models over time. In the previous chapter, we trained our models on the data from the beginning of the Russian full-scale invasion of Ukraine in February-April 2022. Here we compare their performance on the newly collected data 1 year later. We could show that our models were resilient against propaganda evolution throughout the first year of wartime. We also present a method of the SVM model's error analysis, plotting the linguistic feature distributions over the True Positive, True Negative, False Positive and False Negative prediction categories. We use this to analyse the systemic linguistic changes in propaganda leading to errors that manifested in similar distribution predominantly among True Positive and False Positive categories, but also True Negative and False Negative ones. We also create a simplified post-hoc interpretability method “Reverse ablation study”, which is called to interpret the transformer model’s outputs. It is much more stable and computationally less expensive than the classic algorithm, which analyses the attention layer, although it may not account for different non-sequential combinations of words that may have an impact on the prediction. Using this approach we shed light on many similarities between the morpho-syntactic categories important for SVM and the BERT models, showing weak and exploitable sides of the BERT model. In addition, although we reconfirm that the BERT model learns morpho-syntactical information, which has already been shown in the literature (Jawahar et al., 2019; Pérez-Mayos et al., 2021) our results indicate that it may be deemed less important than morphological one for the propaganda prediction task. Finally, using statistical testing, we compare the linguistic variations between the 2022 and the 2023 data sets, detecting several interesting trends that reflect how different societies and media react and adapt to the events of the current war.

3.2 Contributions

Conceptualization, methodology, formal analysis and investigation, visualisation, writing - original draft preparation: [Veronika Solopova];
Writing - review and editing: [Veronika Solopova, Christoph Benzmüller];
Supervision: [Tim Landgraf, Christoph Benzmüller].

I was fully responsible for the conceptualisation, and data collection, while I also carried out all of the evaluation experiments. In this study, I also created a new model-agnostic interpretability method for Language Model predictions and implemented it for the BERT model. Finally, I wrote the first draft, and edited it, with special help from my supervisor, Christoph Benzmüller, and I acted as a corresponding author, responsible for submission and reviewers’ comments implementation.

3.3 Abstract

In the Russo-Ukrainian war, propaganda is produced by Russian state-run news outlets for both international and domestic audiences. Its content and form evolve and change with time as the war continues. This constitutes a challenge to content moderation tools based on machine learning when the data used for training and the current news start to differ significantly. In this follow-up study, we evaluate our previous BERT and SVM models that classify pro-Kremlin propaganda from a pro-Western stance, trained on the data from news articles and telegram posts at the start of 2022, on the new 2023 subset. We examine both classifiers' errors and perform a comparative analysis of these subsets to investigate which changes in narratives provoke drops in performance.

3.4 Introduction and Related Work

Fake news has been shown to evolve over time (Adriani, 2019). A piece of news is often modified as it spreads online by malicious users who twist the original information (Guo et al., 2021), while an imperfect replication process by other users leads to further distortion (Zellers et al., 2019a). Guo et al. (2021) showed that the disinformation techniques, parts of speech, and keywords stayed consistent during the evolution process, while the text similarity and sentiment changed. Moreover, according to their scoring, the distance between the fake and evolved fake news was more prominent than between the truth and the initial fake news. The evolved ones sound more objective and cheerful and are more difficult to detect. Jang et al. (2018) also observed significant differences between real and fake news regarding evolution patterns. They found that fake news tweets underwent a more significant number of modifications over the spreading process.

In the case of fake news and disinformation originating in state-run outlets, we talk about propaganda. In this and previous studies, we focus on Russian propaganda (Kendall, 2014; Chee, 2017; Parlapiano & Lee, 2018). It has been shown that the Russian Presidential Administration exercises coordinated control over media advertising budgets and editorial content whilst maintaining an illusion of media freedom by letting a small number of minor independent media outlets operate (Lange-Ionatamišvili, 2015). Hence, the adaptations to the Kremlin's political agenda are an additional factor that contributes to how Russian fake news evolves. Modern Kremlin propaganda fundamentally appeals to former greatness, glorification of the Russian Empire, the victory in World War II, the Soviet Union's past and the narrative of 'Facing the West' (Khrebtan-Hörhager & Pyatovskaya, 2022). Looking at the key narratives between the beginning of 2022, and the start of 2023, after a year of unsuccessful assault we observe several shifts in the narrative. At the beginning of the war, the official goals and objectives were identified by obscure terms such as "denazification" and "demilitarization" of Ukraine. At the same time, a fight against the Neo-Nazis has become an established rhetoric of the highest officials. "American bio-labs in Ukraine", "8 years of genocide in Donbas" and the claim that the Ukrainian government is responsible for shelling its own cities (Korenyuk & Goodman, 2022; Opora, 2022) became the most frequent topics.

After almost one year, Russian officials now openly recognize shelling of civilian electric infrastructure (Kraemer, 2022; Luke Harding & Koshiw, 2022; Grynszpan, 2022; Ebel, 2022), while propaganda directed to the external audience becomes majorly blackmail threatening Western countries to prevent them from supplying Ukraine (Faulconbridge, 2022a). As for the internal audience, the main objective is to support mobilisation efforts in Russia (Romanenko, 2022).

In our initial study (Solopova et al., 2023b), we proposed two multilingual automated pro-Kremlin propaganda identification methods, based on the multilingual BERT model (Devlin et al., 2019) and Support Vector Machine trained with linguistic features and manipulative terms glossary. Considering the aforementioned transformations, we hypothesised that our models' performance should drop on the 2023 data. In this follow-up study, we measured how the models trained a year ago perform on current news from the same sources. We also analysed how their language changed according to our linguistic feature set.

Model	F1	Cohen's κ	FP%	FN%
SVM 2022 full test	0.88	0.66	8%	3%
SVM 2022 small	0.74	0.5	9.5%	16%
SVM 2023	0.85	0.71	9.5%	4%
BERT 2022 full test	0.92	0.81	2%	2%
BERT 2022 small	0.87	0.74	11%	1.4%
BERT 2023	0.93	0.87	5%	0.8%

Table 3.1: The Table shows the models' performance on 2022 and 2023 subsets.

In Section 3.5, describe the experimental setup and the new data set. We present our results in comparison to those from 2022 in Section 3.6. In Section 3.7 we carried out an error analysis of the SVM and BERT models. For the SVM we contrasted the linguistic feature distributions in the groups of errors. For the BERT model, we applied a simplified word importance approach to gain insight into vocabulary and morpho-syntactical categories. In Section 3.8, we compare the 2022 and the 2023 data sets to see how propaganda evolved overall in our given context. Finally, we discuss our key findings and draw a conclusion in Section 3.9.

3.5 Methods

3.5.1 Models

In our initial study, we implemented a binary classification using the Support Vector Machine model for input vectors consisting of 41 handcrafted linguistic features and 116 keywords (normalized by the length of the text in tokens). For comparison with learned features, we extracted embeddings using a multilingual BERT model (Devlin et al., 2019) and trained a linear model using these embeddings. In this study, we apply the models to the new data from the same sources to see how resistant such systems are to changes in the data provoked by the changing events of war and adaptations from the Kremlin's propaganda campaign. We evaluate the performance of our models using Cohen's κ (Cohen, 1960), F-measure (Powers, 2008), False Positive (FP) and False Negative (FN) rate.

3.5.2 Data

We automatically scraped articles from online news outlets in Russian, Ukrainian, Romanian, French and English language, attributing each source to either pro-Kremlin or pro-Western class. We assigned ground-truth labels without manual labelling, based on journalistic investigations, or, in the case of Romanian data, using proxy websites, which categorize outlets as those containing fake news. We filtered out the news on neutral topics.

For Russian and Ukrainian we also collected posts from Telegram news channels which are the most popular alternative to traditional media. For pro-Western channels, we used those recommended by the Ukrainian Center for Strategic Communications[1], while for the pro-Kremlin stance, we identified one of the biggest Russian channels with a pro-war narrative.

We had 8 data collections from the 23rd of February until the fourth of April, 2022. In 2023, we collected on the 9th of January. Although this particular day can be considered relatively peaceful in terms of war events, this collection contained news about the preceding incidents and overall political analysis.

We made sure to collect from the same sources as the last year. However, French RT was banned from broadcast in Europe. Instead, we scraped a francophone version of the Turkish Anadolu Agency, which evokes Russian versions of the events in its reports. We also completed RainTV with Meduza news in the Russian liberal subset, since at the moment Meduza is a source with the least dubious reputation, widely read by the liberal Russian community. In 2022, we trained the model with 18,229 out of 85k texts to balance out different languages and sources. In 2023, we collected 1400 texts overall. You can find the data and our code in our Github repository[2].

[1] https://spravdi.gov.ua

[2] https://github.com/anonrep/pro-Kremlin_propaganda

3.6 Results

The full test in 2022 corresponds to the performance on 8700 samples of the original test set, while the small is a random sampling of the original 2022 test set to correspond to the size of the 2023 set and make them comparable. Although we also took an average of 5 seeds, the perfect comparison is complicated since we cannot ensure a balanced representation of the test samples from 2022 and 2023 in their complexity. As shown in Table 3.1, both models stayed accurate on the task. The SVM model on the 2023 data slightly outperforms its small test results from 2022 and even the full test as per κ. It seems quite stable in its False Positive rate across the experiments but has a higher False Negative rate, especially seen in the 2022 small test results.

The BERT on the 2023 data outperformed both full and small 2022 tests in f1 and κ. On the 2023 data, there are considerably fewer False Negative, while it shows a slight tendency towards False Positives.

12 out of 12 news from liberal Russian outlets were labelled as propaganda by both SVM and the BERT. The SVM had difficulty with the Ukrainian Telegram, labelling 50% as propaganda. In terms of the Ukrainian outlets which in 2022 we considered as pro-Kremlin propaganda, in 'Newsua' both BERT and SVM found no propaganda, while in 'Strana.ua', almost 100% was found to be propaganda by both models.

3.7 Error analysis

SVM. Regarding the SVM model, some patterns can be observed by looking into the distributions between the True Positives, True Negatives, False Positive, and False Negative. Thus, the number of reports mentioned, positive sentiment, stative verbs and subordinate clauses used all indicate strong similarities in distribution between True Positives and False Positive. In the case of relative clauses, clauses of condition and time, there is a correlation between both True Positives-False Positive and also True Negatives-False Negative pairs. False Negatives also have the highest average sentence length. Finally, we observe the highest number of abstract nouns and adjectives in True Negatives and False Positive, which means it can be a very confusing category in 2023 data. Out of the keywords, the most confusing are 'Europe', 'Kremlin', 'invasion' and to a lesser degree 'Belarus'. For more information see Appendix A.3.1

BERT. We were inspired by the attribution method (Sundararajan et al., 2017b). It is based on integrated gradients and requires retraining of the initial model. This approach is also computationally expensive because it uses back-propagation to calculate word importance. We segmented texts, so that the first segment is the first token of the text, while every next segment will have another next word unmasked until the last segment becomes a full text again. We classify each of them.

$$text = w_0, w_0 + w_1, w_0 + w_1 + w_2... + w_n$$

If the new next word changed the prediction value and its probability, it was recovered into either the list of words inducing pro-Kremlin or pro-Western prediction, separately for 2022 and 2023. We analysed extracted lists with linguistic features extraction script to see if there are some similarities in how experts and BERT choose propaganda features.

Thus, the first finding is that BERT identifies the names of the sources appearing in the text and connects them to the prediction classes. For instance, 'ziua', the name of a Romanian tabloid is one of the most frequent words we extracted for Romanian words, which changes prediction into 'propaganda'. In contrast 'activenews', a neutral Romanian news outlet always changed prediction value into 'pro-Western stance'. Even more, in 2022 french data a link to Russian 'Ria' news also was accurately determinant for propaganda class. In 2023, the main word indicating propaganda in Russian news was 'main/head', for the French 'authority' and for the Romanian 'treaty'. In contrast, the main words for pro-Western prediction for the Russian were 'announce' and 'sovereign default'. In 2023, the main pro-Western words for Romanian are ‘sanctions’, ‘tribunal’ and ‘war’. In 2022, the word ‘war’ was actually a determinant for propaganda, while words describing punishment were not typical topics for Romanian media, they were, however, already present in Ukrainian one. It is possible that keywords BERT learnt in one language are projected to others in the multilingual model. In 2023 pro-Kremlin propaganda in Ukrainian news would focus on the word ‘Putin’ while predicting for pro-Western news are words ‘Ukraine’ and ‘Ukrainians’. In Ukrainian pro-Western news, words connected to national institutions such as ‘government’, ‘minister’, and ‘state’ are significant.

In the Russian language, a keyword most reliable for prediction of the liberal side is ‘orcs’, the way Ukrainians call Russian soldiers (while Russia is called ‘Mordor’ by the analogy of Tolkien’s Lord of the Rings).

By classifying the resulting words according to categories of linguistic features, we can see that many categories are matched. The most popular parts of speech are adjectives, abstract and proper nouns, and high-modality words. Many of them express either strongly negative or positive connotations. Similar to our initial study results, reporting words are highly predictive of the pro-Kremlin stance in the Russian language in 2022.

Syntactical features such as different types of clauses are present to a lesser degree. Hence, morphological information may be used more than syntactical one for predictions.

Some glossary keywords were also used by BERT’s model, e.g., ‘war’, ‘special operation’, ‘DNR’, ‘LNR’, ‘negotiations’, and ‘Kremlin’.

3.8 Comparative Analyses

We decided to look into the evolution of propaganda, by comparing the averages for each feature between 2022 and 2023 for each subset. We used z-score normalized averages. We could not use medians, which are a better choice, because the data is sparse, most of the medians equal 0, which complicates normalization and significance testing. We chose the Mann-Whitney U-test, as the events are not paired and are not normally distributed. See the comparison in Figure 3.1. The most substantial difference is seen for the keyword 'Kiev Regime", which became a lot more frequent in the Russian Telegram, where users also started discussing more negotiations and 'the west', making more claims, and using more assertive words, adverbs and other high-modality words. Russian state-run outlets on the other hand started using considerably less 'Special military operation' wording but also dropped the rhetoric of 'the Republic of Crimea', 'LNR' and 'DNR', which the Russian Federation annexed and considers its own regions, rather than independent republics. It also speaks less of negotiations, sanctions, genocide, fake news and Belorussia.

Russian Liberal news did not change its style and narrative, nor did English-speaking, French pro-Western and French pro-Kremlin news. Romanian pro-Kremlin data became less emotional. We can observe a drop in most negative and positive emotions, especially in 'trust'. There can be seen more abstract nouns and conditional clauses, which are more typical for the pro-Western narrative but also relative clauses and claims, which can usually be seen more in pro-Kremlin news. On the other hand, pro-Western Romanian media has much more negative sentiment than at the beginning of 2022, there is more anger and fear. They talk more about the deceased and the attacks, calling out the Kremlin more directly.

Ukrainian pro-Western news became more neutral, as negative and positive emotions calmed down, particularly trust. There is less mention of genocide, embargo, negotiations and sanctions, which were more important topics for 2022. A rise in the clause of time, adverbs and especially proper nouns is significant, reflecting mostly the discussion around armament supplies.

In Ukrainian Telegram, on the contrary, there is more anger, awaiting, and sadness. The high effect size for the keyword 'fake' reflects Ukrainian efforts to debunk Kremlin propaganda. Stylistically, the language possesses more adjectives, and subordinate clauses of reason, purpose and condition. The potentially pro-Kremlin news in Ukrainian, which seems to have partly changed their allegiance,

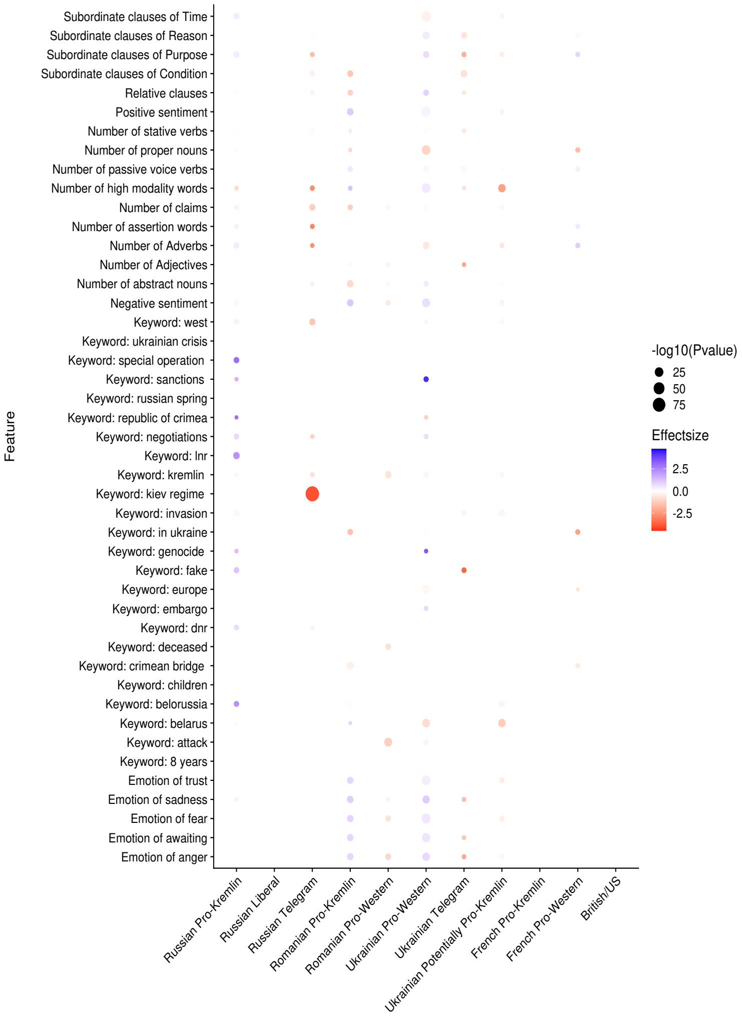


Figure 3.1: The dot plot shows the comparison between 2022 and 2023 subsets according to linguistic features. The dot size shows P-values while the colour shows the effect size. It represents the difference between the 2023 and 2022 averages, with red indicating growth in usage and blue meaning the drop.

shows more emotion of trust and fear, it is in general more expressive, with a higher number of adverbs. It uses the Russian manipulative 'Belorussia' term and 'Belarus' but leans more towards the latter. For comparing the languages see Appendix A.3.1.

3.9 Discussion and Conclusion

We applied an SVM with linguistic features and BERT multilingual model trained on the data from the beginning of 2022 to the new data from 2023. Since it is complicated to balance the complexity of the test sets, the true accuracy of the model lies anywhere between the full and the small 2022 test results, depending on how explicit the propaganda is. However, it is still possible to claim that both models successfully accurately identify a pro-Western stance.

Both classifiers are more prone to False Positives. As we showcased in the SVM model's error analysis, some distributions of significantly important features from our previous study, like abstract nouns and adjectives, are now similarly distributed between False Positives and True Positives.

At the same time, the BERT model is prone to attributing the class according to the news source name mentioned, which can lead to the model predicting everything describing or even debunking these outlets as propaganda. Overall, we observed that morphological information may be used more than syntactical one for predictions in BERT, while according to our initial study, a tendency towards some subordinate types distinguishes well the two stances. At the same time, the rise in temporal clauses in pro-Western stance, which in 2022 was highly significant for pro-Kremlin news may explain the higher miss-classification rate of the SVM.

The word 'war' appeared highly predictive for both SVM and BERT. Indeed, at the beginning of the war, this term was avoided by Kremlin officials and even made illegal in Russia (Troianovski & Safronova, 2022; Faulconbridge, 2022b). Hence, it would usually not appear in pro-Kremlin news that used euphemisms instead.

In the Romanian language, we can see how in 2022, in contrast to other languages, it was a determinant for propaganda, and now it is a determinant for pro-Western news. Consequently, some mistakes may be coming from such terms.

All liberal Russian 2023 news was identified as pro-Kremlin propaganda by both classifiers. However, they did not change their style since 2022, even though we added Meduza.

Meanwhile, Romanian pro-Kremlin sources in 2023 became more neutral. Similarly, in Ukrainian 'Newsua' which according to journalistic investigations was flagged as pro-Kremlin, in 2023 100% of articles were classified as pro-Western, by both models.

The evolution of war news gives us an insight into deeper-rooted differences between the sides of the conflict. The fact that in the Ukrainian language in 2023, in contrast to 2022, pro-Kremlin propaganda focuses on what Putin says, while real Ukrainian news almost does not mention him, but instead focuses on the Ukrainian government and Ukrainians themselves reflects how wartime societies evolve.

Overall, both models managed to draw good results on 2023 data, even considering how much topics and linguistic characteristics changed after one year of the war.

Limitations

The classical attribution method may be a more reliable explainability approach for BERT-like models than the one presented. We cannot be sure that these exact words and not them being present in combination with others, or even the length of the text is what changes prediction. In our future work, we want to expand on the explainability and transparency of our algorithms, add more languages and provide a web application interface. The comparability of the performance of the models on the 2022 and 2023 sets still leaves much to be desired. No cleaning nor filtring was performed over the scraped text which can contain irregular symbols left from the website meta-data. At the same time, collaboration with a fact-checking agency would also increase labelling quality.

Ethics Statement

It should be disclosed that the corresponding author is of Ukrainian nationality, although the study is not funded nor in any way affiliated with any governmental or private Ukrainian agency. Our work seeks to contribute to the automated content moderation efforts to protect human moderators from the constant psychological trauma they have to undergo reading toxic and manipulative posts and news. However, an imperfect automated tool may flag neutral content and should not be used to demonetize or ban internet users on social media.

Unfortunately, such technology can be used to reinforce eco-chambers if users choose to filter out everything that is, e.g. not pro-Kremlin propaganda. It can also help create tools which would be able to produce propaganda which will avoid these specific phenomena we describe, and thus make it more difficult to detect.

We also hope to support the general efforts to strengthen European security in the face of the Russian international propaganda campaign, by scaling defensive capacities and increasing citizens' awareness.

...if you give a man a fish, he is hungry again in an hour.

If you teach him to catch a fish, you do him a good turn.

Anne Isabella Thackeray Ritchie

4

Check News in One Click: NLP-Empowered Pro-Kremlin Propaganda Detection

Joint work with **Viktoriia Herman**.

Supervised by **Christoph Benzmüller** and **Tim Landgraf**.

4.1 Preface

This Chapter will be published as: Solopova et al. (2023a). **Veronika Solopova**, Viktoriia Herman, Christoph Benzmüller, and Tim Landgraf. 2024. Check News in One Click: NLP-Empowered Pro-Kremlin Propaganda Detection. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 44–51, St. Julians, Malta. https://aclanthology.org/2024.eacl-demo.6/

Based on a European External Action Service (EEAS) EUvsDisinfo project, Germany and Italy have been one of the main targets of the Kremlin propaganda campaign. In this database of fake media pieces accumulated since late 2015, German media holds the 1st place with 700 documented fakes, while Italy is third with 170. The German government recognises this problem: according to the German Ministry of the Interior and the Community the Russian Federation seeks to steer public opinion in Germany to its advantage by spreading disinformation and propaganda[1]. As German energetic system appeared vulnerable due to its dependency on Russian gas (Ting Lan & Zhou, 2022), Russian disinformation was resonating the voices promising German households to "freeze" (Delcker, 2022). Moreover, Russian disinformation was focused on demonising Ukrainian refugees (Morris & Oremus, 2022), placing the blame on the Ukrainian army for the potential occurrence of a nuclear disaster at the Zaporizhzhia NPP, narratives about neo-Nazis and ultranationalists in Ukraine (Smart, 2022), NATO forces provoking, according to these sources, the big Invasion. At the same time, Russian influence in Germany is also transmitted through growing links with both right- and left-wing populist parties (Meister, 2022). Meanwhile, Italian media is still actively providing a platform to Russian politicians in the interviews, spreading scepticism about Russian war crimes, and advocating dialogue and cooperation with Russia. According to The Propaganda Diary database (VoxCheck, 2020) fake

[1] https://www.bmi.bund.de/SharedDocs/schwerpunkte/EN/disinformation/disinformation-related-to-the-russian-war-of-aggression-against-ukraine.html

items about neo-Nazis and ultranationalists, Ukrainian authorities' alleged crimes against civilians in Donbas, and discrimination against the Russian population in Ukraine are among the most frequent topics. Most importantly, in both countries, two narratives are prevalent: anti-Americanism and calls against giving more weapons to Ukraine. In this Chapter, we train models based on the same architecture we presented in Chapter 2 for these two languages. Training experiments show us that the best performance is achieved with single-language models, not mixed, multilingual transformer models fine-tuned with Italian and German texts. We also determined that augmentation of the training set with translated news from other languages only helps the Italian model but introduces too much noise into the German one. Investigating the correlation between linguistic features and probability for the model to predict propaganda class, we observe several different tendencies in German and Italian compared to other languages we analysed in Chapter 2.

Another problem we tackle with this paper is the absence of tools available to a lay user that would allow them to check their news and receive detailed feedback on the manipulative patterns present. Hence, as a next step, we build a web application that allows users to check their news through our models. We introduce hybrid AI architecture, where SVM algorithms put constraints on class probabilities predicted by the BERT models and a rule-based determination chooses the final class, while linguistic indicators and keywords, which work as features for the SVM model, are shown to the user as supportive explanatory elements.

We perform a user study analysing user inputs and model outputs and also align them with questionnaire results. Although the results are optimistic, we identified that the users would prefer to use such a system as a browser extension and are more interested in checking social media comment sections and tweets than conventional news. According to our analysis, the linguistic indicators may have been better performing than the BERT predictions based on comparing user labels and the web app outputs, which points to their particular worth for such applications.

4.2 Contributions

Conceptualization, methodology, formal analysis and investigation, writing - original draft preparation: [Veronika Solopova];
Front-end development: and user test data collection: [Viktoriia Herman (50%), Veronika Solopova(50%)];
Writing - review and editing: [Veronika Solopova, Christoph Benzmüller];
Supervision: [Tim Landgraf, Christoph Benzmüller].

In this study, Viktoriia Herman was responsible for the front-end development of checknewsin1.click, website design and implementation of a database for user tests.
I was responsible for the collection and the curation of German and Italian data, while I also created experimental set-up, trained and evaluated models, and interpreted the corresponding SVM outputs, comparing them with the results for other languages from the previous two chapters. In the software development part of the study, I built the logic of the Hybrid AI system and was responsible for the implementation of this AI module into the productive system of the web application, namely the back end and its connection to the front end using the flask framework. I am still responsible for hosting the website on my personal server.
In the user tests part, I conceptualised the user study design and together with Viktoriia, we distributed the study over social media. Finally, I analysed and interpreted the user study results.

4.3 Abstract

Many European citizens become targets of the Kremlin propaganda campaigns, aiming to minimise public support for Ukraine, foster a climate of mistrust and disunity, and shape elections Meister (2022). To address this challenge, we developed "Check News in 1 Click", or CNOC, the first NLP-empowered pro-Kremlin propaganda detection application available in 7 languages. It provides the lay user with feedback on their news and explains manipulative linguistic features and keywords. We

conducted a user study, analysed user entries and models' behaviour paired with questionnaire answers, and investigated the advantages and disadvantages of the proposed interpretative solution.

4.4 Introduction

Evidence that we are living through a global crisis of trust in news is substantial, which inspired many a debate concerning the measures needed to rebuild it (Flew et al., 2020; Gaziano, 1988). An increasing number of people are getting their news online, particularly the younger generation. At the same time, many have started avoiding the news, first those concerning the COVID-19 pandemic and now those about the Russian war in Ukraine, majorly due to low credibility and negativity (Coster, 2022). At the same time, digital platforms are viewed more sceptically than traditional news, especially political ones, as they are believed to be agenda-driven and contain propaganda (Mont'Alverne et al., 2022; Flanagin & Metzger, 2000; Kalogeropoulos et al., 2019).

State-sponsored pro-Kremlin propaganda became a major issue, as reports claim that only a small per cent of Russian bots are being uncovered and detected (Menn, 2022). Geissler et al. (2023) showed that Twitter's (now X's) activity supporting Russia generated nearly 1 million likes, about 14.4 million followers and a substantial proportion of pro-Russian messages that went viral.

To address this issue, we created an accessible online user interface to check news in terms of pro-Kremlin propaganda, general manipulation and non-neutrality in 7 languages. It receives users' news and offers the model's verdict, its probability, as well as an explanation of manipulative keywords, linguistic strategies and indicators, shown to be associated with pro-Kremlin news. In addition to the models from our previous study, we trained new ones for Italian and German languages, exploring the usefulness of the data-augmentation strategy through translation, as well as multi-language versus language-specific pre-trained transformer models for this task. Here, we present our system architecture and the user study we conducted, quantifying user satisfaction and desirable features and analysing user entries and the outputs they received.

4.5 Related Work

Many tools have been developed to warn readers about fake news and "weaponize" them to understand the manipulative news better. The traditional fact-checking tools are curated by human moderators. For instance, the **Disinformation Index**[2] is a web-based tool that gives a real-time score to news outlets based on the probability of disinformation appearing in the source, while **Emergent.Info**[3] tracks and debunks rumours and conspiracy theories online, verifying the claims suggested by the users.

An increasing amount of tools are based on automated text analysis and classification, almost exclusively for English. **The Factual**[4] is rating the credibility of the news each day using the site's sourcing history, the author's track record, and the diversity of sources in a news article as key features. **ClaimBuster**[5] is an online tool for instant fact-checking, allowing users to check the veracity of their texts, by searching for a fact-checked claim similar to user's input. The **Fake News Graph Analyzer** characterises spreaders in large diffusion graphs (Bodaghi et al., 2021). The **Grover** (Zellers et al., 2019b) uses a fake news detection model, which takes on the language of specific publications to detect misinformation more accurately. **Bad News** (Roozenbeek et al., 2022; Basol et al., 2020) is a gamified platform intended to build user understanding of the techniques and tactics involved in disseminating disinformation. They show that attitudinal resistance against online misinformation through psychological inoculation may reduce cultural susceptibility to misinformation.

Considering propaganda detection as a specific case of disinformation, it became a popular NLP task,

[2] https://www.disinformationindex.com/

[3] http://www.emergent.info

[4] https://www.thefactual.com

[5] https://idir.uta.edu/claimbuster/

especially due to the 2016 US Presidential Campaign, Brexit and COVID-19, as well as appearance of several Shared Tasks, such as The SemEval 2020 Task 11(Da San Martino et al., 2020a), WANLP 2022 Shared Task on Propaganda Detection in Arabic (Alam et al., 2022). However, only a few projects develop comprehensive interfaces accessible to the wide public. **Proppy** (Barrón-Cedeño et al., 2019) was trained on known propaganda sources using a variety of stylistic features and is constantly monitoring news sources, clustering them into events, and organizing articles about them based on the probability of containing propaganda. **PROTECT** (Vorakitphan et al., 2022) and **Prta** (Da San Martino et al., 2020b) allow users to explore the articles, texts and URLs by highlighting the spans in which propaganda techniques occur through a dedicated interface.

In the case of pro-Kremlin propaganda detection, interface applications for lay user are limited, while none of them appear to be using AI solutions. **Hamilton 2.0**[6] is a real-time dashboard, created by the project of the Alliance for Securing Democracy, which aggregates analysis of the narratives and topics promoted by Russian, Chinese, and Iranian government officials state-funded and state-linked media accounts and news. **NewsGuard** [7] uses a team of journalists and experienced editors to produce reliable ratings and scores for news and information websites. To the best of our knowledge, no research-based open-source tools using AI to check potential Russian propaganda in a user's specific piece of news and in several languages are currently available.

4.6 Methods

4.6.1 Data

In addition to English, Russian, Ukrainian, French and Romanian, from our previous study, we chose to add German and Italian models to our tool. According to the European Union project EUvsDis-

[6] https://securingdemocracy.gmfus.org/hamilton-dashboard

[7] https://www.newsguardtech.com/special-reports/russian-disinformation-tracking-center/

info[8], “no other EU member has been subjected to such a powerful disinformation attack as Germany has been”. In its database of fake media pieces accumulated since late 2015, German media holds the 1st place, while Italy is in third.

We used fact-checked and attested pro-Kremlin propaganda articles from Propaganda Diary (VoxCheck, 2020) by VoxCheck. Around 5% was also added from the press of political parties associated with pro-Kremlin sympathy. This amounted to 963 articles. As an example of trust-worthy media, we used VoxCheck’s “white list” including sources such as ZDF, Der Welt, Frankfurter Allgemeine Zeitung, and Spiegel (676 altogether). As an augmentation set, we translated 537 neutral news with BBC and The Guardian translations from English to German using translators python API[9] and 565 RT.ru and Ria.news from Russian to German. Together native news set consists of **1639** texts, while the the augmented one is**2741**.

In Italian, we collected **2229** news from the Propaganda Diary, out of them 922 with attested Russian Propaganda and 1307 ones from the “white list” (e.g. Internazionale, La Repubblica, Corriere). We augmented the ‘propaganda’ class by 304 samples with translations from Russian to Italian of Sputniknews, resulting in **2533** texts.

4.6.2 Models

The models from our initial study included one multilingual SVM model trained on morpho-syntactic features and keywords from the glossary of manipulative terms of Russian propaganda curated by the National Security and Defence Council of Ukraine and a fine-tuned multilingual BERT model (Devlin et al., 2019). We followed a similar scheme for the German and Italian models. We first trained both Support Vector Machine (SVM) and BERT multilingual models for both languages together and augmented the data with translated articles. This approach only drew a 0.8 weighted F1-score for

[8] https://euvsdisinfo.eu

[9] https://pypi.org/project/translators/

SVM and a drastically low 0.51 for the BERT model. Training models separately increased performance in each language, except for the Italian SVM model (0.77 on average, and the highest score was 0.78.). The result for the German SVM increased to 0.87 in 5-fold cross-validation, and 0.9 on the best seed. We used the bert-base-german-cased model and dbmdz pre-trained bert-base-italian-cased model (see Appendix A.4.3 for parameters), both implemented through HuggingFace[10] framework. The German model scored 0.94 F1, and 0.99 auroc, with 0.88 mcc, while the Italian one scored 0.90 F1, 0.93 auroc and 0.8 mcc on the best fold, with averages across the folds being 0.88 F1, 0,96 auroc, 0.77 mcc.

We decided to revise our augmentation policies and excluded non-native data. Interestingly, results dropped for both SVM and BERT models in the case of the Italian language (0.73 F1 on average) and drastic to 0.72 mcc, 0.94 auroc and 0.86 F1 averaged over 3 folds, although translations in the training set only accounted for 12% of texts. In contrast, while translations were 40% of the augmented set, the German model's performance slightly increased without them, with SVM achieving 0.91 F1 best and 0.89 on average and the BERT model gaining up to 0.036 in mcc and 0.1 in F1 (see Table 4.1 for training results).

4.6.3 System description

The interface is a web app, written with Python Flask framework for the back-end, and HTML, CSS and JavaScript for the front-end. The proposed news is fetched from the input window. The code for the front- and back-end is available under MIT License in our GitHub[11].

First of all, the language is identified using langid.py (Lui & Baldwin, 2012). If the detected language is one of the languages we support the appropriate BERT model (language-specific for Italian and German and multi-language one for the rest of the languages) predicts the probability of propaganda in

[10] https://huggingface.co

[11] https://github.com/*/propaganda_website

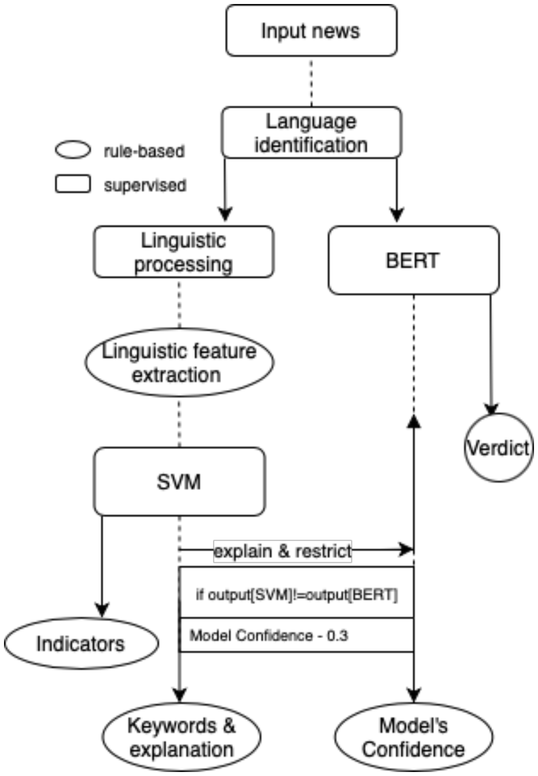


Figure 4.1: The figure illustrates the system's mock-up. The elliptical elements are rule-based reasoners while squared ones are trained models.

Table 4.1: Evaluation of the models used in the study. MCC and AUC results are not given for SVM-multi and BERT-multi as in the previous study Cohen's kappa was used instead. The numbers are rounded to 2 digits after the comma. de- stands for German model, it- for the Italian, w/tr - with augmentation through translation, w/otr- without.

Model	F1	MCC	AUC
SVM-de-it	0.82	0.64	0.82
BERT-de-it	0.01	0.51	0.51
SVM-de-w/tr	0.90	0.80	0.90
SVM-de-w/o-tr	0.92	0.83	0.93
BERT-de-w/tr	0.94	0.88	0.99
BERT-de-w/o-tr	0.95	0.92	0.99
SVM-it-w/tr	0.78	0.57	0.78
SVM-it-w/o-tr	0.75	0.49	0.74
BERT-it-w/tr	0.96	0.80	0.96
BERT-it-w/o-tr	0.94	0.73	0.93
SVM-multi	0.88	n/a	n/a
BERT-multi	0.92	n/a	n/a

the text. If the text is longer than 520 tokens, it is divided into several chunks. If at least one contains propaganda, the whole text is classified as such. If the language is not supported, the news is translated into English using Traslators API. The program saves both the verdict, 'Propaganda' or 'No propaganda', and the probability of the predicted class. In parallel, the linguistic feature extraction script, using Spacy[12] for lemmatisation and part-of-speech tagging, analyses the whole body of the news and passes the feature and keyword vector to the specific SVM model (Italian, German or multilingual). If SVM predicts an opposite class from the BERT model, we deduct 45% probability from the BERT's probability for the predicted class, and if the probability becomes lower than 30%, we change the prediction to the opposite one. The mock-up can be seen in Figure 4.1.

For each RBF-kernel SVM model, we also trained a linear one and looked into the coefficients of features and keywords and their association with a particular stance (Figure 4.2). The top features

[12] https://spacy.io

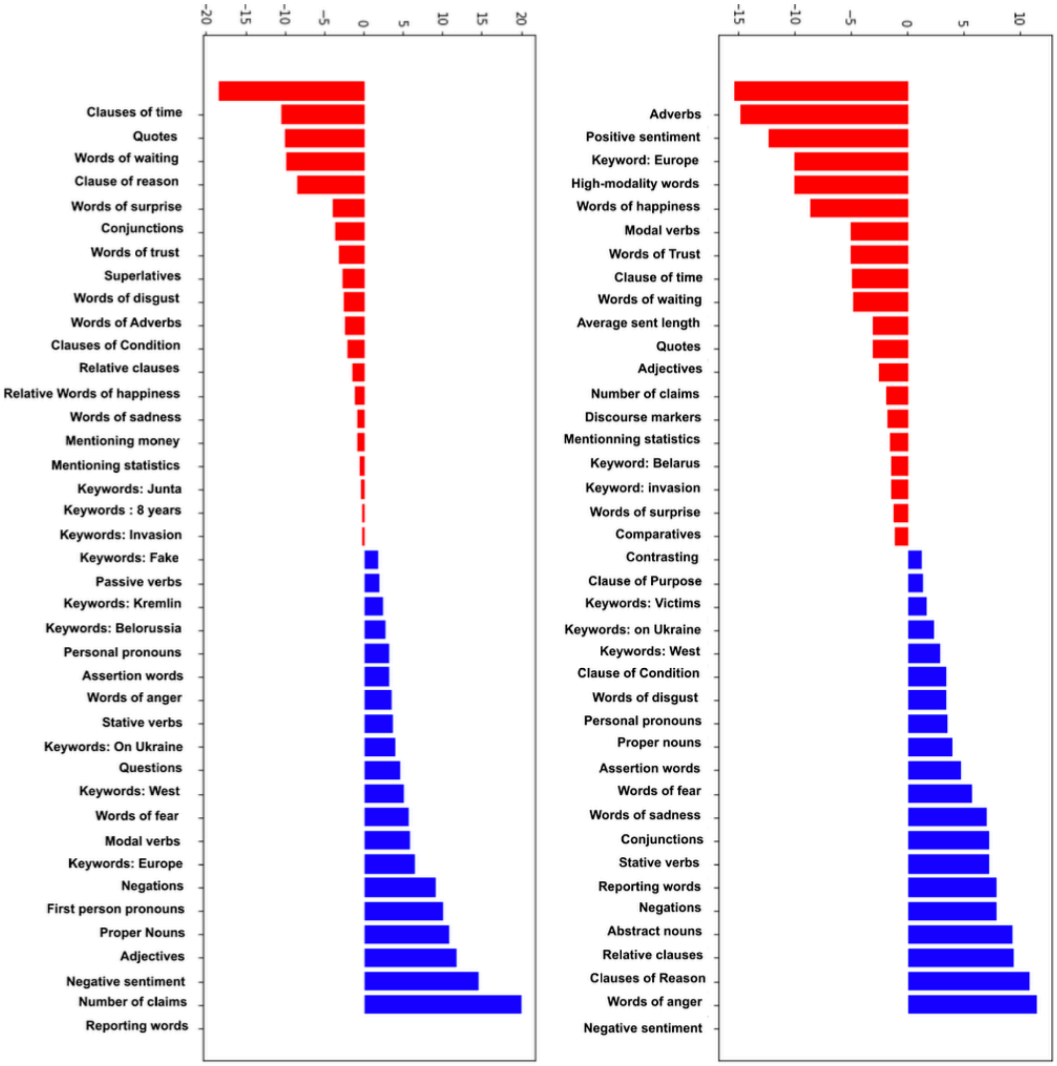


Figure 4.2: The figure illustrates the distribution of the learnt features according to the stance. The upper red side shows the features with the highest negative coefficients for "Pro-Kremlin propaganda" prediction (hence, more likely in Western, Pro-Ukrainian media), while the lower blue side shows the coefficients indicative of "Pro-Kremlin propaganda".

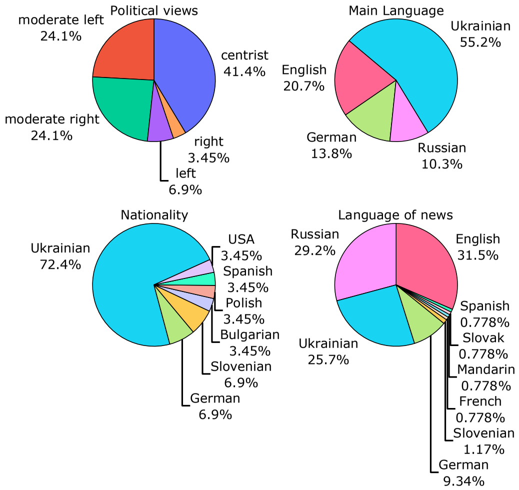


Figure 4.3: The figure illustrates statistics on the users who took part in the survey and used the application.

are then used as linguistic indicators and are shown to the user as warnings of potential manipulative, non-neutral language associated with the stances. Important keywords are presented separately with explanations from the Glossary of the National Security and Defence Council of Ukraine on a click. Comparing the important features and keywords, we discovered, that each language had its patterns of how Pro-Kremlin propaganda manifested itself, so we crafted indicators for each language separately (Appendix A.4.2). Some indicators, such as the abundance of negations, clause of purpose and reporting words, appeared to be universally indicative of pro-Russian propaganda in all of the languages we analysed. However, many features from our previous study, indicative of Pro-Kremlin propaganda were found more predictive of the Western stance in the two new languages. For example, frequent discourse markers, which are highly indicative of the pro-Kremlin side for other languages, are not associated with this prediction in German. The same stands for both German and Italian in terms of a high amount of quotes and clauses of time. In contrast, the clause of reason, highly predictive of a pro-Western stance for most languages, has the same tendency in Italian, but the opposite in German.

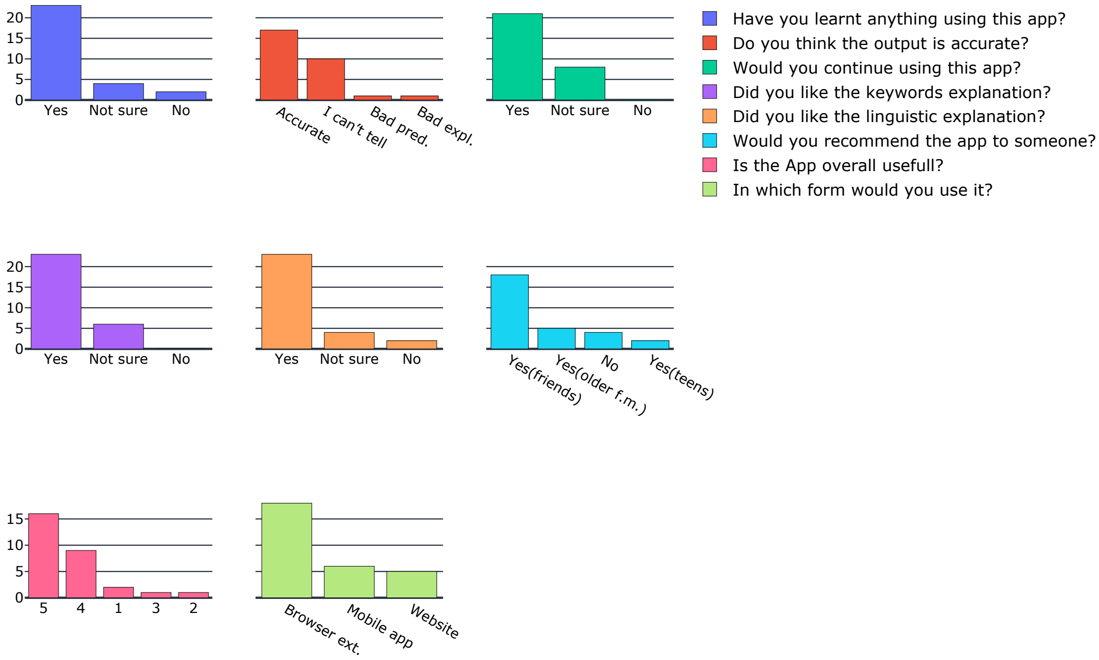


Figure 4.4: The figure shows the results of the user study questionnaire.

4.6.4 User study design

Users were asked to check at least three different news in the app [13] and fill out an integrated user questionnaire (Appendix A.4.1).

To understand the user profile we asked about the nationality, the language they searched in, their political stance, and how many pieces of news they verified. To quantify their experience, we asked their opinion about every element of the news analysis, its usefulness and accuracy, the preferable form (web application, desktop application, browser extension, chatbot), if they learnt something about propaganda and if they would continue using it, as well as the age group they would recommend this tool to (e.g. elder relatives, peers, teenagers, etc). From the back-end side, we collected the news the users entered, their own label ('propaganda' or not) and the analysis that the model provided.

[13] checknewsin1.click

The invitation to the user study was sent to various platforms on social media: several Italian, French and Ukrainian Facebook groups, subreddits r/EuropeanUnion, r/Samplesize, r/takemysurvey, r/YUROP,r/Ukraine, r/Ukraina; Dou.ua, a website for Ukrainian developers and IT workers, Instagram stories. The user study contained the consent form.

4.7 Results

191 users used the app with 257 unique requests, and only 29 out of them participated in the survey. 72% of the users in the survey are of Ukrainian origin, central Europeans (Polish, Bulgarian, Slovenian, Slovak) account for another 15%, 7% German, with one American and Spanish user. Ukrainian was named as the main language only 55% of the time though, while 20% searched news in English, 13% in German and 10% in Russian.

The full pull of users showed further language variety: almost 1/3 of all news entered into the app were actually in English, 1/3 in Russian and a slightly smaller percentage in Ukrainian. Apart from 10 entries detected in German, other languages included French, Spanish, Slovenian and Mandarin. As for the political views of the respondents, 41% self-identified as centrists, 24% as moderate right or left, while only 7% and 3.5% were left or right respectively (see Figure 4.3).

4.7.1 Survey results

As illustrated in Figure 4.4, the majority of users (86%) positively received the tool evaluating its usefulness as four or five on a scale of five, and only four respondents assessed the use as three and below. 79% responded that they learned something new while using the tool. The same per cent liked the keyword explanations and linguistic indicators, whereas 72% said that would continue using this app further. Only 58% of users said that they think the output of the models was accurate, while 34%

could not tell, and 2 users either considered the verdict or the explanation to be wrong. 63% would recommend the tool to their friends, 17% to older relatives and only 7% to teenagers, while 13% said they would not recommend it to anyone.

Talking about the potential formats for the tool, 62% chose that browser extension would be the most preferable form, while mobile application is also slightly more preferred than the website option as it is (20% against 17%).

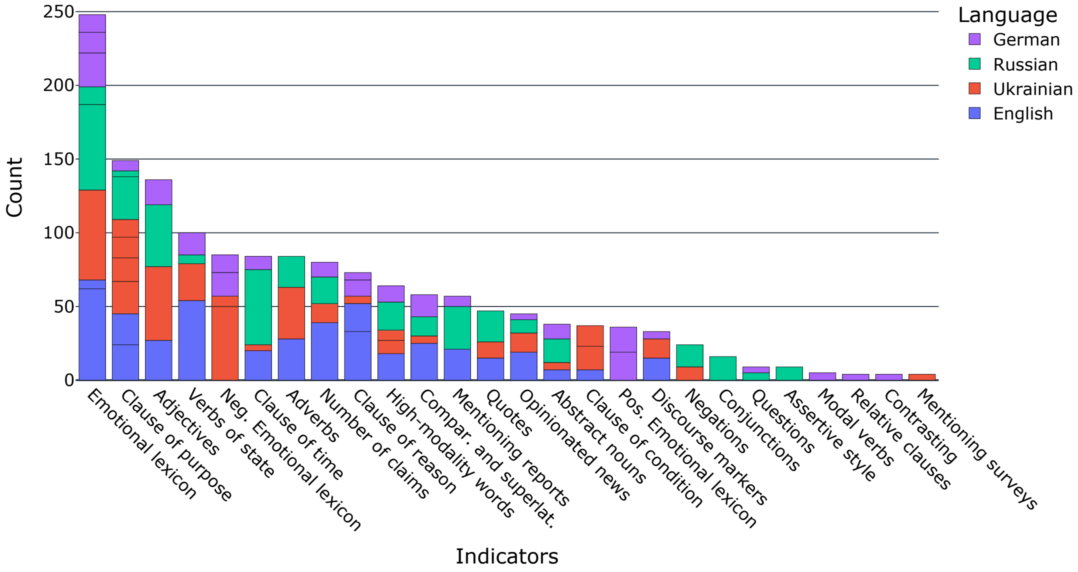


Figure 4.5: The figure shows the most recurrent indicators presented to the users for the 4 most frequent languages.

4.7.2 Keyword and Linguistic indicators proposed

In 21% of user inputs, no keywords were identified, while only 4 requests did not have any indicator proposed. There were on average 7.9 indicators found per request. Most of the indicators (53%) were of the general non-neutrality and manipulative nature, whereas pro-Kremlin indicators were identified slightly more often than “pro-western” ones (25/21%). The keywords (Figure 4.6) show

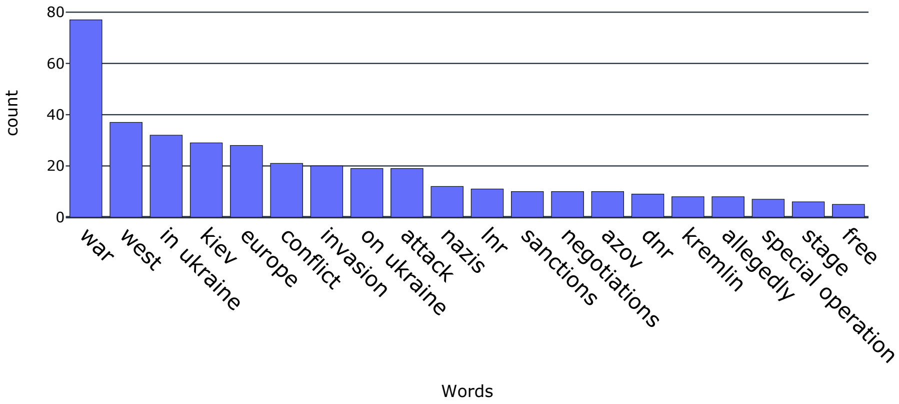


Figure 4.6: The figure illustrates the most present keywords identified and explained to the users.

that most news checked were political and focused on the Russian-Ukrainian war. Its synonyms and words related to war events were also in abundance: “conflict”, “attack”, “sanctions”, “negotiations”, “free/liberate”, “victims”, “war crimes”, and “arms deliveries”. The single most identified keyword was “war”, which would lean the model towards a ‘no Russian propaganda’ prediction. “West”, “in Ukraine” (in contrast to “on Ukraine”), “Kiev” (in contrast to preferred “Kyiv” transliteration) and “Europe” are the second most present pack.

West and Europe form juxtaposition with the term “Kremlin”. Many terms with direct correlation to Kremlin propaganda are present: “nazis”, “fascists”, “on Ukraine” (as on territory and not a country), “LNR”, “DNR”, “Azov”, “to stage”, “great power”, “Kiev regime”, “special operation” (official euphemism for war). “Allegedly” is curiously the only adverb present at the top of the list and is more associated with the Kremlin playbook, but can be present in any non-neutral low-quality journalism. The only adjective is “alarming”.

Speaking about the linguistic indicators shown to the users (see Figure 4.5), by far the most frequent is a warning on the presence of the emotional lexicon, which was identified in an overwhelming 95% of news. Other features associated with generally non-neutral news are a high number of adjectives and high-modality words, adverbs and quotes (except for German), abstract nouns, claims, comparatives and superlatives, personal pronouns of first person singular (opinionated news), and questions. Many features associated with pro-Russian propaganda in the previous studies were often detected: clause of purpose, conjunctions and assertive style (uniquely in Russian language), mentioning a lot of surveys in Ukrainian, and modal verbs for German. The abundance of temporal clauses and discourse connectors is indicative of a pro-Western stance in German but was considered a pro-Russian narrative marker in the 3 other languages in focus. Interestingly, the negative emotional lexicon was identified solely in Ukrainian and German languages, while the lexicon with a positive connotation in German only, which was shown to correlate with pro-western news in Section 4.6.3.

4.7.3 User and model label comparison

The multilingual BERT model showed an imbalanced prediction rate for different languages. The new German model had almost 50/50% positive/negative prediction rate, similar to the labels provided by the users. At the same time in Russian and Ukrainian language the verdict 'propaganda' was issued by the model only 8% of the time, while in English it was 28%. In contrast, the users labelled almost identical amounts of news as 'propaganda' and 'not propaganda' in English and Ukrainian, while in Russian 73% of submissions were claimed to contain it. Overall, only 21% of verdicts and user labels coincide in German, 36% in Russian, almost 50% in English and 52% in Ukrainian.

Diving deeper into the differences between the proposed and predicted labels, in German, there is an almost equal percentage of mismatch (41% model: 'No', user: 'Yes' and 37% model: 'Yes', user: 'No'). In other languages, the model is majorly predicting 'no propaganda'. In the case of Russian, e.g. the model did not predict 'propaganda' any single time when the user would say otherwise, with a similar result in Ukrainian (1.5%).

4.8 Discussion

Two major factors could explain the discrepancies between the user labels and the BERT predictions: either the user was wrong or the model, and here both tendencies seem to be present. If the model did not perform as well as on the testing set, the question is what was so different about the user entries that it could not generalise? We compared the distribution of the text lengths of the conventional news (which are rather large ∼407 tokens), Telegram news (which are rather short ∼32 tokens) in the training set and the news offered by the user (∼205). We could see that the latter distribution with all quartiles falls perfectly in between the 2 training set constituents (see Figure 4.7). Generally, the news can be even larger than the ones in our training set. For instance, the average article length of The New York Times is 622 words and 516 for The Washington Post (Menendez-Alarcon, 2012). A brief qualitative analysis shows that while many inputs are indeed news, they are also majorly Reddit comments, tweets, sentences taken out of one's mind, and even a few random words. We implemented the opportunity for the user to provide us with the link and not only copy-paste a text, which then we scrape using newspaper library[14]. Some inserted a link to Elon Musk's tweets, and while X cannot be scraped, the scraping library output was "This website does not use JavaScript", which was then erroneously analysed by the models. Such extreme instances as the last example only accounted, however, for a small percentage of the entries. On very long entries, the model did not once predict 'propaganda' and coincided in this prediction with the user. It had at least 15% better matching with user labels on very short samples, similar to Telegram posts in length, proving that length can indeed be a reason for some miss-classifications, when the user was correct. However, the length is only the surface description of the underlying genre missing from the training material: the users are not as interested in conventional news checking, as in flagging and quick discovery of bots and malicious actors in social media comments and tweets. A high number of Ukrainian participants and a high number

[14] https://newspaper.readthedocs.io

of certain responses concerning the tool's accuracy also showed that users predominantly were sure of their ability to recognize propaganda, but were interested in ways of quickly eliminating it from the informational eco-sphere. Late X's policies, made the development of such tools extremely problematic.

The indicators and keywords provided an important addition to the main model's verdict. Not only did the constraints we introduced on the main model help mitigate strong language-related biases, but also they appeared to be more reliable. They do not mismatch as often with user annotations. Only in 16% of cases where there were more pro-Russian propaganda features found and 8%, where no pro-Western features were reported at all, would the user consider it a 'no propaganda' sample. With the user label being 'Russian propaganda', there was only 12% with more pro-western than pro-Kremlin indicators identified, and 7% where no pro-Kremlin associated features were offered to the user. The strong performance of the indicators may have had a positive influence on the overall user evaluation of feedback's accuracy.

Another possibility, is also namely that users underestimate their knowledge of propaganda or are not very attentive when providing the label. While we received a lot of negative labels, the linguistic features indicate that most of the news pieces are not neutral. 37% of the news which were strongly not neutral were attributed to the 'no propaganda' label by the users. Only 6% of truly neutral entries were rightfully annotated as such, and 4% of them were called propaganda.

Overall, the results of the user survey, however limited in number, are positive. Both accuracy, recommendation, and interest in continuing to use the app are majorly high and both keywords and linguistic explanations were appreciated. In the free form, where we asked the users what they would like to change, some even suggested putting more stress on the explanations and taking away the overall verdict, but rather showing the percentage of propaganda present. This could be framed in future work as a regression problem, and additional models detecting such propaganda techniques as "red hearing" and "whataboutisms" should be implemented. Apart from minor front-end suggestions,

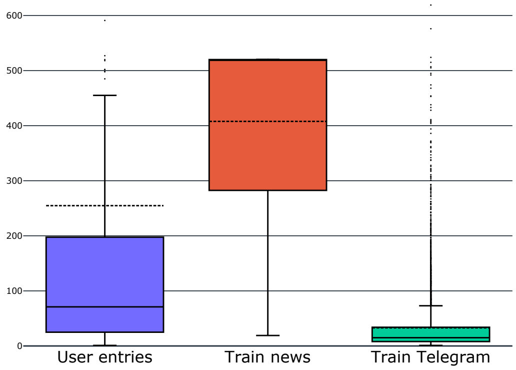


Figure 4.7: The figure illustrates differences in the text length between the training sub-corpora and the user inputs.

such as more visual support and instructions, some users were indicating that there was news with a pro-Western stance which were citing the President of Russia, which contained propaganda, and such cases may have to be dealt with separately. For the same reasons, the field of fact-checking is moving from the direct text-to-label classification towards more fine-grained and multi-featured info-sphere-based prediction Grover et al. (2022). The need to introduce many constraints for the main model through other models in our study is also a reflection of this trend. Including the layer user and human moderators in the research should become standard practice, as it helps better understand the needs of the community and tailor future solutions accordingly.

4.9 Conclusion

In this study, we trained pro-Kremlin propaganda detection models for German and Italian languages based on fact-checked news, pre-trained transformers and SVM empowered by linguistic indicators and keywords. We determined that language-specific BERT worked better than the multilingual model we used in the initial study and that augmentation using translated data is not a universally useful method to reach better performance, as it worked for Italian, but harmed German model performance.

Using these models we built a website with pro-Kremlin propaganda detection, explaining manipulative linguistic indicators and keywords to the lay user. We tested the web application with almost 200 users, 29 of whom took part in our user survey, positively evaluating the website. At every entry, we asked users to label the news they want to check, whether they think there is propaganda. The linguistic indicators and keywords showed fewer mismatches with the given labels, while the transformer-based model struggled, majorly because the entries were of unseen genre and length distributions. The users showed interest in detecting propaganda in more social media types of content: Reddit comments, and tweets.

Limitations

The BERT-family transformers in general seem to be sensitive to the length of the text being very different from the initial distribution. The old multilingual BERT model used in the study, in particular, was trained in April 2022 and over time and with the news topics evolving, it started to express language-specific biases. Several users reported that entering the same text translated into 2 different languages may give two different results. With the rise of the Large Language Model (LLM) of the new generation, the technique may be relatively out-of-data, and future research will undoubtedly focus in this direction. The demographics of our study, although include different nationalities, are still predominantly from Ukraine, and young adults (who are the usual users of the platforms we used to market the study), thus excluding younger and more senior groups. We were also not able to attract Romanian and Italian users, despite targeted marketing in their groups.

Ethics Statement

It is important to state that open-source propaganda research also provides malicious actors with the means to counteract automated tools and adapt the style so that it is even more difficult to detect in the future. We still claim that it is even more crucial to educate the wider public about the instruments to verify the news they consume.

You are what you tweet.

Germany Kent

5

Dissemination Mechanisms of Harmful Speech Online: Anatomy of Two Shitstorms

Joint work **Tatjana Scheffler**, Ruhr-Universität Bochum, Germany and **Mihaela Popa-Wyatt**, University of Manchester.

5.1 Preface

This chapter will be published as: Scheffler et al. (2022). Verbreitungsmechanismen schädigender Sprache im Netz: Anatomie zweier Shitstorms. Rupert Gaderer, Vanessa Grömmke (Hg.): Hass teilen. Tribunale und Affekte virtueller Streitwelten. Bielefeld: transcript 2024 (Virtuelle Lebenswelten 3).

The chapter is a translation of the original German version. http://arxiv.org/abs/2312.07194

One of the major limitations of AI approaches that use linguistic expert knowledge is the fact that features have to be determined for each specific task, which makes them less sustainable than modern statistical algorithms that learn relevant semantics on their own. However, it is still possible that several language phenomena, at least within the social media genre that we are looking at, can be efficiently described with the same linguistic variables. In this work, we show how the same linguistic indicators from Chapter 2, 3 and 4 can be successfully transferred onto the other type of social media modelling and analysis task: online scandal, or so-called *shitstorm* modelling. Using the same feature extraction algorithm, we were able to identify different phases in the evolution and development of the scandal and the emotional reactions of its participants while migrating from different platforms through the recruitment of social media activists. This shows that the scope of possible applications of the developed method goes beyond propaganda detection, and may be a generally useful tool to describe social media language particularities. Based on these features, two pilot models are presented: (1) predicting the stance of the user (whether the user supports or is against the person at the centre of the scandal) and (2) at which chronological period the shitstorm may be based on one message. We believe that if the amount of training data was increased, the models could achieve much better performance and could be useful tools for social media and public relations managers as well as social media language researchers, describing human swarm behaviour on the web and, eventually, human moderators on the platforms.

5.2 Contributions

The paper indicated equal contributions for Tatjana Scheffler and the author of this thesis. Conceptualization: [Veronika Solopova (33%), Scheffler Tatjana(33%), Mihaela Popa-Wyatt(33%)];
Methodology: [Veronika Solopova];
Formal analysis and investigation: [Veronika Solopova];
Writing - original draft preparation: [Tatjana Scheffler(50%), Veronika Solopova(50%)];
Writing - review and editing: [Tatjana Scheffler, Veronika Solopova, Mihaela Popa-Wyatt];
Project administration: [Tatjana Scheffler].
Tatjana Scheffler and Mihaela Popa-Wyatt were responsible for the conceptual idea to investigate shitstorm and their spreading on social media over time. As a philosopher of language, Mihaela Popa-Wyatt contributed to the Discussion and Introduction placing our study in the theory of hate propagation. Tatjana Scheffler administered our project and is the corresponding author, responsible for editing, and communication with the editorial board. As the actual paper is written in German, Tatjana Scheffler was also responsible for the translation of the Methodology and experimental part written by the author of this thesis into German language.
I was responsible for carrying out all of the experiments, while also participating in their planning and conceptualisation. I also proposed the shitstorms to focus on, collected and annotated the data with hate speech and stance labels, and visualised and interpreted the results. Although ML experiments are not central to this work, I also contributed by training the models for the tasks of stance detection and temporal classification of the message based on their content. I also wrote the first draft of the methodology and analysis sections from 5.5 to 5.8.

5.3 Abstract

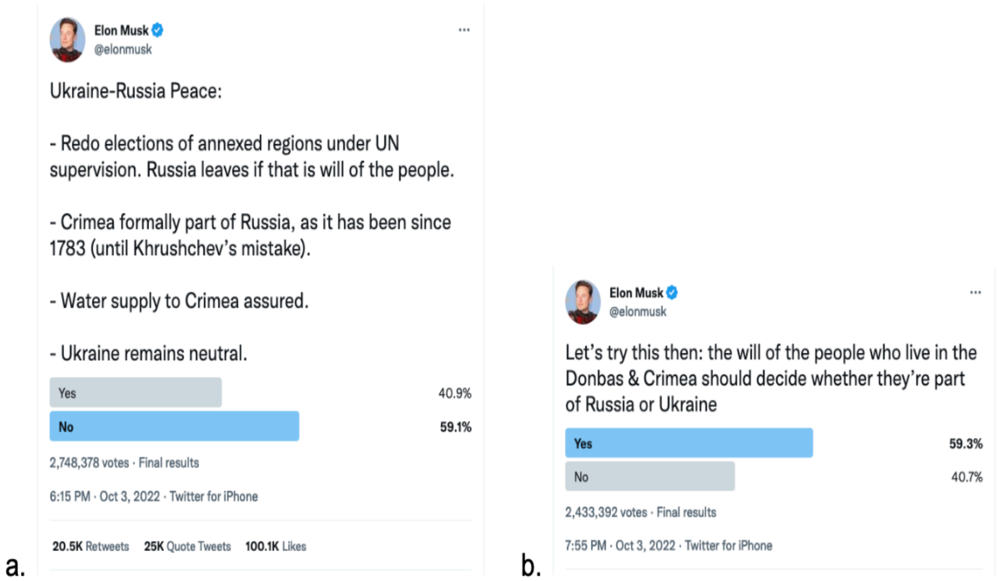
In this study, we analyse two cross-media shitstorms directed against well-known individuals from the business world. We examine the spread of the outrage wave across two media at a time and test the applicability of computational linguistic methods for analyzing its time course. Here, we focus on the distribution of linguistic features within the overall shitstorm and we ask whether one group of supporters and one of the opponents of the target have a distinguished linguistic form. We also look at the dynamics of different group participation and migration patterns between different platforms. Based on our analysis, we train two models predicting the phase of the shitstorm and the stance of the user based on the message.

5.4 Introduction

The shitstorm, "an unforeseen, short-lived wave of outrage in social media" (Gaderer, 2018), has only recently received consideration in linguistic research (Bauer et al., 2016; Bendel et al., 2016; Gaderer, 2018; Haarkötter, 2016; Himmelreich & Einwiller, 2015; Kuhlhüser, 2016; Stefanowitsch, 2020). In this working paper, we turn our attention to two exemplary, cross-media shitstorms directed against well-known individuals from the business world. Both have in common, first, the trigger, a controversial statement by the person who thereby becomes the target of the shitstorm, and second, the identity of this target as relatively privileged: cis-male, white, successful. We examine the spread of the outrage wave across two media at a time and test the applicability of computational linguistic methods for analyzing its time course. Assuming that harmful language spreads like a virus in digital space (Popa-Wyatt, 2022b,a), we are primarily interested in the events and constellations that lead to the use of harmful language, and whether and how a linguistic formation of "tribes" (Deremetz & Scheffler, 2020) occurs. Our research, therefore, focuses, first, on the distribution of linguistic features within the overall shitstorm: are individual words or phrases increasingly used after their introduction, and through which pathways do they spread. Second, we ask whether "tribes," for example, one group of supporters and one of the opponents of the target, have a distinguished linguistic form. We hypothesise that supporters remain equally active over time, while the dynamic "ripple" effect of the shitstorm is based on the varying participation of opponents.

5.5 Empirical Basis: Two Example Storms

We choose two recent shitstorms, which occurred in September and October 2022, as our research base. The first started with a Twitter poll questioning Elon Musk's support of Ukraine in Russia's war of aggression; the second with a tweeted video of the CEO of an eSports platform showing him celebrating with well-known misogynist Andrew Tate. All the data in the platforms involved was automatically scraped using Python: Telegram with the Python library Telephon, Reddit with the Praw library, and Twitter with Tweepy using academic developer credentials.



Elon Musk
@elonmusk

Ukraine-Russia Peace:

- Redo elections of annexed regions under UN supervision. Russia leaves if that is will of the people.

- Crimea formally part of Russia, as it has been since 1783 (until Khrushchev's mistake).

- Water supply to Crimea assured.

- Ukraine remains neutral.

Yes	40.9%
No	59.1%

2,748,378 votes · Final results

6:15 PM · Oct 3, 2022 · Twitter for iPhone

20.5K Retweets 25K Quote Tweets 100.1K Likes

Elon Musk
@elonmusk

Let's try this then: the will of the people who live in the Donbas & Crimea should decide whether they're part of Russia or Ukraine

Yes	59.3%
No	40.7%

2,433,392 votes · Final results

7:55 PM · Oct 3, 2022 · Twitter for iPhone

Figure 5.1: Tweets from Elon Musk about the war in Ukraine.

5.5.1 Elon Musk

The shitstorm began on Oct. 3 when Elon Musk tweeted a poll along with his proposals to end the war in Ukraine (Fig. 5.1.a), and later that day (Fig. 5.1.b). Later that same day, the President of Ukraine posted a tweet with his poll, “Which @elonmusk do you like more?”. One who supports Ukraine/one who supports Russia,” with 78.8% of 2.4 million voting for the first option. Another important tweet, which was eventually deleted, appeared on October 6. Within the Ukrainian community, the scandal spread to Telegram channels. We have evaluated one of them, namely the user thread on the blog of Ukrainian MP Alexey Goncharenko, who also commented on the situation several times from October 3 to 6.

Especially interesting are the messages in the Telegram channel, in which Ukrainian readers are asked to participate in the polls and to fend off Russian bots: “Once again, Elon Musk presents himself as an expert on geopolitics on Twitter. I’m already replying to him, join in too,” it reads on October 6. This invitation is responsible for the spike in participation that day (see Figure 5.2). In total, 413 tweets and 539 messages were collected from Reddit.

Comparing the platforms, we found that Telegram activity was much more short-lived, but more intense (Fig. 5.2). The last Telegram post on this topic was found on October 7: “About the Musk. Sci-

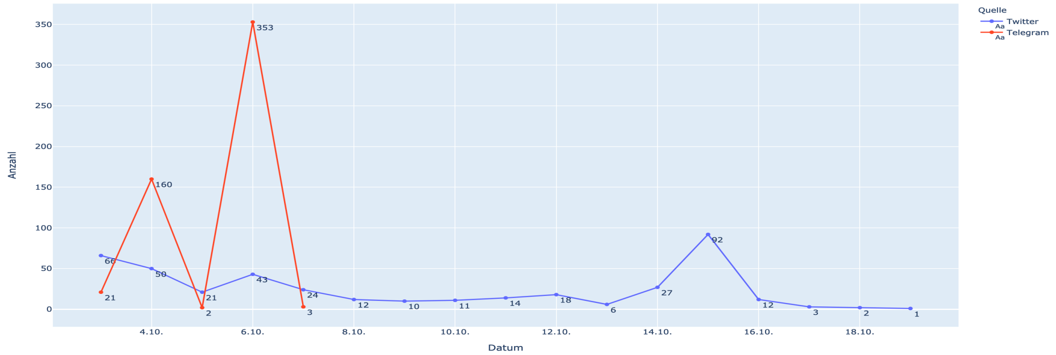


Figure 5.2: Timeline of the Elon Musk shitstorm on Twitter and Telegram.

entists have solved the mystery of Mona Lisa's smile. She's just a fool." On Twitter, the debate continued throughout October, ending mostly on October 19, with Elon Musk's last tweet on October 15: "The hell with it … even though Starlink is still losing money & other companies are getting billions of taxpayer $, we'll just keep funding Ukraine govt for free," which drew a lot of grateful Ukrainian tweets. Only 8.8% of Twitter users* participate in these threads more than once, and the replies are mainly responses to the original tweet rather than to any other reply. On Telegram, 28% of users participate more than once and 13% more than twice. The hashtags used mostly position themselves not pro Elon Musk, but against the Ukrainian community, such as directly "#standwithrussia" or "#americanpropaganda" and contextually such as "#MelnykSeiStill", "#MelnykShutUp" addressing Andrij Melnyk, the Ukrainian ex-ambassador to Germany, known for his active political involvement in the German debate on the supply of heavy weapons. In the Telegram data, we found no evidence of the use of hashtags, as this is neither typical for the Ukrainian community nor for Telegram as a platform in general (Scheffler et al., 2021). To characterize the content of the discourses, we determined the most frequent words over time. The keywords "Ukraine," "Russia," "peace," "war," and "Elon" are at the core of the conversation; they are frequent from the first to the last day we documented, with "Russia" occurring more frequently than any other keyword and much more frequently than the keyword "Ukraine" on October 13 and 15. The word "support" remains only until October 10, while "stop" in the sense of "stop funding this corrupt country" occurs only until the day Musk announces that Starlink will continue to operate in Ukraine, and "want" in the sense of "explain what Elon Musk wants" begins the next day of the debate and continues throughout. The remaining words stand for subordinate issues. For example, "Zelensky" was discussed from Oct. 4-9 after his tweet, "thank" shows the gratitude of the Ukrainian community; then they talk about how expensive the Starlink operation is in Ukraine, "Tesla" and "Crimea" are discussed only later (see Appendix A). Comparing Twitter and Telegram activities, there are a number of differences, starting with emojis. In general, emojis are not as present in either dataset. However, we can see a more political or supportive bias in the Twitter data and mocking, and laughing emojis are more common in Telegram (Figure 5.3). In

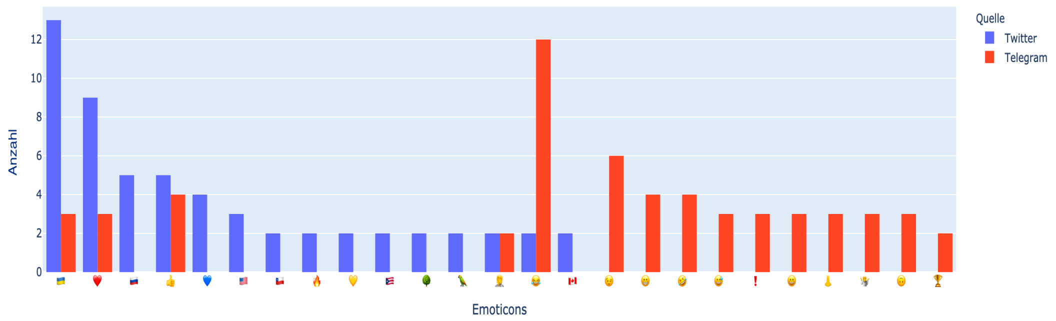


Figure 5.3: Use of emojis in Elon Musk's shitstorm, Twitter versus Telegram.

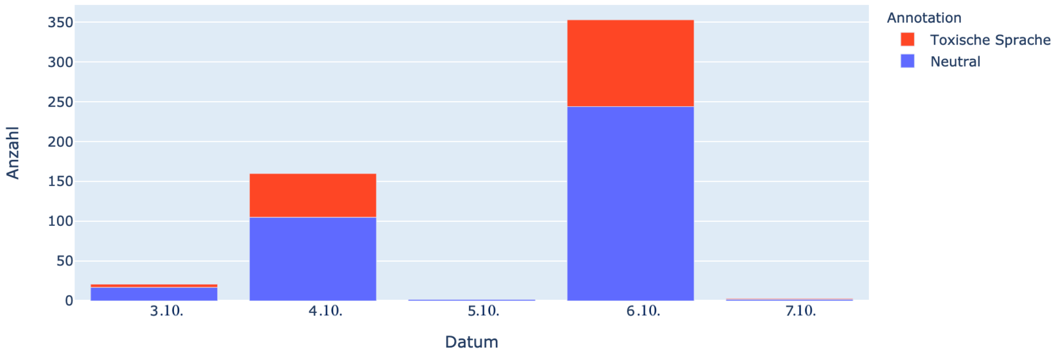


Figure 5.4: Distribution of toxic language on Telegram in Elon Musk's shitstorm (toxic language was automatically detected).

terms of offensive language, the conversation in Telegram is much more prone to hurtful language (31% of posts) than Twitter (5%), see Figs. 5.4,5.5.

5.5.2 eSport

The second shitstorm began on September 17, when G2 eSports CEO Carlos "Ocelote" posted a video on Twitter celebrating with Andrew Tate. What may have been short-lived outrage turned into

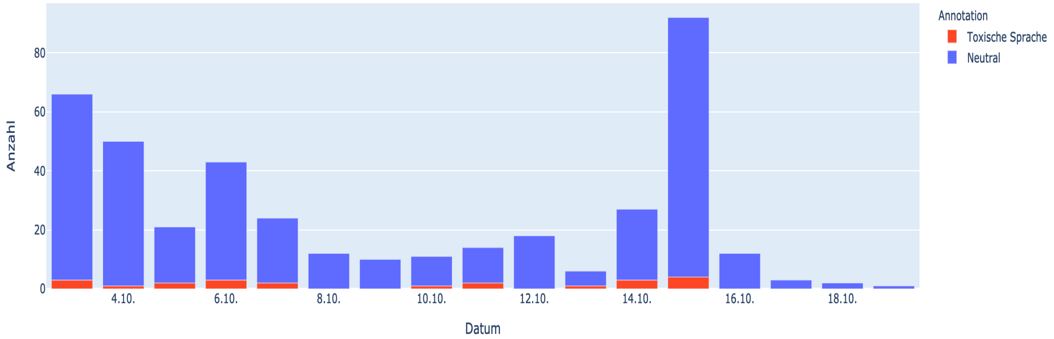


Figure 5.5: Distribution of toxic language on Twitter in Elon Musk's shitstorm.

a shitstorm when Ocelote reiterated his support for Tate, tweeting, "No one will ever be able to control my friendships, I draw the line here, I party with who I want." Later the next day, although a PR statement was published on his behalf, at the same time he liked tweets that supported his innocence. This contradiction was received very negatively by the community. That evening, it was announced that Ocelote would take eight weeks of unpaid leave as CEO, but this did not satisfy the public. The situation escalated until September 23, when it was announced that the organization had lost a very important franchise opportunity in a new game due to the situation. Most recently, Ocelote was forced to leave the CEO position for good and even sell his shares. Similar to the Musk shitstorm, here on September 18 and 20, users* from Twitter came to another platform (here an eSports subreddit) to inform others about new tweets. However, they did not directly invite participation in the shitstorm.

While we don't have user information on Reddit, 7% again participated more than once in this Twitter scandal. Emojis are not very present in the Reddit data, as they were only used 13 times in total. The most commonly used Twitter emoji is the clown, which again shows the general attitude of the community towards the target (Fig. 5.7). Hashtags are also not typical for this Reddit thread: Only two were used to poke fun at the use of "#MeToo," and the creative example "#TOPG2," which combines Andrew Tate's tagline and the name of the G2 team. On Twitter, hashtags were used a bit

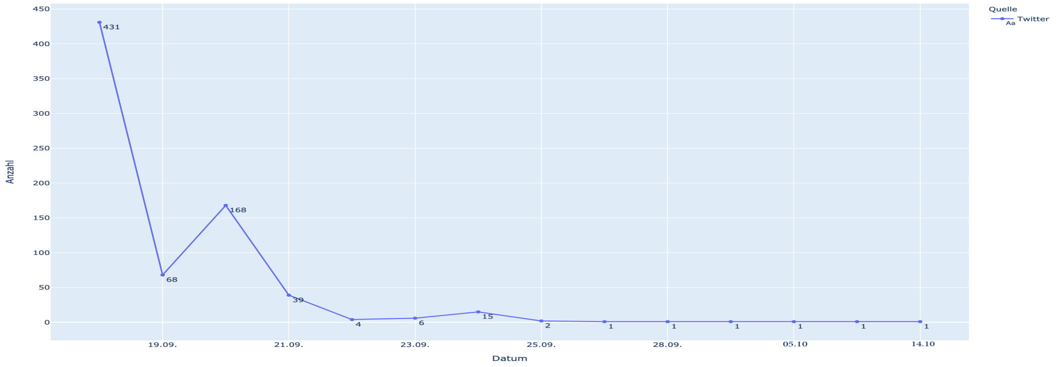


Figure 5.6: Timeline of the eSports shitstorm on Twitter. Unfortunately, the exact timestamps of the Reddit posts could not be retrieved.

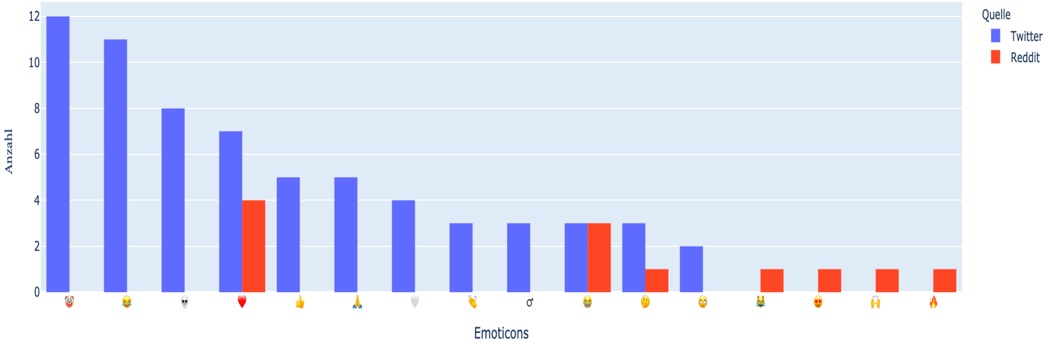


Figure 5.7: Use of emojis in eSports shitstorm, Twitter vs. Reddit.

more frequently: 2 times “#Respect”, 2 times “#G2ARMY”, “#AndrewTate” and some unexpected examples such as “#StandWithUkraine”, “#RespawnRecruits”, “#EINS” and “#ABetterABK” referring to an association of Activision Blizzard Kind employees. Twitter language overall in this discourse is twice as toxic as in the Elon Musk thread, with 10.4% of harmful language. Reddit, the second platform that is geared more towards the eSports community, also has more harmful language with 22% of all messages. This can also be seen by the fact that “fuck” is one of the most common keywords (Figure 5.8). In terms of keywords (Appendix B), the main keywords on both social media platforms reflect the main topic: Carlos, Tate, g2, and people. On Twitter, users* later also discuss what is the

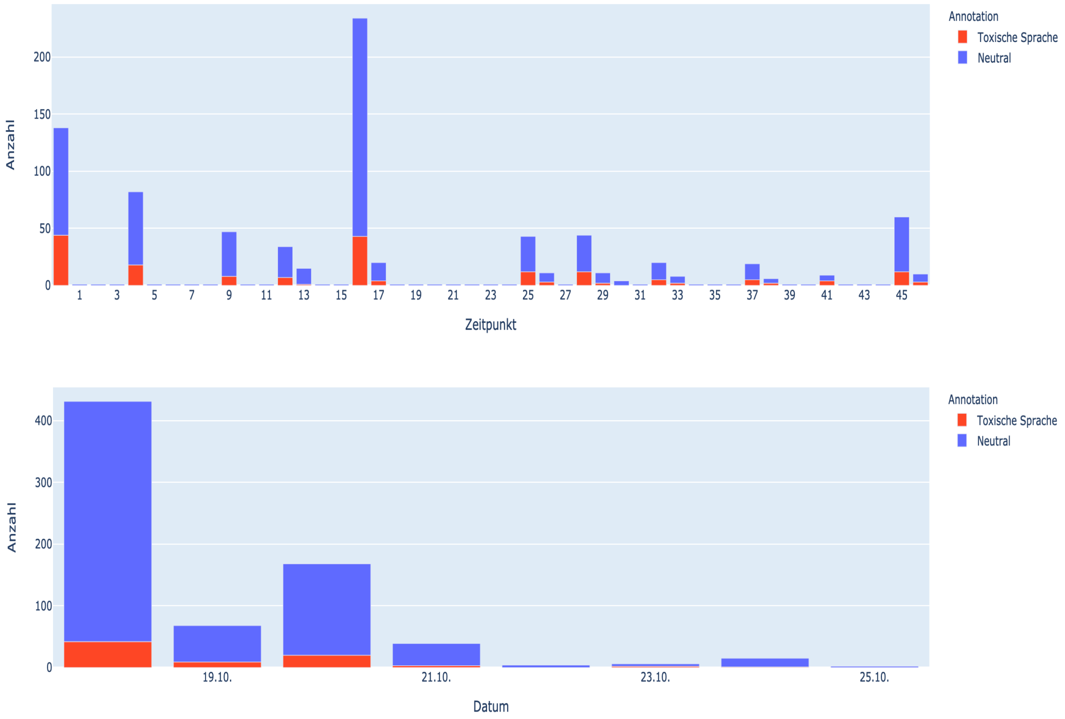


Figure 5.8: Distribution of toxic language on Reddit (above) and Twitter (below) in the eSports shitstorm.

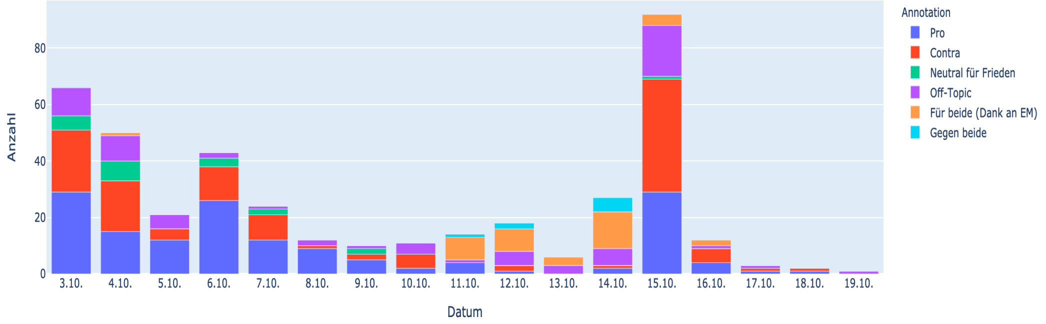


Figure 5.9: Distribution of the different groups over time in the Elon Musk shitstorm on Twitter.

most important aspect of the scandal: the party itself or the tweet in which the CEO reiterates his support for Andrew Tate. The discussion on Reddit includes many other subtopics: how the image of the organization was damaged, how the female part of the fan base feels in this context, etc. At the same time, many participants express gratitude and apology, criticizing a "cancel culture" that takes away a person's entire life because of one mistake.

5.6 Groups: Supporters and opponents in the shitstorm

We analyze the individual posts to find supporters and opponents.

5.6.1 Elon Musk Twitter

. In Fig. 5.9 we can see that at the beginning from 3 to 9.10, there is a neutral peace stance, which disappears in the middle of the scandal. In contrast, on 11.10, the group "for both Ukraine and Elon Musk" appears, represented mainly by Ukrainians who thank Musk for funding the Starlink program in Ukraine. At the same time, the number of tweets against both sides (Musk and Ukraine) is growing, which could be due to Russian bots. One of the examples of tweets supporting this theory is the following one, posted in English but clearly showing Russian syntax and containing a pro-Kremlin narrative with anti-Semitic conspiracy theory and Slavic supremacy: "Here the only problem is that the person in charge of Ukraine is not Slavic, many are not, they are Jews with the face of capos who have revived Nazism, and now carry the eschatic, it is not Nazi, but Force Ukraine. Jews who are now Nazis". We can also see that at the beginning there were more supporters of Elon Musk (blue in Fig. 5.9), while together with the grateful tweets after October 10 the share of people who were against him increased considerably, indicating the influx of Ukrainians from Telegram to Twitter on 10.10, triggered by Telegram posts.

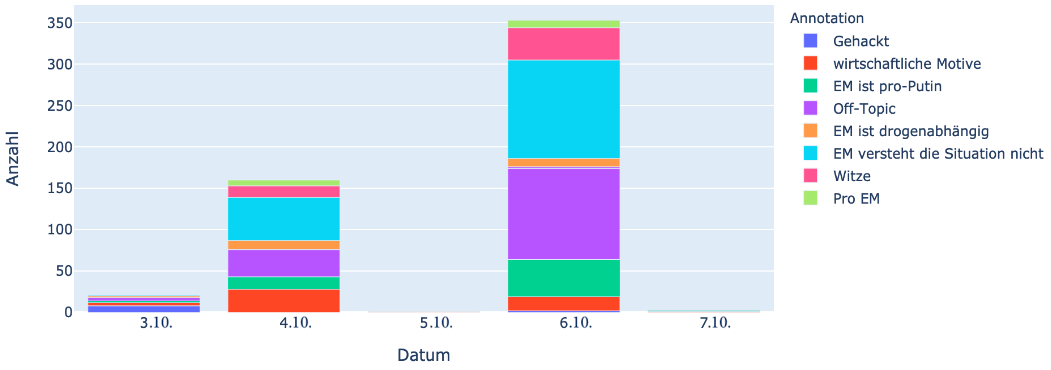


Figure 5.10: Distribution of the different groups over time in the Elon Musk shitstorm on Telegram. The first 7 are all against Elon Musk but have different explanations for his behaviour. The last group supports him because he helps Ukraine.

5.6.2 eSports Twitter

The Twitter scandal in the eSports community shows much less sympathy for the target. Initially, only 26% are on his side, but later, after unsuccessful damage management and possibly more people learning about it through Reddit posts, the number drops to 10%. Interestingly, the group that is against both sides grows slightly in percentage. This is related to the fact that the CEO publicly apologized, which looked like betrayal and weakness to this group of his supporters, and they wanted him to "not to bow to the mob." If you look at the two sides in the eSports community on Twitter, the opposition to G2's CEO uses a lot more personal pronouns and conjunctions and a bit more adverbs. It is also the only group that expresses disgust, while the only features that are slightly more relevant to supporters are abstract nouns and the emotional vocabulary of trust. The remaining features only show differences between tweets attributable to one of the two sides. Exact timestamps are missing on Reddit, but we observe similar spikes as on Twitter on the first day and 9/20. The main groups here are supporters of Ocelote and his opponents, while the "against both sides" group is responsible for less than 1% of the posts. After the initial, almost unanimously negative reaction, Reddit users* are initially much more sympathetic to Ocelote, until this changes and there are more opponents*, even though he was fired. A peculiarity of this subreddit is also that people often discussed many other things (in green) and not just this situation, just like on Telegram. Closed, issue-oriented communities (like on Telegram or Reddit) seem to be much less prone to shitstorms, as these communities have many other issues, whereas on Twitter it is like things that very different participants* focus on a specific, controversial message. However, the two groups seem to use more similar linguistic expressions

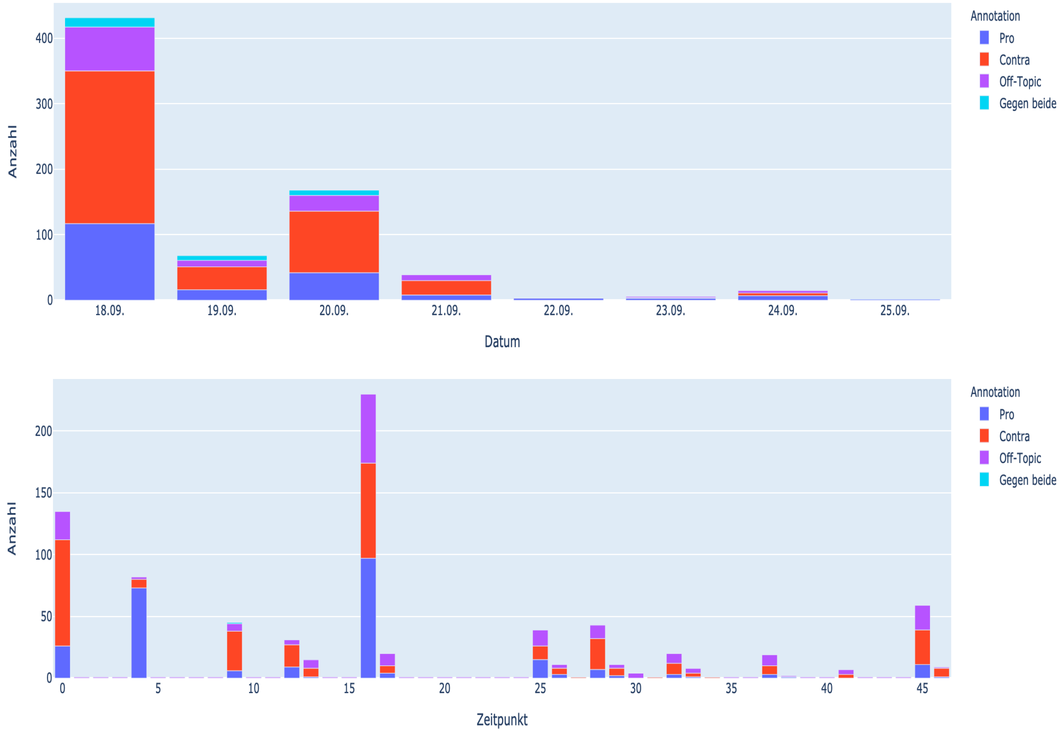


Figure 5.11: Distribution of different groups over time in the eSports shitstorm on Twitter (top) and Reddit (bottom). Reddit is given in points in time.

on Reddit. Supporters again use slightly more adverbs and conjunctions, as well as condition words, subordinate clauses, abhorrence, and interestingly, the conjunction “but.” The proponent group uses more personal pronouns here.

5.7 Chronological analysis

Looking at the distribution of message frequency over time, we can see a similar pattern between the two shitstorms. A high intensity at the beginning drops off shortly after and is followed by two smaller spikes each. We looked at several linguistic indicators to further examine how the language within the shitstorm evolves. We divided each corpus into 4-time spans, with the beginning being the first peak, the second part being the second peak, the third part being the data between the second and third peaks, and the last peak being the end. From each message, we extract the following linguistic features that are useful in determining opposing sides in a debate (Solopova et al., 2023b):

1. Morphological features: Number of adverbs, adjectives, verbs, proper nouns, conjunctions, negations, number of comparatives, superlatives, personal pronouns, passive forms;
2. Syntactic features: contrasting use of "but," number of subordinate clauses of concession, reason, and purpose; relative, temporal, and conditional clauses.
3. Punctuation features: Quotation marks, question marks;
4. Superficial semantics: abstract nouns, modal verbs, state verbs, assertions, high modality words, and counts consistent with NRC Emotion Lexicon emotions, such as fear, surprise, anger, expectation, disgust, happiness, sadness, trust, negative and positive emotions, words expressing an assertion or opinion. Average sentence length.

5.7.1 Chronology eSports shitstorm.

Reddit. In the first and last parts of the storm more adverbs, adjectives, personal pronouns, abstract nouns, subjunctive verbs, and conjunctions are present. Purpose expressions and 1st person singular personal pronouns are used almost exclusively in the aforementioned sections, while relative clauses occur almost exclusively in the last period, similar to the emotion of disgust. The emotion of trust increases slightly in the last period, while the emotion of surprise is almost absent at the beginning of the scandal, but is present as it progresses.

Twitter. In the last period, the average sentence length is significantly shorter, while the use of personal pronouns is significantly higher. We see a similar tendency toward "quadratic function" for abstract nouns and state verbs. Assertions are made only in the first period, where the highest use of conjunctions is also observed (which then appear only to a lesser extent in the third period). Justification sentences are again found only in the first period and especially frequently in the last period of the scandal, as are feelings of happiness. In general, positive emotions are most prevalent in the first third and last periods, while they are completely absent in the second period. We can also see an increase in the use of proper nouns in the third period, where people mainly discuss the lost sponsors and the hypocrisy of the reaction of Riot (the big game company), which itself has a lot of sexual harassment allegations, but imposed such a harsh punishment on the CEO of G2. We can also see that the number of adverbs used slowly decreases from day 1 to the last day.

5.7.2 Chronology Elon Musk shitstorm.

Twitter. A similar first decreasing and then increasing trend as above can be observed for various features in the Elon Musk Twitter subset for the emotion of surprise, fear, negations and abstract nouns. The emotion of anger and the subordinate clauses of reason are also present only at the beginning and end of the scandal. Sadness and surprise are significantly present only at the beginning of the shitstorm, while assertions are mainly made in the first and second periods. Finally, the emotion of anticipation grows from the beginning to the second period, where it increases and then gradually decreases towards the end.

Telegram. The third time frame is very significant for Telegram interaction (here October 7), with a slightly higher number of adjectives, conjunctions, and many more emotions such as expectation, anger, disgust, surprise, high proportions of modality words and fear. However, anger and fear are completely absent at the beginning of the storm; the lowest proportion of subjunctive verbs is also found there. Subordinate clauses of reason are present only in the second period and with the highest number of verbs of state. The falling-increasing tendency is present in proper nouns, but it is reversed in verbs of the state. The number of personal pronouns increases from the beginning to the last period, where it is highest. The average sentence length is much higher than on Twitter. Elon Musk's first tweet received 59 comments at the time of data collection. In the highest period, it reached about 90 replies. At the same time, the Telegram community initially assumed a much lower response, reaching 150 and 350 responses in the peaks, but then apparently quickly moved away from the topic altogether.

5.8 Automatic linguistic differentiation

Based on the chronological and group analysis presented above, we used the data in a pilot experiment to train two models that could help in the quantitative analysis of shitstorms in the future. We used the Huggingface multilingual BERT model trained on more than 100 languages (Devlin et al., 2019) and a classification model based on simpletransformers. First, we classified messages according to the time in which they appear in the shitstorm. We distinguished three classes: intense beginning, middle to last peak, and end. We obtained a classification accuracy of F1=67%. This shows that the phases of the shitstorm differ linguistically. With more fine-tuning and larger training data, this can make predicting the phase the shitstorm is currently in, and thus how far it is from the end, a very achievable reality. Second, we classified the position of the contributor into three classes: Supporter, Opponent, and Neutral to the target of the shitstorm. We obtained a lower result on this task, with F1=62%. Looking at the different groups in the above analysis, we think this low performance is largely because we grouped a lot of heterogeneous subclasses in the "for" and "against" classes.

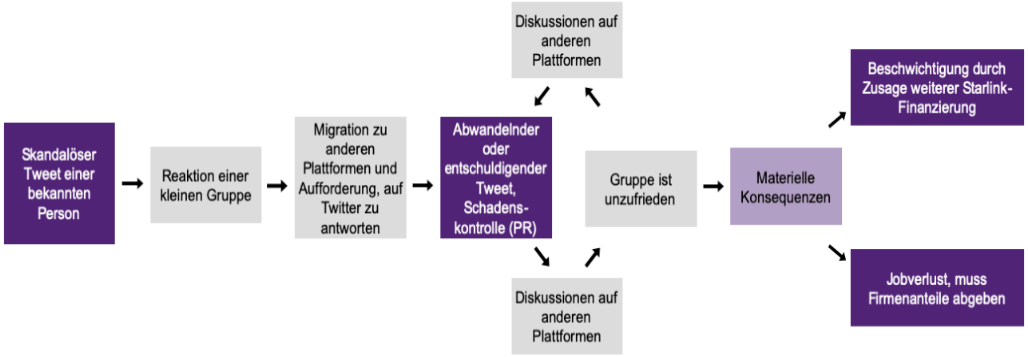

Skandalöser Tweet einer bekannten Person
Reaktion einer kleinen Gruppe
Migration zu anderen Plattformen und Aufforderung, auf Twitter zu antworten
Abwandelnder oder entschuldigender Tweet, Schadens-kontrolle (PR)
Diskussionen auf anderen Plattformen
Diskussionen auf anderen Plattformen
Gruppe ist unzufrieden
Materielle Konsequenzen
Beschwichtigung durch Zusage weiterer Starlink-Finanzierung
Jobverlust, muss Firmenanteile abgeben

Figure 5.12: Time course of the cross-platform shitstorm.

5.9 Discussion and Conclusion

Online shitstorms usually arise when controversial topics are discussed. This sets the stage for heated arguments in which users either defend or reject opposing positions on a topic. An important factor that arguably amplifies the shitstorm is the sharing of outrage. Outrage and negative emotions have been shown to have advantages in competing for attention by encouraging posters to produce and consume more content that evokes outrage, or by polarizing content. This increases both the duration and intensity of engagement because content that evokes outrage tends to be viewed and shared more frequently, so users keep coming back. Outrageous content is, therefore, more likely to go viral by creating emotional contagion across opposing ideological communities (Popa-Wyatt, 2022b,a). The network structure of social media platforms satisfies our preference to belong to an "in-group" and to mingle with like-minded people in clique structures (i.e., homophily). This satisfies our search for identity, which gives us a sense of purpose: We align our values and interests with those of others who share a common identity and similar preferences. Yet our sense of tribalism also makes us vulnerable to harmful narratives and ideologies. This has the effect of promoting segregation in communities and conformity to their practices, which can lead to polarization of content. Network structure divides communities into in-groups and out-groups. This can have two correlated effects: (a) alienating and excluding the voice of perceived out-groups and (b) creating a bubble effect where perceived in-groups reinforce each other's beliefs in the absence of challengers and dissenters. Critically, this division of content not only limits the range of voices that can be heard but also leads to a polarization of opinions, resulting in increasingly extreme versions of those opinions. In the data we looked at, clique formation is most evident on the content-based platforms Telegram and Reddit, where one opinion strongly prevails. On Twitter, on the other hand, outrage builds up in the interplay between supporters and opponents, who come together in a public arena (sometimes after explicit prompting). It is striking that the course of both shitstorms follows the same pattern (Fig. 5.12).

More often, there's a compromise between ethics and expediency.

Peter Singer

6

Implications of the New Regulation Proposed by the European Commission on Automatic Content Moderation

Joint work with **Vera Schmitt**, **Vinicius Woloszyn** and **Jessica de Jesus de Pinho Pinhal**, Technische Universität Berlin, Germany.

The chapter has been removed from the online version for copyright reasons.

6.1 Preface

This chapter was previously published as: Schmitt et al. (2021). Implications of the New Regulation Proposed by the European Commission on Automatic Content Moderation. Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication, 47-51, doi: 10.21437/SPSC.2021-10, https://doi.org/10.21437/SPSC.2021-10. A big part of work on malicious content detection on the web is the area of concern of Ethics of AI and the legal frameworks which are built and constantly improved in order to efficiently regulate the trade-off between the democratic freedoms of the Internet users and the health and security of this informational eco-system. Development of this legal framework is significant for those we seek to defend the most: the children, the historically oppressed minorities harassed by internet users hiding by the mask of anonymity. However, the average internet user also often becomes a victim of toxic behaviour, in the case of fake news and propaganda, as well as any other type of misinformation or disinformation. All of us, being the voters in democratic elections, can be constant unknowing targets of online manipulative and persuasive content. The scandal surrounding Cambridge Analytica (Boldyreva, 2018; Boerboom, 2020; Hu, 2020) is only one of the famous examples. The New Regulation[1] proposed by the European Commission in 2021 seeks to provide the legal framework for how AI applications, including those used for automated content moderation, should function in the EU. In this position paper, we investigate the weak sides of this proposal, contributing to the scientific debate about how AI can be allowed to regulate automated content moderation, such as fake news and hate speech detection. We discuss the existing social media platform's countermeasures to malicious content (e.g. flagging, banning, demonetisation) and in which AI risk categories they would fall under the proposed regulations if moderated automatically. We identify the problems with their enforcement, broad interpretation possibilities, and difficulty

[1] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

translating them into concrete guidelines. This paper rounds up the part of this thesis concerning automated content moderation in the context of social media, giving it a further reaching ethical and legal perspective for the potential implementations of the methods proposed in the previous chapters on social media platforms.

6.2 Contributions

Conceptualization: [Vera Schmitt(50%), Veronika Solopova(30%), Vinicius Woloszyn(20%)];
Methodology: [Vera Schmitt(50%), Veronika Solopova(50%)];
Formal analysis and investigation: [Vera Schmitt(40%), Veronika Solopova(40%), Jessica de Jesus de Pinho Pinhal(20%)];
Writing - original draft preparation: [Vera Schmitt(35%), Veronika Solopova(35%), Vinicius Woloszyn(20%), Jessica de Jesus de Pinho Pinhal(10%)];
Writing - review and editing: [Vera Schmitt(70%), Veronika Solopova(30%)];
Project administration: [Vera Schmitt].
Vera Schmitt coordinated this position paper, gathered the consortium and as a corresponding author was responsible for the final editing, submission and presentation of the paper at the symposium. Vinicius Woloszyn, as a more senior researcher, was actively helping with strategic decisions on the paper structure and contributed to writing the first draft. Jessica de Jesus de Pinho Pinhal, as a specialist in Ethics and Epistemology of AI, was especially responsible for the discussion part, ethical and legal considerations.
I contributed to the paper on all levels, including full responsibility for the hate speech moderation side of the study, analysis of the existing and newly proposed legislation and its consequences for the automated moderation task, in the context of various countermeasures. I participated in both the writing of the first draft and the final editing, contributing to each section of the paper.

Without reflection, we go blindly on our way, creating more unintended consequences and failing to achieve anything useful.

Margaret J. Wheatley

7

A German Corpus of Reflective Sentences

Joint work with **Oana-Iuliana Popescu**, German Aerospace Center, Jena, and **Margarita Chikobava**, German Research Center for Artificial Intelligence, Berlin, Germany.

Supervised by **Ralf Romeike**, **Christoph Benzmüller** and **Tim Landgraf**.

7.1 Preface

This Chapter was previously published as: Solopova et al. (2021). A German Corpus of Reflective Sentences. In Proceedings of the 18th International Conference on Natural Language Processing (ICON), pages 593–600, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI). https://aclanthology.org/2021.icon-main.72/

Although analysis of student essays of various genres has been of great interest to the researcher community for years, the scope of the task was primarily framed as essay scoring. With the rise of the LLM's and the opportunities they brought, many realised the mentoring and coaching potential of the technology for the educational sector and a possibility of solving the two Σ problem (Bloom, 1984), central to Didactics. The two Σ problem is defined as the distance between students' performance in the conventional educational setting (regular classroom) versus one-on-one tutoring, which can be measured in 2 Σ, with individual lessons being that much more productive. The current hope is that the AI assistants could play the role of the one-to-one tutors as an addition to the conventional classroom, which would potentially help to close the gap at least to what Bloom called "Mastery Learning".

Hence, as a second use-case of automated content moderation, I investigate automated moderation of student reflective practice. My eventual goal presupposes an iterative conversation with a student, which would help a student not only get better at writing in some particular genre but also become a better professional and self-sustained learner. In the context of reflective practice, a compulsory part of education for many university specialisations such as Teacher Education and many medical care professions, this tool can help alleviate the immense workload on university tutors who are supposed to give personal feedback to each reflection. The workload also prevents the tutors from giving timely, high-quality feedback, which is necessary for the best progression.

One of the general problems in NLP, but also the field of reflective writing analysis in particular, is a lack of labelled corpora suitable for training the models and even more for benchmarking. The second problem is also the lack of the aforementioned resources in languages other than English. In this Chapter, we present a corpus of reflective sentences, which we collected and annotated as a first step of our project. This data is used in further chapters as training and evaluation material, which we further enriched with new annotations and analyses in the following chapters. Due to many factors, such as the difficulty of collection and the complicated consent procedure to fulfil the data-privacy regulations, which we ensured to fulfil, this corpus is the first open-source instance in the field. Additionally, we present a corpus analysis with the help of a linguistic features extraction script, which will be further used as a rule-based hybrid component of the overall system to give students feedback in the following Chapter.

7.2 Contributions

Conceptualization: [Veronika Solopova(50%), Oana-Iuliana Popescu(50%)];
Methodology: [Veronika Solopova(50%), Oana-Iuliana Popescu(50%)]. Formal analysis and investigation: [Veronika Solopova];
Writing - original draft preparation: [Veronika Solopova(50%), Oana-Iuliana Popescu(30%), Margarita Chikobava(20%)];
Writing - review and editing: [Veronika Solopova, Oana-Iuliana Popescu, Margarita Chikobava, Christoph Benzmüller];
Supervision: [Ralf Romeike, Tim Landgraf, Christoph Benzmüller];
Project administration: [Ralf Romeike, Tim Landgraf, Christoph Benzmüller].

Oana-Iuliana Popescu contributed by curating data collection with German Computer Science students, while she also co-annotated all of the data with the author of this thesis. She also equally contributed to the writing of the first draft and editing. Margarita Chikobava curated the collection of a subset, which was not published in the end as part of this study, because of the unfulfilled consent form requirement. This unpublished subset was, however, used as part of the training data for many a model described in Chapter 8. Margarita also created the visualisation used in this paper and contributed to sections of Didactic content.

In addition to procuring and translating data from Dundee University and Ethics of AI students, as well as co-annotating it, I had a leading part in the creation of annotation guidelines and questions for the Google Forms collection. I also was fully responsible for the analysis part of the paper, performing statistical testing and describing my results in the paper. My contribution also included writing the bigger part of the first draft, as well as revisions for the camera-ready version. I also am the corresponding author, and I continued curation of this dataset after O.I.P and M.Ch. left the project, adding new annotations, which were used to train models in the next chapter.

7.3 Abstract

Reflection about a learning process is beneficial to students in higher education (Bubnys, 2019). The importance of machine understanding of reflective texts grows as applications supporting students become more widespread. Nevertheless, due to the sensitive content, there is no public corpus available yet for the classification of text reflectiveness. We provide the first open-access corpus of reflective student essays in German. We collected essays from three different disciplines (Software Development, Ethics of Artificial Intelligence and Teacher Training). We annotated the corpus at the sentence level with binary reflective/non-reflective labels, using an iterative annotation process with linguistic and didactic specialists, mapping the reflective components found in the data to existing schemes and complementing them. We propose and evaluate linguistic features of reflectiveness and analyse their distribution within the resulting sentences according to their labels. Our contribution constitutes the

first open-access corpus to help the community towards a unified approach for reflection detection.

7.4 Introduction

Consciously experienced and reflected practice is a prerequisite for professionalization (Donald, 1983). For pre-service teachers, reflection is crucial because it belongs to the core competencies of prospective teachers (Combe & Kolbe, 2004; Hänsel, 1996; Shandomo, 2010). In literature, several types of reflection can be found. Core reflection deals with the core of one's personality: mission and identity (Korthagen & Vasalos, 2005), while self-reflection refers to thinking about one's own behaviour, actions, thoughts or attitudes (Bubnys, 2019). The reflection process can be either guided using prompts to indicate the structure of the reflection (Allas et al., 2020), or free, where the reflection process follows no given structure (Sturgill & Motley, 2014). In our corpus, we mainly focus on guided self-reflection.

Educational staff must assess students' reflection texts, yet this is a non-trivial and time-consuming task. Machine learning methods can provide possibilities to create such applications. However, the first step towards this is identifying whether reflection is present in a text or not. Collections of student essays in machine-readable formats have been created for the last two decades for various machine-learning tasks, such as automated essay scoring (Foltz et al., 1999), argumentation mining (Wang et al., 2020a), reflection detection and automated feedback (Wulff et al., 2020). However, to the best of our knowledge, there is no open-source corpus of reflective essays currently available. The reason, in our opinion, lies in the challenges that this kind of data brings. From an ethical point of view, these data are sensitive, since they can be highly personal. In addition, essays are usually collected in an educational setting, and it might be against regulations to publish them. Furthermore, inspiring students to reflect is difficult. As a literature review shows, students mostly write descriptive sentences when journaling (Dyment & O'connell, 2010).

We thus contribute a publicly available, balanced text corpus of reflective and descriptive sentences

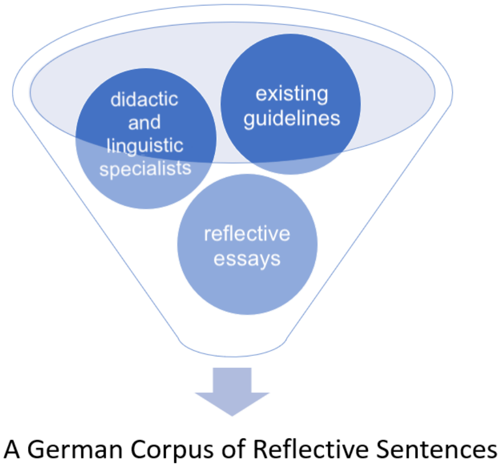

didactic and linguistic specialists
existing guidelines
reflective essays
A German Corpus of Reflective Sentences

Figure 7.1: The main components of our approach.

from students of various universities and disciplines as the first step towards a benchmark for reflection detection in texts. For this, we collected essays from three different sources and anonymized them. We then pre-processed texts into sentences and added manual sentence-level annotations according to a synthesised taxonomy, engaging professional linguists and didactic specialists to refine our criteria. We present our quantitative and qualitative linguistic analysis of the resulting corpus. The link to our data can be found in Appendix A.6.1.

7.5 Related Work

In the context of the multi-genre essay collection, significant works include the British Academic Written English (BAWE) (Nesi & Gardner, 2012, 2013), the Uppsala Student English Corpus (USE) (Axelsson & Berglund, 2002), and the Michigan Corpus of Upper-Level Student Papers (MICUSP) (Römer & Swales, 2010). Several efforts were undertaken to create a specialized reflective corpus of students' essays at sentence level, namely in pre-service and early teachers settings (Wulff et al., 2020; Murphy, 2015) or medical students and personnel (Liu et al., 2019b; Olex et al., 2020). For the didactic case specifically, there has been increasing work in automated detection of reflective sentences in the didactic context (Geden et al., 2021; Jung & Wise, 2020; Liu et al., 2019c; Wulff et al., 2020; Ullmann, 2019, 2017, 2015). However, none of the used corpora are publicly available.

7.6 Data Collection

We collected essays of different lengths in both English and German from students and pre-service teachers. We used the sentence segmenter of SpaCy (Honnibal et al., 2020) to obtain a total of 4232 sentences. During the annotation process, we performed manual anonymization and eliminated all the occurring personal information, including mentioned social media accounts, as well as student and teaching staff names. We describe below how data from the individual sources were collected. For more details on the segmentation, anonymization, and consent processes, see Appendix A.6.1.

Dundee teaching placement essays With the agreement of the University of Dundee, we scraped 122 reflective essays in English written by students in teacher training during their placements in primary and secondary school in 2018. The students had to upload their essays in the form of an e-Portfolio on Glow Blogs[1], a provider of WordPress tools used by the Scottish educational centers. The data reflect their impressions of the Scottish educational system in general and school approaches in particular, the acquired skills, their background, role models, insecurities, and motivations to become a teacher.

We translated the essays into German using DeepL[2] and manually corrected conflicting translations that occurred due to inconsistent formatting. After segmentation into simple sentences, we obtained a total of 3595 sentences.

[1] https://blogs.glowscotland.org.uk/glowblogs/

[2] https://www.deepl.com/translator

Ethics of AI and Software development Using a questionnaire, we collected a set of guided reflective essays in German and English from students of the Free University and the Technical University of Berlin taking a Software Development project or the Ethics of AI lecture. Data was collected repeatedly at an interval of a few weeks.

The students were asked to reflect on the learning outcome since the previous collection. They were guided by a set of questions developed using Gibbs' reflective cycle (Gibbs, 1988), thus spanning the following topics: description of the action they took during their work/learning process, evaluating what they have learned and how to apply it further, what challenges they encountered, and which feelings they note. Additionally, they had to rate how their perception and their competencies of the topic changed and describe why. After segmentation, we obtained a total of 637 sentences.

7.7 Annotation Guidelines

7.7.1 Reflection on the topic

Reflection on the topic accompanies the complex learning process and helps to integrate new knowledge into the existing one and further elaborate on it. In contrast to self-reflection, the object of reflection is part of the subject domain.

We developed our annotation criteria based on the Structure of the Observed Learning Outcome (SOLO) taxonomy (Biggs & Collis, 1982), which was proposed to assess the quality of learning. This taxonomy allows us to identify successful criteria, as it clearly defines the reflection steps. We adapted the three last levels of the taxonomy: multistructural, relational and extended abstract level. At the multistructural level, learners understand the relationship between different aspects but its relationship to the whole remains unclear. At the relational level, aspects of knowledge are combined to form a structure. At the extended abstract level, knowledge is generalized to build a new domain. From the multi-structural level, we adapted the 'combine' action to the following criteria: (1) putting entities into relation (e.g., part of, opposite, but not providing an example). From the relational level, we adopted several criteria: (2) criticism, (3) evaluation and comparison between methods or objects, (4) analysis (e.g., causality, purpose, contributions), (5) classification and assessment of entities. Based on the last extended abstract level, we developed the two following criteria: (6) generating and formulating hypotheses and theorizing, (7) proposal of alternative implementation (suggestions on how something could have been done in a different way).

7.7.2 Self-reflection

To annotate self-reflection, we adapted the schemes proposed by Shum et al. (2017) and Ullmann (2017), searching for evidence of the categories proposed by the authors in our own data. If the sentence met one or more of these requirements, we annotated it as reflective.

From Ullmann (2017) we included: (1) emotions and feelings if they were followed by the cause or description of the circumstances which provoked them; (2) strategy adaptation based on previous experience, (3) different perspectives, and (4) outcome (lessons learned, future intentions, and action plans). From Shum et al. (2017), we implemented rhetoric components and expressions denoting: (5) learning something specific, (6) experimentation and ability, (7) increased confidence or ability, (8) applying theory into practice, (9) retrospection (e.g., *'it would have helped us'*, *'I should have done it'*), (10) expressions of reflecting specifically and (11) shifts in perception and beliefs. From the intersection of both schemes, we included (12) personal beliefs, assumptions, self-assessment and (13) recognition of difficulties, which we aligned with rhetorical expressions of challenge and expressions describing the unexpected to prior assumptions.

We also introduced new categories based on our data and the didactic nature of our project: (14) rhetoric questions, (15) decisions (motives and the decision-making process), (16) motivation. We also determined conditional categories, that, similar to feelings, are annotated, taking into consideration the broader context and given reasons. These are opinions, evaluations, rendition of the words of others, generalisations, doubts (e.g., ‘it seems’, ‘it may be’), ‘even if A, not B’ patterns, own interpretations of definitions, recommendations.

Contrary to Ullmann (2017) and Shum et al. (2017), we categorize descriptive sentences that describe the context of the event that triggers reflection as non-reflective. We support this decision by contrasting their linguistic feature distributions in Section 7.9.

7.8 Annotation Process

We manually annotated the collected sentences according to the synthesised guidelines presented in Section 7.7. If a sentence met at least one of the enumerated criteria we annotated it as reflective, even if it was a long sentence which also consisted of non-reflective components. The sub-corpora from the Software Project and Ethics of AI lectures were annotated in parallel by four annotators (the first authors and our two collaborating didactic specialists from the Friedrich-Alexander University Erlangen-Nürnberg). The initial inter-annotator agreement was low: 0.64 between first authors, 0.32 between first authors and didactic specialists, and 0.33 between the didactic specialists. Consequently, we refined our annotation guidelines and re-annotated the dataset. The Dundee sub-corpus was annotated by the first author, while the third author annotated 100 random sentences in order to verify consistency. The inter-rater agreement between the annotators was 0.66, which is considered substantial (Landis & Koch, 1977; Stemler & Tsai, 2008). Overall, we see that the annotation of reflectiveness is a problematic and tedious task, rather impossible using crowd-sourcing and requiring rounds of discussions and criteria harmonization among interdisciplinary professionals, as also addressed by Ullmann (2019).

7.9 Analysis

7.9.1 Methodology

We investigate morphological features inspired from Ullmann (2015); Liu et al. (2019a); Murphy (2015). However, we hypothesize that reflective sentences also differ in syntactic categories. Using a list of respective subordinate conjunctions and punctuation, we extracted the main types of subordinate clauses and their length, e.g. clause of purpose ('Within the framework of our group, we additionally met online on average once a week *to share research results and plan the next project steps.*', len=10); clause of reason ('I volunteered *because I want to learn to make better slides and I want to get better at presenting.*', len=17).

We compared the feature distribution in reflective versus non-reflective sentences. The resulting distribution of classes is balanced, with 2177 reflective and 1970 non-reflective sentences. We normalized feature counts according to the number of tokens per sentence, transforming them into frequency counts. As our features were mostly non-normally distributed, we applied non-parametric U-tests (Wilcoxon-Mann-Whitney) and multiple-test correction with Benjamini-Hochberg Procedure (N=45 tests). Since we found a large number of significant features, we further restricted our criteria. We filtered out features with medians lying on 0 (i.e., where more than 50% of the counts are 0), which is not taken into consideration by the U-test. Instead, it considers mean ranks, i.e., the arithmetic average of the positions in the list.

7.9.2 Results

The number of tokens in the sentence appears to be one of the most discriminating factors: reflective sentences tend to be longer, while non-reflective sentences are often nominal and/or contain short enumerations. At the same time, reflective sentences tend to be complex (with both subordinate or coordinate clauses using respective conjunctions). Relative clauses are the most frequent in reflective sentences, as they bring additional details describing the subject. Contrary to our expectations, the clauses of reason and purpose, typically used in justifications, show only a slight positive trend for reflective sentences in the Dundee sub-corpus, possibly because it often illustrates a situation and can contain descriptive causes and goals, e.g., *'We did not go outside because of the rain'*. The trend becomes stronger in the self-reflection sub-corpora.

We can observe the presence of solid justification with our 'claims' feature, which checks matches with opinion words (e.g., 'standpoint', 'sure', 'convinced', 'opinion'), and 'supports', which is a collective count of subordinate clauses of reason, purpose, concession, condition and adversation. All subordinate clauses we measured are generally more present in the reflective part of the data set, and the mean length of clauses of reason and purpose is also generally longer. Concessive clauses appear to be the most numerous in this kind of text. Reflective sentences also show a higher probability of explicit coherency markers with discursive connectives (e.g., 'although', 'however').

As for the tenses used, reflective sentences are more often written using the Future tenses, while non-reflective utterances show a slight preference for the Past tenses.

Our 'personalizing' marker, which shows usage of first person singular and plural of pronouns (personal, possessive and reflexive), is found to be significantly more present in reflective sentences, as also found by Ullmann (2015), as well as a number of adverbs, verbs and adjectives (Murphy, 2015). However, we also measured usage of the German indefinite impersonal pronoun *'man'*, which similarly to the English pronoun 'one' can be considered a tool to generalize, distance the authors from the opinion they express, and make it less personal (hence, 'distancing' feature). Counter-intuitively, it was also found slightly more used in reflective sentences, rather than in descriptive ones.

Interestingly, our data also shows a negative trend for lexical words in reflective sentences and a positive one for stop words, which means that reflective sentences tend to be wordier, but less informative.

High modality words (e.g. 'actually', 'categorically') strongly correlate with sentence reflectiveness, while modal verbs and subjunctive mood (German *Konjunktiv* I and II) show the same trend in all but Dundee sub-corpus. This trend of discrepancies between the original German and translated English data calls for further investigation into differences between reflection articulation in different languages.

7.10 Conclusion

With the proposed corpus, we aim to make the first step towards a more unified approach to reflection detection. At the moment, it is not possible to compare existing models, as there is no publicly available benchmark for this task. To address this issue, we created an open-source annotated text corpus of reflective and descriptive sentences from students of various universities and disciplines. We also provide the quantitative and qualitative analysis of the gathered data and describe the annotation procedure and quality assurance measures we took.

Our work has several limitations. Our annotators are not native German speakers, which can influence labelling. However, this will be re-visioned with later versions of the corpus, as we plan to increase the number of annotators and include native speakers. Another drawback is the automatic translation of the English data into German. While we plan to quantitatively increase our corpus with German data in the future, the Dundee sub-corpus provides a valuable addition. This way, however, it largely influences the results for language-specific features such as subjunctive mood presence, which can appear in translations, but are still much more common in German than in modern English.

We address the low inter-annotators agreement problem with harmonization sessions and refinement of the coding scheme to ensure coverage of complicated instances. We report that with each iteration, inter-annotator agreement increased significantly. Thus, we reckon that a fruitful discussion of linguists and specialists in the field on the focus of the task, which is a time-consuming process, is the only probable answer to the annotation of cognitive, subjective categories.

7.11 Future work

Sentence-level segmentation has a significant disadvantage compared to text-level processing. Nevertheless, for modern classification algorithms, there is a need for an immense amount of data points. Thus, we decided to trade off context for the sake of robustness. In the future, we aim to prove the hypothesis that textual-level reflection can still be reconstituted, computing an overall reflectiveness score. Finally, binary classification is only the first step, while we plan to add more granular reflection level categories according to Fleck & Fitzpatrick (2010a), sentiment polarity, emotions and the position of the sentence in Gibb's cycle (Gibbs, 1988). We also plan to expand the corpus with a larger number of guided reflections from different disciplines. Our overall goal is automated reflective essay analysis, which we plan to compare to the existing results by Ullmann (2019); Wulff et al. (2020), in order to propose an adequate level of feedback that matches the student's needs.

7.12 Acknowledgements

We would like to thank our collaborators from Friedrich-Alexander University Erlangen-Nürnberg, namely Chengming Zhang and Jessica Schiessl for their participation in the co-annotation process, as well as Prof. Dr. Michaela Gläser-Zikuda and Prof. Dr. Florian Hofmann for the sanity check of our guidelines from the pedagogic perspective. We are also very grateful to Derek Robertson and Dundee University for granting us permission to process and publish their Glow Blogs essays.

Stochastic parrots are haphazardly stitching together sequences of linguistic forms according to probabilistic information about how they combine but without any reference to meaning.

Emily M. Bender (Bender et al., 2021)

8

PapagAI: Automated Feedback for Reflective Essays

Joint work of Petrakip project consortium, including: **Eiad Rostom**, **Adrian Gruszczynski**, **Fritz Cremer**, **Sascha Witte**, and **Fernando Ramos López**, supervised by **Ralf Romeike**, **Christoph Benzmüller** and **Tim Landgraf** at Freie Universität Berlin, Germany and **Chengming Zhang** and **Lea Plößl** **Florian Hofmann** and **Michaela Gläser-Zikuda** at Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany. The chapter has been removed from the online version for copyright reasons.

8.1 Preface

This Chapter was previously published as: Solopova et al. (2023c). PapagAI: Automated Feedback for Reflective Essays. In: Seipel, D., Steen, A. (eds) KI 2023: Advances in Artificial Intelligence. KI 2023. Lecture Notes in Computer Science(), vol 14236. Springer, Cham. https://doi.org/10.1007/978-3-031-42608-7_16

The first problem of the Large Language Model (LLM)s is that the naturally looking output comes at a sacrifice of the possibility of control. Recent works in prompt engineering (Huang et al., 2023; Zhang et al., 2023b; Lin et al., 2022), shed light on the abundance of evidence of biases and hallucinations, which are distorted facts presented as real (Das et al., 2022), non-existent citations (Gao et al., 2023), and sociolinguistic or cultural biases (Kolisko & Anderson, 2023), while numerous benchmarks are already present to quantify this task (Li et al., 2023; Liu et al., 2022). These hidden and openly present misbehaviour in LLM outputs points to the so-called stochastic parrots' problem (Bender et al., 2021): the model generates the content "it considers" truthful, or one that could realistically appear in human language, with no regard to its actual existence. Although temperature parameters may control this to a certain degree, current research does not give perfect recipes.

The second problem of the LLM's is their cost and sustainability, especially in the academic context. The quickest API implementations, such as ChatGPT API, come at the expense of willingly submitting your material for the next GPT model generations, which cannot be the preferred course of action in case of an obligatory task performed by students. Meanwhile, the open source models, however numerous and accurate they appear, need extensive computational power to be fine-tuned, to infer and to be implemented in a production flow of an application, wherein a chat-like environment, a user expects quick responses and not a 4 to 15-minute ones. However, this is the usual time needed for CPU implementations on a big input prompt, such as a student essay (250+ words). GPU implementation does not seem to be much better either, as while one can speed up inference using multiple

GPUs, the model should be able to be fully loaded on each GPU used; thus the RAM of each graphic card is crucial. At the moment, as this thesis is written, the most impressive implementations, such as Llama, Wizard, and Falcon, all weigh more than 26-45+Gb, which requires the project to have highly expensive high-memory GPUs. To put this into perspective, the most recent state-of-art Falcon with 180 billion parameters (Penedo et al., 2023) openly states that one needs "at least 400GB of memory to run inference swiftly" and recommends 8 GPUs with 80GB each, with only 4X23GB GPUs available in our case. To put things into perspective, to run its previous less robust iteration, Falcon with 40 billion parameters (Xu et al., 2023) needs at least 85-100GB of memory. The ecological impact is even more concerning than the economic one, with the increasing amount of energy consumption LLMs require (Rillig et al., 2023).

In this work, we present a hybrid approach, where we do use a standalone LLM, but predictions of various previous generation Small Language Model (SLM) classifiers (BERT, RoBERTa and ELECTRA) and linguistic processing module (from the previous Chapter) for text understanding and test profiling and a reasoning rule-based template creation module on top. Our classifiers outperformed the GPT-3.5 model on several tasks, namely emotion detection and cognitive cycle stage identification. We describe our system and report the individual models' performances. In the next Chapter, we use this system with two different feedback formats to evaluate how its usage improves student performance.

8.2 Contributions

Conceptualization:[Veronika Solopova(50%), Chengming Zhang, Christoph Benzmüller, Tim Landgraf, Ralf Romeike, Florian Hofmann,Michaela Gläser-Zikuda];

Methodology: [Veronika Solopova(50%), Eiad Rostom(15%), Adrian Gruszczynski(15%), Fritz Cremer(15%), Chengming Zhang(5%)];

Formal analysis and investigation: [Veronika Solopova(40%), Eiad Rostom(20%), Adrian Gruszczynski(20%), Fritz Cremer(20%)];
Application software development: [Sascha Witte(60%), Veronika Solopova(40%)];
Writing - original draft preparation: [Veronika Solopova];
Writing - review and editing: [Veronika Solopova, Tim Landgraf, Christoph Benzmüller];
Supervision: [Ralf Romeike, Tim Landgraf, Christoph Benzmüller,];
Project administration: [Ralf Romeike, Tim Landgraf, Christoph Benzmüller, Michaela Gläser-Zikuda, Veronika Solopova].

The PetraKIp project was managed in the interdisciplinary consortium of two universities: Freie Universität Berlin, administered by Ralf Romeike, Tim Landgraf and Christoph Benzmüller, and Friedrich-Alexander Universität Nürnberg-Erlangen, under the supervision of Michaela Gläser-Zikuda. Christoph Benzmüller also, later on, moved to a position in Otto-Friedrich-Universität Bamberg, with a third university informally joining the project. However, for the bigger part of the project, after Margarita Chikobava and Oana-Iuliana Popescu left in 2022, I became the main coordinator, fully responsible for the implementation of the AI deliverables of the project and the communication between the two teams. Hence, although this study is the work of a big consortium, I became the corresponding author and also I am the main contributor to the conducted study. It is also sensible to mention that my contribution to the project started already at the grant proposal writing stage in 2020, when together with Christoph Benzmüller, I proposed potential Natural Language modules for the system. While the sketch of the application and its modules were decided by the whole consortium, our Didactics partners (Michaela Gläser-Zikuda, Florian Hofmannand and Chengming Zhang) were consulting the AI team on the nature and categorisation of important features in the Reflective essays, which actual tutors consider during the evaluation process. From this fruitful collaboration and constant negotiation, we were able to formulate realistic tasks and determine labels for the models the AI team trained.

All models were trained either by me or under my supervision in the frame of Master's and Bachelor theses in support of the PetrKIp project. Fritz Cremer trained the Sentiment analysis model within his Bachelor thesis "Comparison of Machine Learning Models for German Sentiment Analysis" (2022) under my close supervision. We co-annotated our corpus from the previous chapter with Sentiment labels. The Reflective level detector was built by Adrian Gruszczynsk for his Master's thesis "Evaluation of Machine Learning Approaches for Assessment of Reflective Level" (2022). I annotated the data for this study alone, with slight support from the Didactics team and supervised all experiments. The emotion classifier was trained by Eiad Rostom, in the context of his Master's thesis "Emotion Recognition in Reflective Text Using Transformer Models and Transfer Learning" (2023). I created guidelines and co-annotated the data with Eiad and Fernando Ramos López, also supervising the thesis. I trained the models for Gibbs cycle detection and topic modelling tasks, created the feature extraction script and connected all models to the production system of the PapagAI application. I also conceptualised and built the whole Hybrid AI architecture presented in this study and performed GPT experiments. Together with Sascha Witte, who was responsible for the software development of the application, I curated and updated the AI module of the application for the most of duration of the PetraKIp project. Fernando Ramos López contributed to the creation of templates for the AI answers in different languages. Additional data used for this model was collected by Tim Landgraf and me in his Software development course which took place in the summer semester of 2022 at Freie Universität Berlin, and also provided by Chengming Zhang and Lea Plößl.

In addition to the aforementioned shared contributions, I prepared the first draft of this paper, which was edited with Christoph Benzmüller and Tim Landgraf's help, and I was also responsible for the submission and implementation of the reviewer's suggestions for the camera-ready version.

Knowing yourself is the beginning of all wisdom.

– Aristotle, Metaphysics (350 BC)

9

Automated Feedback Can Foster Deeper Reflections

Joint work with **Chengming Zhang** at Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany,

supervised by **Florian Hofmann** and **Michaela Gläser-Zikuda**.

As well as **Fernando Ramos López** supervised by **Tim Landgraf** and **Judith Meinschaefer** at Freie Universität Berlin, Germany.

9.1 Preface

Part of this Chapter (Study II) was previously published as a Master thesis: "Measuring the effects of linguistic formality on user perception of virtual assistants using computational linguistic methods", by Fernando Ramos López (2023), supervised by Judith Meinschaefer, Tim Landgraf and Veronika Solopova. It was included in this thesis with the primary author's agreement.

While the beneficial impact of AI assistants on student performance and motivation has been investigated from various angles Benotti et al. (2017); Chen et al. (2019); Wambsganss et al. (2021), a positive influence of AI guidance on reflective practice is yet to be proven, and the difference of performance after the human and automated feedback is not yet calculated. In this Chapter, similarly to the evaluation we performed with Check News in 1 Click in Chapter 4, we study users' reactions to the PapagAI automated feedback we describe in the previous Chapter. In the two user studies we present, we analyse the students' reflective performance after tutor and AI feedback. We also look into the language register preferences of users when it comes to AI feedback. Our findings suggest an absence of marked preference for human feedback among students, with comparable performance metrics observed in the treatment (AI feedback group) and control group (human feedback group). Our data also indicate a potential enhancement in the quality of student reflection attributable to our feedback mechanisms. However, the robustness of these conclusions is tempered by limitations inherent to the small scale of the study groups and the single-blind methodology employed.

9.2 Contributions

Conceptualization:[Veronika Solopova(30%), Chengming Zhang(30%, Fernando Ramos López(20%),Florian Hofmann(20%)];

Methodology: [Veronika Solopova(30%), Chengming Zhang(40%), Fernando Ramos López(30%)].
Formal analysis and investigation: [Veronika Solopova(20%), Chengming Zhang(40%), Fernando Ramos López(40%)];
Writing - original draft preparation: [Veronika Solopova].

For Study I, I developed the software infrastructure and adapted and implemented the feedback format from the previous chapter to the one proposed by Florian Hofmann, with a professional topic of class management. Chengming Zhang curated and statistically analysed the user-study results and created the tables presented in this study. I analysed the broader impact of the results and interpreted the trends we detected.

For Study II, Fernando Ramos López co-created the prompting questions and sentences used in the system's outputs, curated the user study and analysed the results, while the author of this thesis implemented the feedback format and the user-study infrastructure into the PapagAI, wrote the code for the feature extraction and analysis and supervised Fernando Ramos López work.

9.3 Abstract

In this report, we attempt to evaluate the usefulness of the PapagAI application. We perform two studies: (I) We compare student performance following human tutor and AI feedback; (II) We verify what formality level of language the students prefer and if they improve their reflective level after one treatment. Our results indicate a positive influence of the PapagAI application on student reflective skills. However, the inability to gather relevant control/treatment groups prevents us from statistically quantifying all of the results. Thus, we mostly describe positive trends.

Keywords: Automated feedback, Human feedback, Formality

9.4 Introduction and Related Work

Numerous attempts were undertaken to measure how timely feedback improves students' capacities to be self-regulated learners and excel in their studies. In a 250-studies meta-analysis, Black & Wiliam (1998) concluded that feedback improved student learning and satisfaction. According to Butler & Winne (1995) and Zimmerman (2002) learning diaries with individual feedback improve Self-Regulated Learning (SRL). Bellhäuser et al. (2023) investigated the effects of automatically generated adaptive feedback on daily SRL. Using randomly assigned experimental groups with and without feedback, they showed that students in both groups improved in planning, self-motivation, self-efficacy, volition, and reflection. However, students receiving feedback set more ambitious goals claimed higher self-efficacy and made better plans. At the same time, feedback did not affect intrinsic motivation, effort, or procrastination. They hypothesised that the learning diary constituted an effective intervention on its own, irrespective of feedback provision. Wambsganss et al. (2021) evaluated whether a chatbot providing adaptive tutoring helped students in the treatment group write more convincing texts. Based on their results, students using the chatbot wrote more convincing texts and argued their claims better from a formal point of view. In Chen et al. (2019) most students improved their code and positively assessed the automated prompts to foster deeper reflection on their codes' possible security problems. In Benotti et al. (2017), students also improved their learning and became more interested in the learning topics.

Nonetheless, both students and their instructors are often reported to perceive AI systems negatively, and even as invasive (Seo et al., 2021). While they appreciate AI's ability to give quick responses, students are concerned with loss of privacy (Luckin, 2017; Chan, 2019; Bajaj & Li, 2020; Lee, 2020) and potential algorithmic bias (Crawford & Calo, 2016; Murphy, 2019). The instructors were worried about AI's limiting factor on students' ability to learn independently and think critically (Wogu et al., 2018).

At the same time, students are often unaware of the positive influence that AI brings (Fichten et al., 2022), making the user an unreliable source if we consider measuring application performance only using the questionnaire. AcaWriter (Knight et al., 2020) is the only tool to the best of our knowledge that offers automated feedback to student reflections and was tested with positive results (85.7% of students perceived the tool positively). In contrast, the impact on the reflection quality over time was not measured and remains unclear. Thus, we designed several studies to evaluate the effectiveness of the PapagAI app and determine its impact on student reflectivity. In collaboration with Chengming Zhang and Friedrich-Alexander-Universität Erlangen-Nürnberg, in a single-blind study, we verified if Didactic's students perform better after automated or human tutor feedback if structured similarly and compared these results with the perceived usefulness and credibility of the students unknowingly attributed to the feedback they received. For this study, we used adapted feedback templates for the teacher education use case. We also conducted a study to investigate if the students preferred and found more credible informal or formal language in the feedback we describe in Chapter 8 and if the reflective depth improved after one feedback treatment.

9.5 Methods and Study Designs

9.5.1 Study I. Human versus machine feedback

Using the same features and AI modules as in Chapter 8, the feedback was reformatted and adapted to the Didactics' Class Management course needs. We formulated four levels of feedback on a scale from 0 to 4 according to ten parameters as illustrated in Table 9.1. To produce a score evaluating the structure, the algorithms verified the balanced presence of discourse markers (as in Chapter 7) and the appropriateness of Gibbs cycle components (as in Chapter 8). We also analysed the appropriateness of the Gibbs cycle (Chapter 8.6) component according to the part of the text where it is found: Description is expected majorly in the introduction, Analysis in the middle part, while Conclusion and

Table 9.1: Adapted pedagogic feedback constituents in Study I

Feedback aspect	Example	Level
Structure	The structure is fully comprehensible	3
Size	You have written 356 words. The size is fully appropriate	3
Language	The language is mostly understandable	2
Goal of reflection	a goal of reflection is indicated	2
The topics of the reflection	The topics you mentioned have been sufficiently addressed. The topics you mentioned seem to be sufficiently covered.	2
Subject-specific knowledge	In the reflection an assessment of the used school pedagogical expertise is fully recognisable.	3
Multiple perspectives: Cognitive aspects	Only one perspective is present in your reflection.	1
Multiple perspectives: Affective aspects	In the reflection, your emotions and feelings were not mentioned	0
Learning results	In the reflection, consequences for future work processes/learning processes are fully indicated.	3
Overall reflective level	Descriptive reflection: The teaching situation is not only described but also evaluated - on the basis of objective professional or own subjective views. When analysing the teaching situation, reasons for a pos./neg. assessment is given. Personal assessments, judgments, etc., are part of the reflection. The different aspects are not coherently linked and are not strictly causally related.	1

Future Plans by the end of the essay.

The size of the essay is counted through tokenisation performed with Spacy. The language is also supposed to be neither overloaded, with too many embedded complex sentences, nor too simple, when only simple sentences are used. Whether the goal was mentioned is detected through vocabulary matching. Here, a dictionary of all frequent synonyms is matched to lemmatised sentences. The topics are extracted as described in Chapter 8.6 and are enumerated in Table A.2. We verify whether the student wrote at least three sentences on each mentioned topic and if at least one contains analysis according to the Gibbs Cycle model. The subject-specific knowledge module checks how many of the mentioned topics are actual pedagogical topics, meaning they are topics from Clustering 1 and not general ones (Clustering 2).

"The multiple perspectives" constituent includes cognitive aspects that consider how many perspectives were present and affective ones, responsible for the presence of personal emotions. We detect multiple perspectives through the presence of various pronouns and vocabulary associated with school actors: students, pupils, interns, teachers, school management, and headteachers being recalled.

We identify emotions and sentiment with models from Chapter 8.6. "Learning results" are extracted through the presence of Gibbs cycle's Future Plans and Conclusion classes and the number of sentences attributed to these labels. Overall reflective level according to Fleck and Fitzpatrick's scheme is identified with the model from Chapter 8.6.

In this study, which took place in June 2023, the target group is the teacher trainees (45 students) participating in the Class Management course in Friedrich-Alexander-Universität Erlangen-Nürnberg, which are predominantly female (80%). Most (64%) have no previous AI-related experience, 13.34% have certain AI experience in the context of teacher education and 22% claim to have experience outside of their current field of education. The participants are mostly 20 years old with a standard deviation of 2.65 and are mainly at the end of their third semester (with SD of 1,43). The AI feedback group has 22 students, while the Teacher feedback group accounts for 23 participants. The feedback was given in 2 rounds (1st RW and 2nd RW), with feedback coming after each reflection after around one month of delay, as it was decided to send human input produced by the course lecturer and the AI feedback simultaneously.

9.5.2 Study II. Formality level interaction Study

In this study, framed as a Master thesis of Fernando Ramos Lopez, we investigated the degree of formality that participants prefer when interacting with an automated feedback system with a chatbot interface. We also analysed if they improved after one treatment. The study required the participants to write two 200-word reflections and complete two surveys. Every participant would write a reflection and receive personalised automated feedback.

As users were shown to dislike very informal registers when revealing personal information, and a fully formal style would not look harmonic in the chat-like design, two levels of formality were tested in the study: slightly formal and slightly informal. Informal feedback included expressive punctuations (e.g "!!!", "???"), word elongations (e.g. "yeees"), interjections, smilies and emojis and the personal pronoun "you". Apart from the absence of the aforementioned features, formal feedback also used the impersonal pronoun "one".

The participants were divided into two groups. Both groups wrote one first reflection. Then, on stage (1), Group A received formal feedback, while Group B received informal feedback. Both groups were asked to answer Google forms[1] questionnaire and to write the second reflection implementing the questions proposed by the automated feedback. Then, contrary to the first stage of the study, in the second round (2), Group A received informal feedback, and Group B received formal feedback. The participants were asked to answer the questionnaire again. Swapping the order of the first type of feedback ensured that we could control the influence of the novelty of the first interaction with

[1] https://docs.google.com/forms/d/11iwe7SV1lnQCrACOGzIYzYtKLNszYOFE8GR2EzORriA/edit#responses

the application, minimising the order bias and measuring the formality influence as a more isolated variable.

The participants were recruited from linguistics and romance language institutes, as well as a Computer science software development project class from Free Universität Berlin. The participants accepted the terms and conditions of the study by accepting the AGB of the PapagAI app. The instructions were provided as a video[2]. The reflections could be written in German, Spanish, or English and focus on some experience from school or university.

After each reflection-feedback stage, the survey included eight questions detailed in Table 9.2. A five-level Likert scale was used in the study to rate the questions. The collected reflections were categorised into four groups, namely "before formal" and "before informal", which are reflections written before receiving any feedback, "after formal," and "after informal," which refer to the second reflections written after receiving the input for Reflection № 1.

We adapted a feature extraction module from Chapter 8 to analyse the reflection. The values were normalised according to the text length of each reflection to receive relative values, also known as term frequencies (TFs). P-values were calculated using the Mann-Whitney U test and the T-test with a p-value threshold ≤ 0.05 based on the distribution of the feature (gaussian or not) along with the corresponding effect sizes measured with Cohen's d for each linguistic feature and multiple test correction with False Discovery Rate (FDR) procedure using Hochberg's method was applied to control for the risk of type I errors. Following Pavlick & Tetreault (2016) we also decided to measure if users mimic the level of formality of the feedback they obtain after the first reflection in their second reflection. Thus, we also compared several features between reflections 1 and 2, using Dewaele and Heylighen's (Heylighen & Dewaele, 2002) formula for cross-validation:

$$F = \frac{f(nouns)+f(adjectives)+f(preposition)+f(articles)-f(pronouns)-f(verbs)-f(adverbs)-f(interjections)+100}{2}$$

[2] https://www.youtube.com/watch?v=QdmZHocZQBk

Table 9.2: Survey questions in Study II

Study questions
Q1. Does the language used by the chatbot sound formal?
Q2. Would you like to interact with a chatbot that replies in this manner?
Q3. Does the language used by the chatbot make the content of the messages sound trustworthy?
Q4. Does the language used by the chatbot sound appropriate to the chat setting?
Q5. Is the language of the chat appropriate to the topic? (a chatbot designed to improve your reflective skills)
Q6. The content of the feedback provides helpful advice to improve the depth of my reflection.
Q7. The content of the feedback was relevant to my reflection.
Q8. The content of the feedback provides valuable general advice.

where f is the frequency of all terms of this part of speech in the document.

Based on this score, the formal word category, which is not deictic (nouns, adjectives, prepositions, and articles), is expected to rise in frequency with increasing text formality. In contrast, deictic categories (pronouns, verbs, adverbs, and interjections) are expected to decline. Thus, the bigger the score, the more formal the text.

9.6 Results

9.6.1 Study I

We performed statistical testing of the results between the two groups based on several dimensions, such as reflective level based on Fleck and Fitzpatrik's scheme, sentiment polarity and subjectivity level, as well as linguistic features including morphological ones like several different parts of speech, as well as Cognitive (words related to mentioning causes, differences, insights etc.) and Temporal indicators (focus on the past, future or present tense) using Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 1999). As illustrated in Figure 9.3, there is no substantial difference in terms of improvements over the reflective level between the AI and the Teacher feedback groups. In fact, in the AI group, one student wrote a worse reflection in the second round, and three students showed lower performance the second time after the Teacher's feedback.

Considering affective indicators such as various sentiment polarities and levels of subjectivity, we can see that overall, in both groups, students became slightly less pessimistic and more optimistic about their performance. At the same time, extreme positivity and negativity also disappeared. Slight differences lie in the more significant number of students changing their reflection's sentiment from negative to neutral. The teacher feedback group decreased in neutrality while the AI feedback one increased. Subjectivity slightly increased for the AI group and decreased for the Teacher group. The difference is, however, not statistically significant according to the T-tests. Regarding linguistic indicators, they mostly decreased from round 1 to round 2. The only one that increased in both groups was the presence of casual markers like justifications and subordinate clauses of reason. The teacher feedback group also started using more coordinated sentences. Generally, the number of complex sentences dropped for both groups, which may mean that many students were criticised for overly loaded language by human tutors and automated systems.

Table 9.4 shows the survey with a standardised questionnaire. It consisted of credibility (four items, $\alpha = 0.61$), usefulness (four items, $\alpha = 0.69$), support level (six items, $\alpha = 0.79$), and result-oriented dimension (four items, $\alpha = 0.83$), using a five-point modified Likert scale (Willems et al., 2020). This scale ranged from 1 ("strongly disagree") to 5, denoting (strongly agree.) Based on the questionnaire, AI feedback still showed slightly higher credibility and usefulness level and a Result-oriented dimension. In contrast, Teacher feedback was considered better in terms of support level. The median of the answers lies on the 3, with a low Standard deviation, which shows that students were uncertain about their answers.

9.6.2 Study II

A total of 32 participants completed the study fully. In terms of survey results, the differences between the formal and informal groups are not statistically significant for any of the questions Figure 9.1. As illustrated in Table 9.6 the results indicate that Group A has a significant difference in the use of Gibbs

Table 9.3: Affective indicators analysis. PST = Paired Sample t-test; IST = Independent Samples t-test; RW = Reflective Writing. ***p < .001; *p < .05

.

	AI Group **(n = 22)**			**Human Group** **(n = 23)**			IST
Affective indicators	1st RW	2nd RW	PST	1st RW	2nd RW	PST	
Very negative	0	0	-	2	1	-	-
Negative	11	8	-	9	8	-	-
Neutral	14	17	-	12	9	-	-
Positive	4	7	-	2	5	-	-
Very positive	5	2	-	2	4	-	-
Positive Emotions	3.79/1.43	4.35/1.95	-1.96	3.61/1.57	4.56/1.37	-2.95***	-0.47
Negative Emotions	1.91/1.32	1.45/0.91	2.15*	1.34/1.08	1.60/1.01	-1.24	-0.57
Polarity	0.23/0.21	0.30/0.24	-1.39	0.29/0.17	0.30/0.19	-0.18	-0.11
Subjectivity	0.08/0.08	0.11/0.12	-1.26	0.10/0.09	0.06/0.07	1.84	2.34*

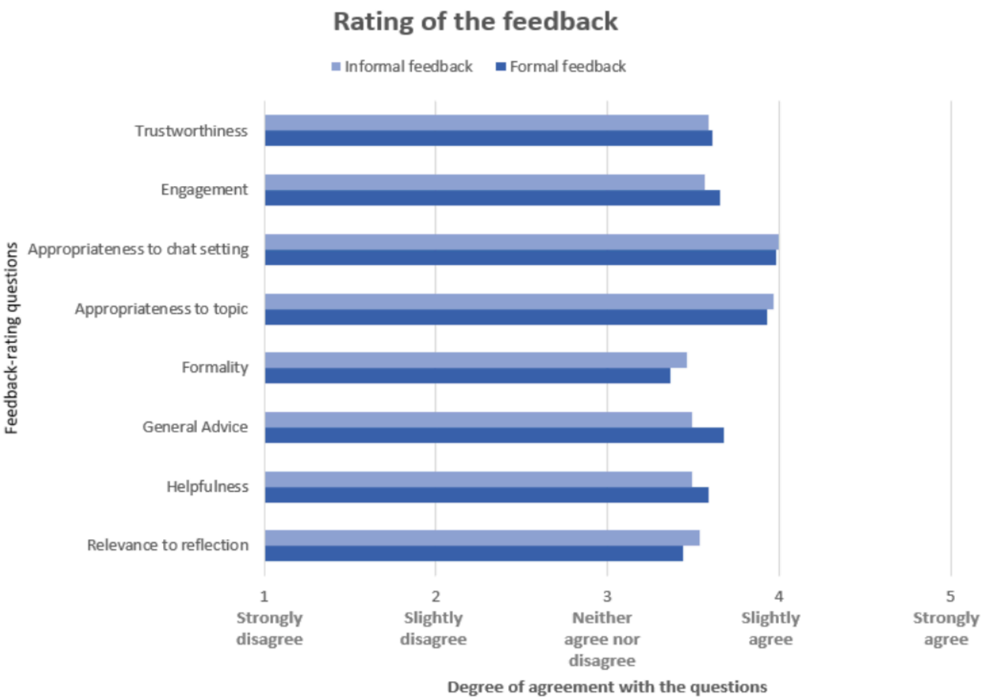


Figure 9.1: The figure illustrates the eight feedback rating questions results. Each upper bar represents the informal feedback while the formal one is the bar below.

Table 9.4: A large effect size (d≈0.8), moderate effect size (d≈0.5). When the effect size is positive, it suggests that an increase or change in the variable, in this case, "Gibbs description", has a strong positive influence on the results. A large negative effect size indicates a substantial impact in the opposite direction.

	AI Group (n = 22)			Human Group (n = 23)			IST
	1st RW	2nd RW	PST	1st RW	2nd RW	PST	
Readability test							
F–K Scores	16.55/2.53	19.01/13.33	-1.13	16.46/2.13	16.62/2.09	-0.44	0.92
Lexical density							
WC	310.97/167.32	298.82/134.86	0.58	317.22/154.30	312.81/154.84	0.20	-0.38
Adverbs	52.44/25.68	**50.20/26.90**	0.57	51.48/29.99	49.26/32.92	0.51	0.12
Verbs	31.91/16.58	29.91/15.19	0.76	32.48/17.74	30.44/15.79	0.83	-0.13
Adjectives	14.29/11.00	**13.03/9.27**	0.69	15.81/9.27	14.67/9.34	0.65	-0.68
Nouns	62.32/36.39	61.24/27.99	0.25	65.22/30.31	**66.93/31.20**	-0.32	-0.75
Sentence structure							
Coordinating	12.79/9.98	12.38/7.51	0.28	11.93/7.03	**13.22/7.06**	-0.86	-0.45
Subordinating	7.94/5.26	**7.50/4.55**	0.53	8.19/5.24	8.11/4.92	0.08	-0.50
Simple_sentences	10.11/7.30	**10.17/6.69**	-0.04	9.93/5.54	9.89/5.82	0.03	0.18
Complex_sentences	3.02/2.83	2.79/2.40	0.54	2.81/2.29	3.22/2.62	-0.69	-0.66
Cognitive							
LIWC.cogproc	22.84/4.10	21.46/3.77	1.71	24.23/3.98	21.28/4.29	3.88***	0.17
LIWC.insight	5.31/1.19	4.84/1.35	1.66	6.89/1.67	5.08/1.63	3.49**	-0.61
LIWC.cause	2.51/1.11	**2.97/1.15**	-2.12*	2.81/1.00	**2.92/1.41**	-0.37	0.17
LIWC.discrep	2.12/1.22	2.10/1.14	0.06	2.20/1.36	**2.26/1.41**	-0.20	-0.47
LIWC.tentat	3.14/1.55	3.17/1.66	-0.11	3.04/1.74	2.70/1.78	1.42	1.07
LIWC.certain	3.82/1.60	3.31/1.61	1.43	3.40/1.34	3.16/1.19	0.77	0.40
LIWC.differ	4.23/1.76	3.97/1.58	0.76	4.79/1.99	4.07/1.86	1.90	-0.21
Temporal							
LIWC.focuspast	4.71/1.81	4.57/1.50	0.42	4.91/1.53	4.80/1.50	0.35	-0.61
LIWC.focuspresent	5.48/1.40	5.66/1.66	-0.48	5.19/1.59	4.91/1.57	0.85	1.79
LIWC.focusfuture	0.76/0.64	0.94/0.75	-1.35	0.92/0.84	0.85/0.69	0.31	0.49

Table 9.5: User questionnaire results.

	AI Group (n = 22)		Human Group (n = 23)		α	Welch's *t*-test
Variable	*M*	*SD*	*M*	*SD*		
Credibility	3.34	0.32	3.29	0.57	0.61	0.340
Usefulness	3.06	0.49	3.02	0.43	0.69	0.254
Support level	3.22	0.57	3.25	0.48	0.79	-0.217
Result-oriented dimension	3.27	0.54	3.25	0.51	0.83	0.073

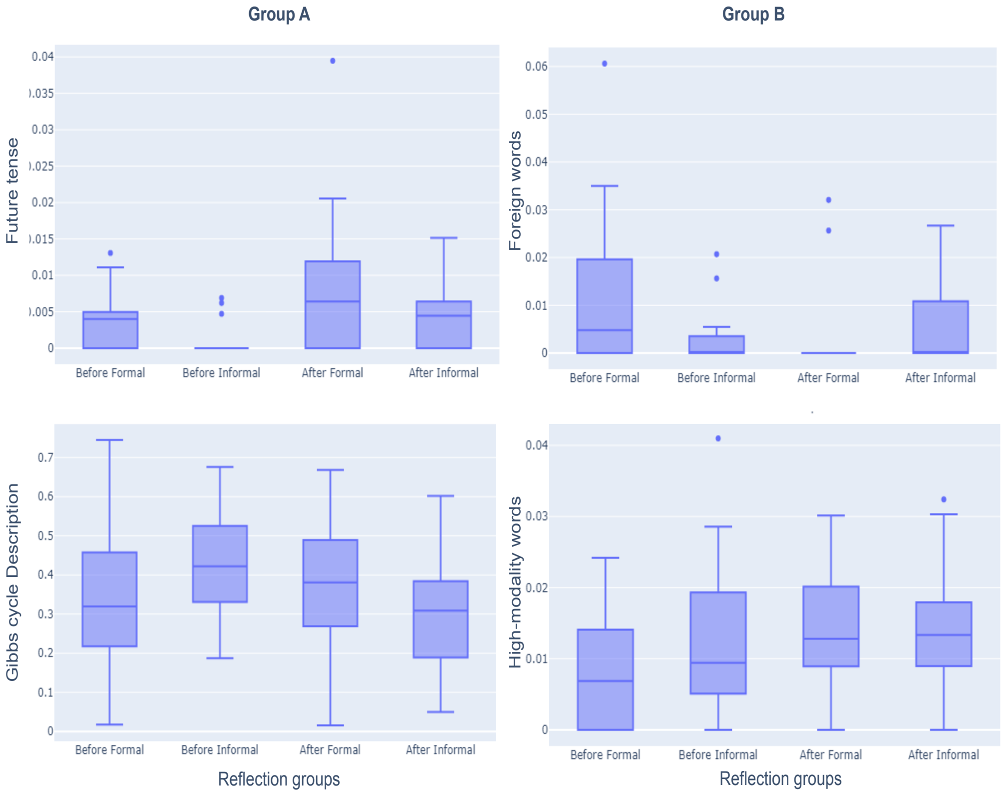


Figure 9.2: The box plot shows the distribution of the most statistically significant features for Group A and Group B.

description and future tense between the first and the second reflections. In the case of Group B, there is a significant difference in the number of foreign words and high-modality words employed between the first and second reflections. As shown in Figure 9.2, the amount of description actually rose in the second reflection, after the formal feedback, while it slightly decreased after receiving the informal feedback. The usage of future tense rose for both groups, showing the overall effect of the feedback. There is a significant decrease in the number of foreign words within Group B's reflections after receiving feedback and a notable increase in the use of "high modality words" among participants in Group B following the feedback they received, although both groups started using more of them in

Table 9.6: A large effect size (d≈0.8), moderate effect size (d≈0.5). When the effect size is positive, it suggests that an increase or change in the variable, in this case, "Gibbs description", has a strong positive influence on the results. A large negative effect size indicates a substantial impact in the opposite direction.

Feature	P value	Effect size
Group A		
Gibbs cycle description	0.02	0.85
Future tense	0.03	-0.82
Group B		
Foreign words	0.02	0.64
High modality words	0.02	-0.70

after first feedback.

If we look at the less statistically significant trends illustrated in Figure 9.3, one can see that, generally, recommendations based on the Gibbs cycle indicators might have had the most influence on the next reflection: "Action plan" grew for both formality groups, while "Conclusion" only raises after informal feedback. The usage of clauses of purpose was also a frequent tip, and indeed, in Group A, participants started using clauses of purpose more frequently, and the length of the clause also grew. Subordinate clauses of reason slightly grew after receiving informal feedback and decreased after the formal one. Finally, all participants expressed more emotions the second time. Counter-intuitively the effect is greater after the formal feedback.

The results of the formality formula show that the average score of the text before formal feedback is 50.077, while after feedback it rises to 50.078, and in the case of informal feedback, the average score before receiving feedback is 50.087, with 50.089 after. In both cases, the difference is insignificant and thus no mimicking effect is apparent from the results. This means that students were not imitating the style of the feedback for their second reflection.

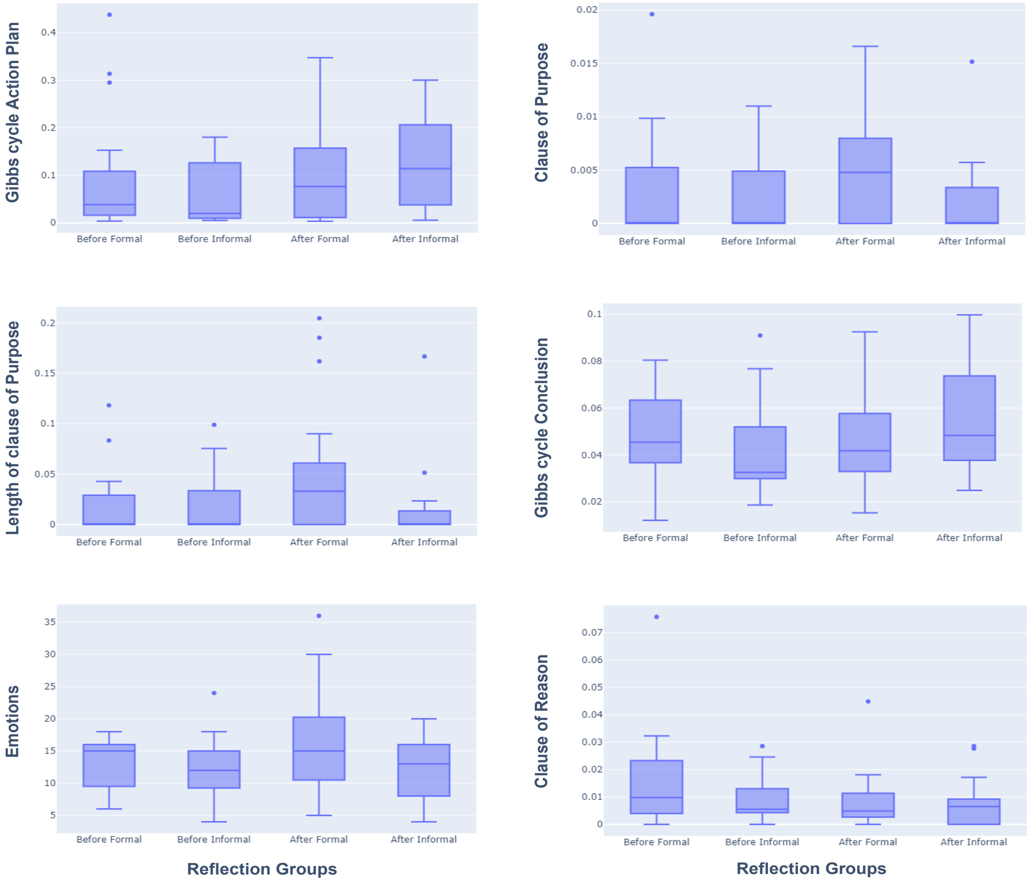


Figure 9.3: The box plot shows the distributions of linguistic features with positive trends, which were not statistically significant.

9.7 Discussion and Conclusion

One significant limitation of both our studies is the lack of participants and often their intrinsic motivation to do this kind of task on an obligatory basis. We also planned Study III, asking students to use the application over a longer period of time, but we were unable to motivate the users and did not gather enough data. Thus, we can mostly speak of the trends, which are still quite promising, and with addition of more participants may prove to draw a positive outcome of the automated feedback on student reflectivity. In the case of study I, we can see no big difference between students' results after human feedback in comparison to the automated one. We can, however, also see no strong improvement after receiving the feedback in general, which may indicate that the feedback format developed was not very useful, which is also reflected by the perceived usefulness scores. We can, however, see that students become slightly more positive about their performance, while the subjectivity level also increased slightly higher for the AI group, and good reflections should possess a high level of subjectivity. In the case of the second study, its major limitation lies in the overly complicated study design with multiple reflections required over a short period of time and survey questions in between, which made it difficult to persuade the users to finish the study and to control that they do not miss any step. English and Spanish texts were automatically translated and processed with German models, where miss-translations may lead to miss-predictions. Ideally, training a multi-lingual model or several language-specific models could yield more accurate results and allow for a cross-linguistic comparison of the features in focus.

The findings do not support any evidence of the user's preference towards certain formality levels, the users even rated the informal feedback as being more formal than the actual formal one. Another study with a higher variety of linguistic indicators may be needed to understand this behaviour. We can, however, see certain changes in reflection indicators between the two reflections. The only indirect indicator to support the users perceiving and mimicking the formality may be the number of foreign words, which increases after informal feedback and decreases after the formal one, and for informal language is it more natural to use more anglicisms e.g. The decrease in “Gibb´s description” result for Group A could be explained by students assuming that the system would have a memory of the previous reflection, and thus less circumstances were described the second time. Additionally, they might have focused more on the suggested cycle stages to improve, and indeed, the “Action Plan” and “Conclusion” categories see a general increase. Users also seem to improve in terms of usage of adverbs, namely high-modality words, which express the degree of possibility, necessity, or likelihood, and similarly to study, I indicate a higher subjectivity level of the essays. Also interestingly, users, after informal feedback, used more clauses of reason, while after formal one, users wrote more clauses of purpose. These clauses of purpose are also longer than the ones present in the initial essays before the feedback was received. Both are positive improvements in terms of quality of reflection. Overall, users are also slightly more emotional in both groups.

Future work should realise the planned Study III, assessing how students improve in reflection over time with a double-blind control/treatment group study, where the control group receives random feedback that looks realistic, and the treatment group receives the real feedback produced by the PapagAI analysis. This would also allow us to determine if students’ interest and engagement may decrease with time as Fryer et al. (2017) suggests.

Logic and mathematics are nothing but specialised linguistic structures.

Jean Piaget

10

Automated Content Moderation Using Transparent Solutions and Linguistic Expertise

The chapter has been removed from the online version for copyright reasons.

10.1 Preface

This Chapter was previously published as: Solopova (2023). Automated content moderation using transparent solutions and linguistic expertise. In E. Elkind (Ed.), Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23 (pp. 7097–7098).: International Joint Conferences on Artificial Intelligence Organization. Doctoral Consortium.

https://doi.org/10.24963/ijcai.2023/823

While the modern NLP trends have long departed from Linguistics as a source of inspiration, one of the purposes of this research was to find and show the use cases where the performance trade-off is minimal and transparency gain is of a particular value. This Chapter can be seen as an "umbrella" publication, unifying, outlining and concluding the research done in this thesis and beyond in the context of using linguistic expertise in Automated Content Moderation. The choice of the hybrid AI architectures in this thesis was often motivated by the interest in applying linguistic knowledge. The result of this quest is both the language quality analysis module in the PapagAI and the SVM models in the Check News in One Click.

This study also includes an outline of my work on Hate Speech detection, which was not included in this thesis since it did not focus on employing a hybrid AI system. At the same time, it gives an overview of the current open-source methods for hate speech detection, both advanced deep learning methods trained on tweets and simple dictionary matching, used by human moderators on Telegram. In this study, we identified that both approaches performed poorly on Telegram data with many errors. Most interestingly, the numerous errors produced by these two approaches did not coincide even once. This again showed that certain hybrid unions of several methods could draw better results in the future.

11

Discussion

11.1 Contributions

In this thesis, I designed and implemented into production two hybrid AI systems in different content moderation applications: social media moderation (namely propaganda detection) and education (reflective essay moderation). Both systems exemplify novel and pioneering open-source approaches, distinguished by our commitment to transparency and interpretability. This is further augmented by a steadfast adherence to ethical AI principles and establishing a controlled environment firmly anchored in theoretical frameworks - Didactics in the case of PapagAI and Linguistics for Check News in One Click. This holds even in times of increasing popularity of generative models, which launch the community into another direction: less explainability, less control, and infringements of data privacy and right of authorship. The collective research presented in this thesis demonstrates the effective coexistence and synergistic relationship between algorithms grounded in linguistic expertise and robust transformer-based classifiers, which can still form productive partnerships. Check News in 1 Click became the first-of-a-kind, researched-based, easy-to-use, multi-language pro-Kremlin propaganda detection tool used by hundreds of users since its launch. The works have already been cited by Maarouf et al. (2023), who developed a human-annotated dataset for detecting online propaganda, by Vanetik et al. (2023), who also works on detecting propaganda in Telegram Posts in the Scope of the Russian Invasion of Ukraine, and by Kloo & Carley (2023), who performed social cybersecurity analysis of the Telegram and Burovova (2023) in the context of the Computational Analysis of Dehumanization of Ukrainians on Russian Telegram.

While PapagAI is the second such system to appear, with many more seemingly underway, it is the first one to implement a wide range of symbolic and sub-symbolic AI models and the first one available for both German and English-speaking users.

Considering the goals I set for my thesis, all of the objectives were at least partially met:

1. In Chapters 2-4 I created models based on linguistic features as well as deep-learning techniques

and evaluated their performance differences with respect to their interpretability. I argued that while in this case, the linguistic model may fall short by approximately 5% in accuracy compared to its deep-learning counterparts, it significantly gains in terms of transparency. However, I acknowledge that the capabilities of explainable AI methods for my deep-learning-based model have not been fully explored or utilized to their maximum potential in this context.

2. I successfully developed, deployed, and conducted user tests for "Check News in One Click" and "PapagAI," employing hybrid AI architectures. I engaged in various ethical considerations throughout their development, acknowledging that these may not be exhaustive. For instance, the issue of cost-sensitive loss functions, which holds relevance to my thesis, was not explored within its scope. However, we addressed similar methods, such as metrics adaptations, in Chapter 8, where we adjusted the Hamming loss to reflect the varying importance of different errors.

   Moreover, there is potential for conducting more comprehensive and prolonged user testing for both applications, which could offer deeper insights into their performance, user experience and especially user-friendliness.

3. Both systems and all of their building components, such as data, models and the code, were made accessible to the public.

Now, let us delve into a more comprehensive examination of the drawbacks mentioned above and the potential improvements they call for.

11.2 Limitations and Future Prospects

Undoubtedly, both implemented systems face several open challenges: Check News in one click is already facing the adaptability problem. The initial study, presented in Chapter 3, which investigated how the model works on the news coming from a one-year-later distribution, was promising. Nonetheless, the system evaluation in a user study presented in Chapter 4 already shows the perpetuated problem of Automated Content Moderation. Variability of genres present within social media, big historical events changing info-space, and the malicious intention of independent individuals and synchronized agents of the information campaigns, who aim to adapt their targeting and harmful content in order to escape platform moderation, and many other factors make social media moderation a challenging task without a human-in-the-loop. That is why recent research shifted towards intelligent helper systems for human fact-checkers and moderators instead of independently acting detection algorithms and sub-subsequent automated policing measure determination. The systems of the new generation find and build a priority list of suspicious claims worth checking, as well as offer humans news with the highest similarity scores of the database of previously fact-checked content.

In the case of propaganda detection, the features reflect the stylistics of manipulative content, which should be more stable over time. At the same, our system evaluation with users shows general polarization and non-neutrality of all content offered by the users, which led to many misclassifications. Hence, in the future, Check New in One Click could benefit from model retraining with the addition of newer data representing a wider range of genres. At the same time, manual annotation or semi-automatic annotation with LLMs could improve label quality. If the technology sees a major improvement in the future, some forms of online learning techniques can undoubtedly be useful when working with news moderation.

A fact-checking module to support manipulation detection, and potentially a group of human moderators supporting the project and building the appropriate database, could be important next steps. The potential for applying such LLMs as Llama 2 (Touvron et al., 2023) for both detection and more variability of the output, should also be investigated, while other user feedback indicates that a visual presentation of the output, e.g. with dynamic charts may be better accepted. Explainability using the attribute method could also be added for the BERT models to highlight the words which were important for the decision.

More linguistic features correlated with manipulative news can be investigated. A multi-modal classification allowing for image processing to recognise deep fakes, detecting old photos used in new contexts and identifying manipulative images would be a strong addition. The scope of the detected propaganda could become wider, as not only the Pro-Kremlin propaganda is of major concern today. While many authoritarian countries consolidate against the democratic world, informational wars are intensifying, so that users would benefit from tools which could identify different origins of propaganda, both internal and external ones. Moreover, in the future, more attention should be given to training models for languages of the so-called "Global South", Latin America, and Central Europe, such as Hungarian, Polish, and Slovakian.

Regarding user-friendliness improvement of the propaganda detection tool, the browser extension has the highest interest score among our participants, even though browser extensions raise security concerns. Nevertheless, they constitute the most accessible and easy way to warn users while browsing.

Finally, the system will benefit from an educational module explaining to a lay user how these models work on a basic computational level. At the same time, as the system is open-source and well-documented through publications, the more the functionalities behind the front end are explained, the more cyber-security of the website against adversarial attacks should be ensured.

As for the PapagAI, one of the significant limitations is scalability, as the AI module cannot process many student requests at once, with the processing time growing exponentially for further students. AWS migration was envisaged and explored to some extent in parallel to the work presented in this thesis. However, this potential solution was not attained, primarily due to financial limitations. Another significant issue arises from the chat-like interface, which inadvertently sets expectations that the system, not yet being a chatbot, struggles to meet. These include rapid responses and the ability to answer questions and react dynamically, along with providing more creative and adaptive feedback. In light of this, a possible solution could be the implementation of a controllable Large Language Model LLM-based chatbot. This could still draw upon the initial system analysis presented in Chapter 8 of this thesis. Alternatively, revising the front-end design is another viable option. Although we originally envisioned moderation through conversational engagement, as previously mentioned, the application of LLM in educational settings introduces specific technical and ethical challenges that need to be addressed. The preferred open-source models require computational power, often unavailable in academic settings. Whereas an API, such as ChatGPT, could indeed solve the scalability, quickness and adaptability problems, it does, at the same time, raise concerns about whether making use of such an application can be made an obligatory and integral part of educational processes. For example, it would send student essays about intimate experiences and emotional struggles to a third-party company which uses this data to train next-generation models. A solution could be to use APIs only to rephrase the feedback produced by the system demonstrated in this thesis and allow for a subsequent conversation with a student. Otherwise, the application can only be used voluntarily. This, in turn, creates problems with the students' intrinsic motivation to do a task that is perceived to create additional work without influencing the final grade, and this, in turn, is a deal-breaker for Teacher Education, where reflection is a central task. The student, refusing to use the application, would still have to write a paper version, which defies the purpose of using the AI to liberate the tutors from the tedious feedback process.

The system implementation of PapagAI could also be improved in many ways. A more robust reflection-level model trained with more data is needed. In particular, more labelled data representing the essays with the highest level of reflection should be added to the training set. The additional annotated data created within the project but not yet used to train the models is not fully significant. The appropriate consent form to use this data to train models was not always requested.

Topic detection is another way of immediate improvement. A generative system to recognise topics more adaptively, from a larger and preferably not pre-defined pool could be the right choice to go. Another important improvement for the back end is to implement a database providing a reflective history of the user, e.g. in the form of feature vectors. It could then be evaluated which parameters improve or degrade over time, which could be an adaptable and personalised addition to the feedback. A summary of previous essays could also be collected in the database and considered for future feedback, for example, using Retrieval Augmented Generation.

Finally, the templates which glue the model outputs together could be produced in higher numbers and with higher variability. As this thesis created a proof of concept and set a goal to evaluate a minimal prototype, it was decided against a broader range of templates to keep the user experience more consistent for the user tests. However, as in the case of Check News in 1 Click, receiving repeatedly similarly formulated feedback may influence user experience and satisfaction.

It should be highlighted that after specific envisaged improvements are made, both systems require more extensive studies with more participants to re-evaluate the applications' effectiveness with statistical significance. This can be done both subjectively with questionnaires and interviews and objectively by answering whether the users became better reflective practitioners in the case of the PapagAI and whether the users learnt to recognise propaganda and fake news better after they started using Check News in 1 Click. More studies are needed to determine how human and automated feedback differ, the long-term consequences for the student's performance, and how to motivate them to do reflective practice because it is practical and not because it is obligatory.

In the end, the hybrid nature of both systems seems advantageous. Even the possible addition of generative components should not make the overall structure obsolete since hybrid building blocks give control and flexibility. Alternatively, both systems can be enriched with other traditional components of hybrid systems: semantic parsing, symbolic reasoners, and ontologies. Evaluation of the logic and argument structure would be useful for both propaganda detection and essay evaluation. Meanwhile, knowledge bases could help both fact-checking and professionalising feedback, giving users professional literature recommendations and specialisation-specific advice.

Recapitulating the aforementioned, both presented system demonstrations provide fruitful areas for future work, and they can grow to become state-of-the-art examples of hybrid AI systems.

12

Conclusion

In this thesis, I presented nine interlinked studies on content moderation and analysis using hybrid AI system infrastructure. The first use case involves social media moderation. It is represented by a multilingual Propaganda detection system using a union of state-of-the-art transformer models and post-hoc explainability through a classical SVM algorithm trained on morpho-syntactic, stylistic and other linguistic features of manipulation. I also showed how this approach is transferable to other social media moderation tasks, namely shit-storm modelling and analysis. I successfully integrated the developed propaganda detection module into the web application "Check News in 1 Click" and conducted a user study. The feedback from users was largely positive, although they recommended several adaptations. Additionally, in order to evaluate the regulation of automated content moderation in the context of the impending European Union legislation, we conducted a thorough analysis of the ethical and legal implications of the European Commission's new proposal on AI.

I developed and implemented a second system focused on reflective essay moderation within the PapagAI application. This AI module employs seven diverse models to analyze each reflective essay, creating a comprehensive profile. A rule-based mechanism then selects the most appropriate suggestions and populates them into a predefined template. User testing revealed that students who received AI-generated feedback performed equally well compared to those who received human guidance. Moreover, the students regarded the AI feedback as credible and as valuable as human-written advice. We also observed certain user improvements already after only one feedback.

In my thesis, I identified and discussed both systems' limitations and proposed directions for future work, highlighting their potential for broader impact. Additionally, I explored innovative ways these systems could be enhanced, examined their scalability, and evaluated how they could be adapted for real-world applications. With this work, I intend to contribute to the popularisation of hybrid AI systems as opposed to end-to-end implementations, which lack sustainability, flexibility and control but also do not seem to make us any closer to the final goal of the field: reproducing the human level intelligence and beyond.

I also hope that Check News in One Click helps foster a sense of personal responsibility for verifying the internet content we consume and the daily usage of AI-based tools for this purpose. I also wish reflective practice becomes a standard one for many more educational fields, creating a new generation of specialists capable of practical and analytical thinking and self-reflection. I am convinced that both skills, which are being advanced by the presented applications, are essential for the future of democratic processes and the ethical evolution of human society.

A

Appendices

A.1 German summary

In dieser Arbeit habe ich neun miteinander verknüpfte Studien zur Moderation und Analyse von Inhalten unter Verwendung einer hybriden KI-Systeminfrastruktur vorgestellt. Der erste Anwendungsfall betrifft die Moderation von sozialen Medien. Es handelt sich um ein mehrsprachiges System zur Erkennung von Propaganda, das eine Kombination aus modernsten Transformationsmodellen und Post-Hoc-Erklärbarkeit durch einen klassischen SVM-Algorithmus verwendet, der auf morphosyntaktische, stilistische und andere linguistische Merkmale der Manipulation trainiert wurde. Ich habe auch gezeigt, wie dieser Ansatz auf andere Aufgaben der Moderation sozialer Medien übertragbar ist, nämlich die Modellierung und Analyse von Shitstorms. Ich implementierte das daraus resultierende KI-Modul in die Webanwendung Check News in 1 Click und testete es in einer Nutzerstudie. Die Nutzer bewerteten die Anwendung positiv, schlugen aber viele Verbesserungen vor. Wir fuhren auch ethische und rechtliche Prüfungen des neuen Vorschlags der Europäischen Kommission zur KI durch, um zu sehen, wie die automatische Moderation von Inhalten unter der kommenden Gesetzgebung in der Europäischen Union funktionieren würde. Der zweite Anwendungsfall betraf die Moderation von reflektierenden Aufsätzen als Teil der PapagAI-Anwendung. Das KI-Modul verwendet 7 verschiedene Modelle unterschiedlicher Art, um ein Profil des gegebenen Essays zu erstellen, woraufhin ein regelbasierter Klassifikator die besten Vorschläge auswählt und sie in eine Vorlage einfügt. Die Benutzertests haben gezeigt, dass die Studierenden, die ein KI-Feedback erhalten haben, genauso gut abgeschnitten haben wie diejenigen, die von Menschen beraten wurden, und dass sie das KI-Feedback für genauso glaubwürdig und nützlich hielten wie das von Menschen geschriebene. Wir haben auch gezeigt, dass die Schüler bereits nach einer Rückmeldung bestimmte Verbesserungen vornehmen.

A.2 Chapter 2

A.2.1 Hyper-parameters of the models

We performed a Grid search and found out that the best results are achieved with Radial basis function kernel, gamma=100, and C=46 parameters.

Our best setup for the neural network with linguistic features was achieved with two hidden layers, a Limited-memory BFGS solver, tanh activation function, and alpha=1e-5.

For the linear model used with BERT, we used a learning rate of 1e-4, four epochs and batch size 16.

A.3 Chapter 3

A.3.1 Additional analysis

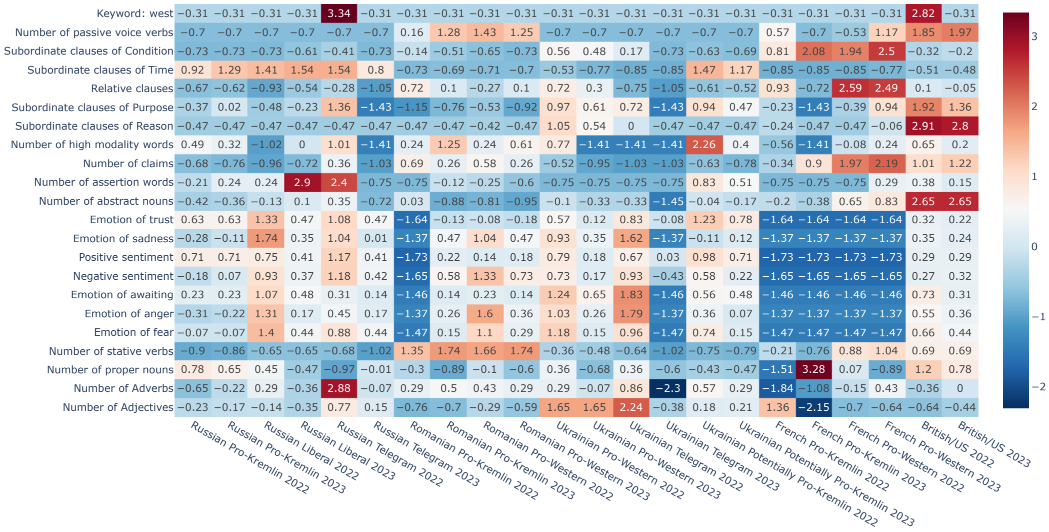
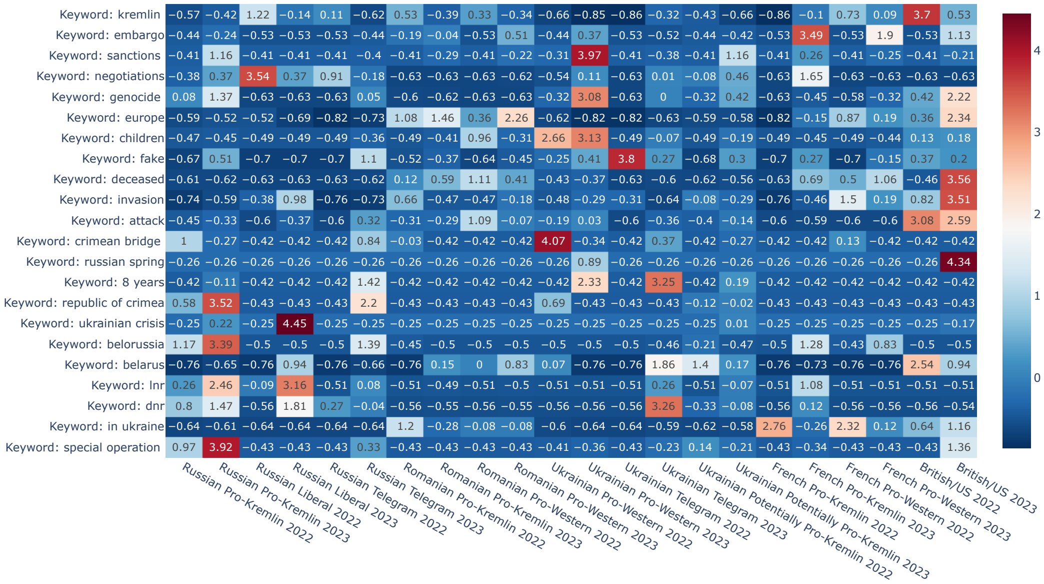


Figure A.1: Normalized averages from the comparative analysis. Linguistic features.



	Russian Pro-Kremlin 2022	Russian Pro-Kremlin 2023	Russian Liberal 2022	Russian Liberal 2023	Russian Telegram 2022	Russian Telegram 2023	Romanian Pro-Kremlin 2022	Romanian Pro-Kremlin 2023	Romanian Pro-Western 2022	Romanian Pro-Western 2023	Ukrainian Pro-Western 2022	Ukrainian Pro-Western 2023	Ukrainian Telegram 2022	Ukrainian Telegram 2023	Ukrainian Potentially Pro-Kremlin 2022	Ukrainian Potentially Pro-Kremlin 2023	French Pro-Kremlin 2022	French Pro-Kremlin 2023	French Pro-Western 2022	French Pro-Western 2023	British/US 2022	British/US 2023
Keyword: kremlin	−0.57	−0.42	1.22	−0.14	0.11	−0.62	0.53	−0.39	0.33	−0.34	−0.66	−0.85	−0.86	−0.32	−0.43	−0.66	−0.86	−0.1	0.73	0.09	3.7	0.53
Keyword: embargo	−0.44	−0.24	−0.53	−0.53	−0.53	−0.44	−0.19	−0.04	−0.53	0.51	−0.44	0.37	−0.53	−0.52	−0.44	−0.42	−0.53	3.49	−0.53	1.9	−0.53	1.13
Keyword: sanctions	−0.41	1.16	−0.41	−0.41	−0.41	−0.4	−0.41	−0.29	−0.41	−0.22	−0.31	3.97	−0.41	−0.38	−0.41	1.16	−0.41	0.26	−0.41	−0.25	−0.41	−0.21
Keyword: negotiations	−0.38	0.37	3.54	0.37	0.91	−0.18	−0.63	−0.63	−0.63	−0.62	−0.54	0.11	−0.63	0.01	−0.08	0.46	−0.63	1.65	−0.63	−0.63	−0.63	−0.63
Keyword: genocide	0.08	1.37	−0.63	−0.63	−0.63	0.05	−0.6	−0.62	−0.63	−0.63	−0.32	3.08	−0.63	0	−0.32	0.42	−0.63	−0.45	−0.58	−0.32	0.42	2.22
Keyword: europe	−0.59	−0.52	−0.52	−0.69	−0.82	−0.73	1.08	1.46	0.36	2.26	−0.62	−0.82	−0.82	−0.63	−0.59	−0.58	−0.82	−0.15	0.87	0.19	0.36	2.34
Keyword: children	−0.47	−0.45	−0.49	−0.49	−0.49	−0.36	−0.49	−0.41	0.96	−0.31	2.66	3.13	−0.49	−0.07	−0.49	−0.19	−0.49	−0.45	−0.49	−0.44	0.13	0.18
Keyword: fake	−0.67	0.51	−0.7	−0.7	−0.7	1.1	−0.52	−0.37	−0.64	−0.45	−0.25	0.41	3.8	0.27	−0.68	0.3	−0.7	0.27	−0.7	−0.15	0.37	0.2
Keyword: deceased	−0.61	−0.62	−0.63	−0.63	−0.63	−0.62	0.12	0.59	1.11	0.41	−0.43	−0.37	−0.63	−0.6	−0.62	−0.56	−0.63	0.69	0.5	1.06	−0.46	3.56
Keyword: invasion	−0.74	−0.59	−0.38	0.98	−0.76	−0.73	0.66	−0.47	−0.47	−0.18	−0.48	−0.29	−0.31	−0.64	−0.08	−0.29	−0.76	−0.46	1.5	0.19	0.82	3.51
Keyword: attack	−0.45	−0.33	−0.6	−0.37	−0.6	0.32	−0.31	−0.29	1.09	−0.07	−0.19	0.03	−0.6	−0.36	−0.4	−0.14	−0.6	−0.59	−0.6	−0.6	3.08	2.59
Keyword: crimean bridge	1	−0.27	−0.42	−0.42	−0.42	0.84	−0.03	−0.42	−0.42	−0.42	4.07	−0.34	−0.42	0.37	−0.42	−0.27	−0.42	−0.42	0.13	−0.42	−0.42	−0.42
Keyword: russian spring	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	0.89	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	−0.26	4.34
Keyword: 8 years	−0.42	−0.11	−0.42	−0.42	−0.42	1.42	−0.42	−0.42	−0.42	−0.42	−0.42	2.33	−0.42	3.25	−0.42	0.19	−0.42	−0.42	−0.42	−0.42	−0.42	−0.42
Keyword: republic of crimea	0.58	3.52	−0.43	−0.43	−0.43	2.2	−0.43	−0.43	−0.43	−0.43	0.69	−0.43	−0.43	−0.43	−0.12	−0.02	−0.43	−0.43	−0.43	−0.43	−0.43	−0.43
Keyword: ukrainian crisis	−0.25	0.22	−0.25	4.45	−0.25	−0.25	−0.25	−0.25	−0.25	−0.25	−0.25	−0.25	−0.25	−0.25	−0.25	0.01	−0.25	−0.25	−0.25	−0.25	−0.25	−0.17
Keyword: belorussia	1.17	3.39	−0.5	−0.5	−0.5	1.39	−0.45	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.46	−0.21	−0.47	−0.5	1.28	−0.43	0.83	−0.5	−0.5
Keyword: belarus	−0.76	−0.65	−0.76	0.94	−0.76	−0.66	−0.76	0.15	0	0.83	0.07	−0.76	−0.76	1.86	1.4	0.17	−0.76	−0.73	−0.76	−0.76	2.54	0.94
Keyword: lnr	0.26	2.46	−0.09	3.16	−0.51	0.08	−0.51	−0.49	−0.51	−0.5	−0.51	−0.51	−0.51	0.26	−0.51	−0.07	−0.51	1.08	−0.51	−0.51	−0.51	−0.51
Keyword: dnr	0.8	1.47	−0.56	1.81	0.27	−0.04	−0.56	−0.55	−0.56	−0.55	−0.56	−0.56	−0.56	3.26	−0.33	−0.08	−0.56	0.12	−0.56	−0.56	−0.56	−0.54
Keyword: in ukraine	−0.64	−0.61	−0.64	−0.64	−0.64	−0.64	1.2	−0.28	−0.08	−0.08	−0.6	−0.64	−0.64	−0.59	−0.62	−0.58	2.76	−0.26	2.32	0.12	0.64	1.16
Keyword: special operation	0.97	3.92	−0.43	−0.43	−0.43	0.33	−0.43	−0.43	−0.43	−0.43	−0.41	−0.36	−0.43	−0.23	0.14	−0.21	−0.43	−0.34	−0.43	−0.43	−0.43	1.36

Figure A.2: Normalized averages from the comparative analysis. Keywords.

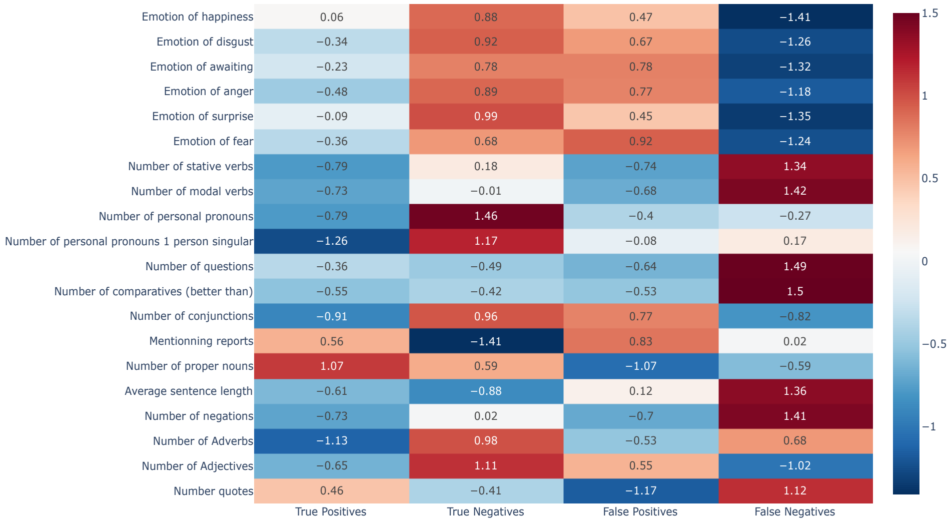


Figure A.3: Error analysis. Normalized averages of linguistic features for the groups of errors.

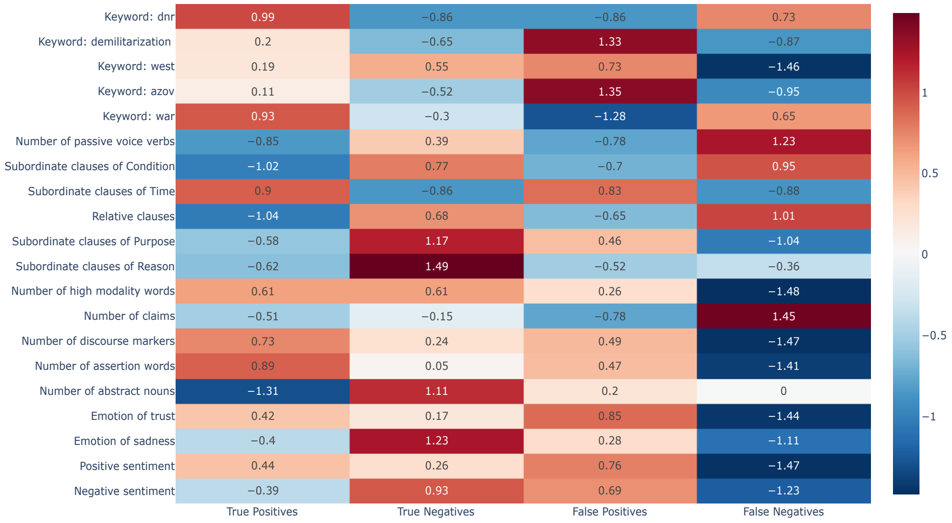


Figure A.4: Error analysis. Normalized averages of keyword occurrences for the groups of errors.

A.4 Chapter 4

A.4.1 User survey questions

- I voluntarily give my permission for my answers to be used for improvement in the areas of study. All information is confidential and anonymous. | Yes, No.
- The username you used while using the App | free form
- Your nationality | free form
- What political views do you have? | left, moderate left, centrist, moderate right, right
- How much news did you check? | free form
- Did you like the keyword explanation? | Yes, No, No sure
- Did you like the linguistic explanation? | Yes, No, No sure
- If you can tell, did you find the output accurate? | Accurate, Can't tell, Not accurate (bad prediction), Not accurate (bad explanation)
- On a scale of 5 (with five being very useful), how useful did you find the tool? | 1,2,3,4,5
- Did you learn something useful about how to detect manipulations? | Yes, No, Not sure
- What else would you want to have in such an app/what would you change? | free form
- Would you continue checking your news with such a tool? | Yes, No, Not Sure
- Would you recommend someone such a tool? | Yes - to friends, Yes - to peers/colleagues, Yes - to older relatives, Yes - to younger acquaintances (teenagers)

- In which form would you rather have such a tool? | Browser extension, Website as it is, Desktop program, Mobile app.

A.4.2 Linguistic indicators per language

- **English**

  Frequent usage of quotations may indicate fake news.

  Mentioning statistics a lot is frequent in fake news.

  A large number of adjectives indicates non-neutral news.

  A large number of adverbs indicates non-neutral news.

  A lot of negations are associated with pro-Russian news.

  Frequent usage of verbs of state is more typical for pro-Western, pro-Ukrainian news.

  Frequent mention of reports or surveys may indicate fake news.

  The usage of comparative or superlative adjectives indicates non-neutral news.

  Usage of questions in news may indicate fake news or whataboutisms.

  This piece of news contains opinions.

  The emotional lexicon shows non-neutral news.

  High usage of abstract nouns (e.g. freedom, liberty) is associated with pro-Western news.

  A high number of claims is associated with pro-Western news.

  Usage of high modality words (e.g. obviously, certainly) adjectives indicates non-neutral news.

  A high number of connectors (e.g. finally, firstly) is typical for pro-Russian news.

  Subordinate Clauses of Concession are markers of pro-Russian news.

  A high number of subordinate clauses of reason is associated with pro-Western news.

  Subordinate Clauses of Purpose are markers of pro-Russian news.

  Subordinate Clauses of Time are markers of pro-Russian news. They justify their actions through temporal references.

A high number of subordinate clauses of the condition is associated with pro-Western news. Pro-Russian news does not think in conditions but rather presents their vision of the future as an inevitable fact.

A.4.3 Best training parameters

The models were trained on 1 RTX5000, with 3 hours each. German and Italian BERT model parameters: “layer_norm_eps”: 1e-12, “num_hidden_layers”: 12, “pad_token_id”: 0, “vocab_size”: 30000, “learning_rate”: 4e-05, “batch size”: 8, “epochs”: 4, 80/20% train/test split.

A.5 Chapter 5

A.5.1 A

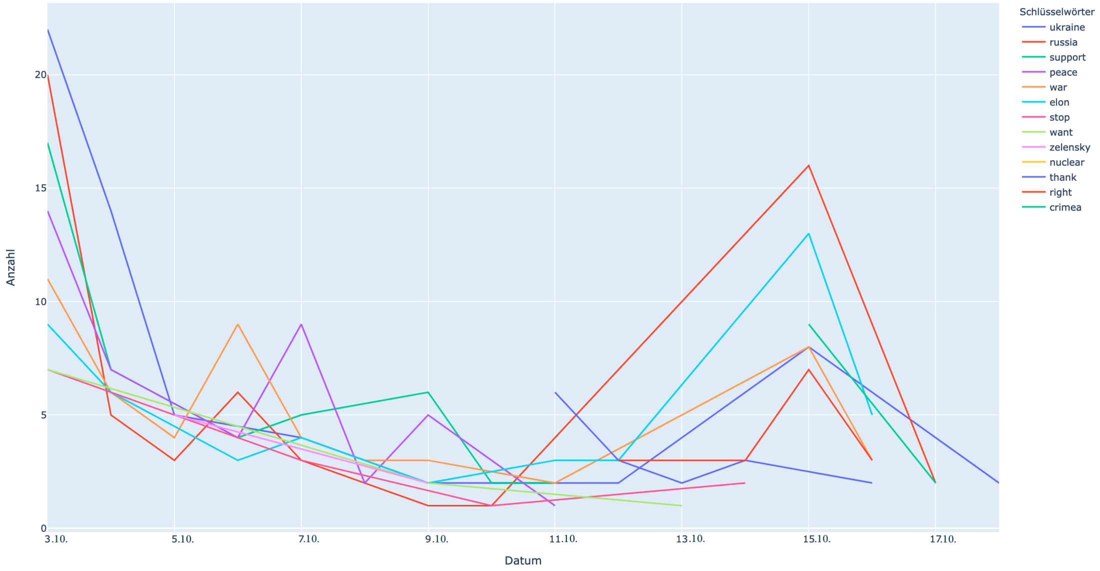


Figure A.5: Elon Musk shitstorm: time distribution of keywords on Twitter. Time course of the Elon Musk shitstorm on Twitter and Telegram.

A.5.2 B

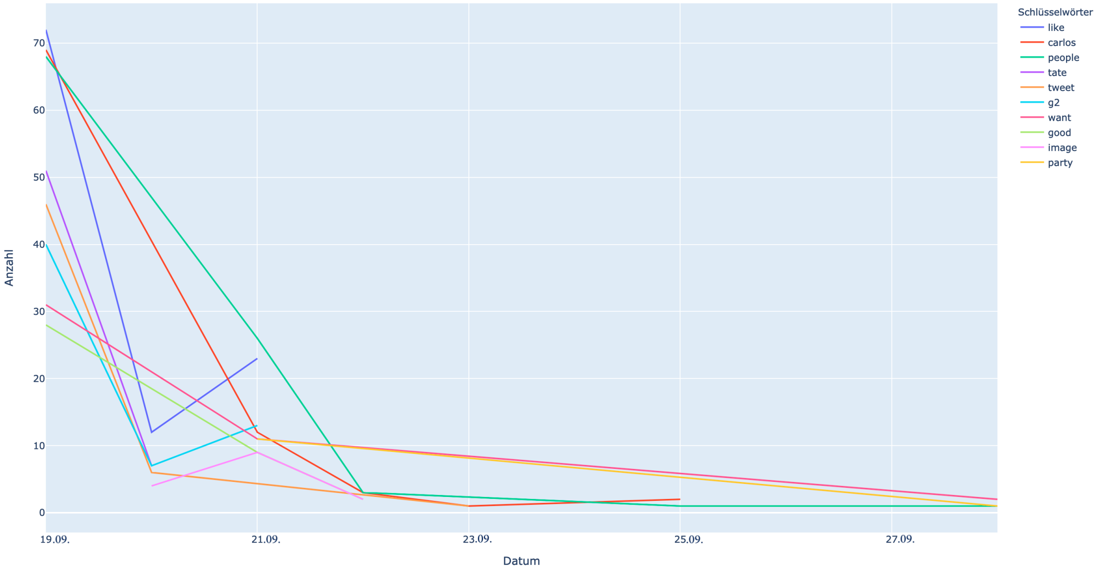


Figure A.6: eSports shitstorm: temporal distribution of keywords on Twitter.

A.5.3 C

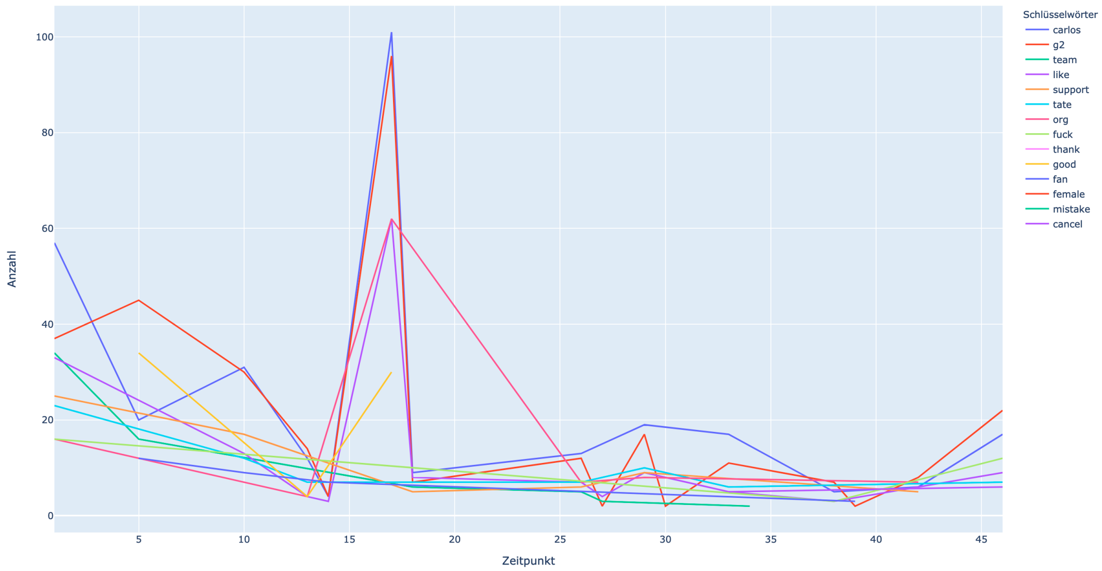


Figure A.7: eSports shitstorm: temporal distribution of keywords on Reddit

A.6 Chapter 7

A.6.1 Data collection

The entire questionnaire, including the consent form, the code for linguistic feature annotation and the data set divided into training and test sets for benchmark purposes, are available on the OSF depository: https://osf.io/ug9r8/ and Github: https://github.com/oanaucs/german_reflective_corpus.

A.6.2 Guided reflection questions (German)

1. Bitte denken Sie an die Erfahrung die Sie während der Aufgabenlösung gemacht haben - aus Ihrer Perspektive. Wer war dabei, was haben Sie gelöst, wann und wo? Erklären Sie bitte welche Entscheidungen und warum Sie sie getroffen haben. Bitte schreiben Sie vollständige Sätze.

2. Bitte reflektieren Sie über das Gelernte durch die Aufgabenlösung. Was haben Sie gelernt? Sind Sie selbstbewusster geworden? Werden Sie das Gelernte in der Praxis anwenden? Was haben Sie vor? Was hätten Sie besser machen können? Bitte schreiben Sie vollständige Sätze.

3. Bitte denken Sie jetzt an die Schwierigkeiten die während der Aufgabenlösung aufgetaucht sind. Was waren die Herausforderungen? Ist etwas unerwartetes passiert? Haben Ihre vorherige Annahmen (z.B. Zeit für die Aufgabe) doch nicht gestimmt? Bitte schreiben Sie vollständige Sätze.

4. Erklären Sie bitte wie Ihre Wahrnehmung gegenüber das Thema verändert hat. Bitte schreiben Sie vollständige Sätze.

5. Erklären Sie bitte wie Ihre Wahrnehmung gegenüber Ihre Kompetenzen verändert hat. Bitte schreiben Sie vollständige Sätze.

6. Erklären Sie bitte wie sich während und nach der Aufgabenlösung gefühlt haben. Welche Emotionen haben Sie erlebt? Wie haben sich Ihre persönliche Überzeugungen verändert? Bitte schreiben Sie vollständige Sätze.

A.6.3 Guided reflection questions (English)

1. Please think about the experience you had while solving the task - from your perspective. Who was there, what did you solve, when and where? Please explain your decisions and why you made them. Please write complete sentences.

2. Please reflect on what you have learned through the assignment. What was new? Have you become more confident? Will you apply what you have learned in practice? What do you plan to do? What could you have done better? Please write complete sentences.

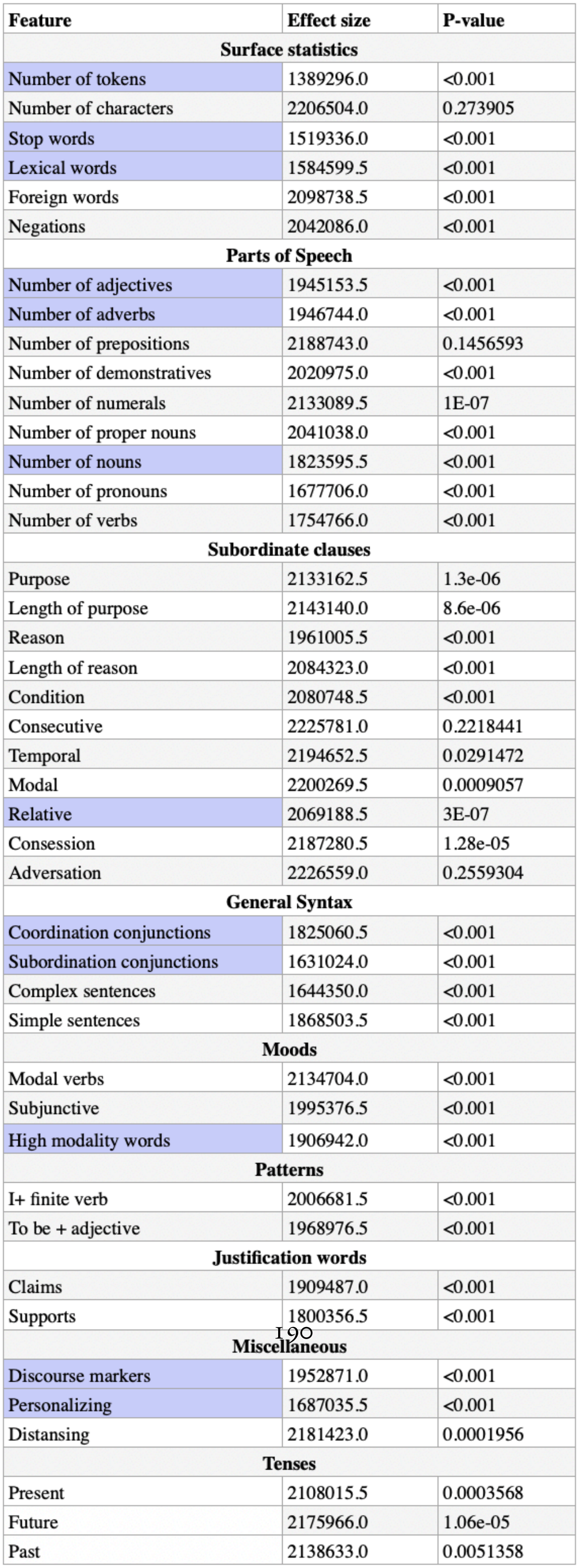
3. Please think now about the difficulties that arose during the task solution. What were the challenges? Did something unexpected happen? Were your previous assumptions (e.g., time for the task) not correct after all? Please write complete sentences.

4. Please explain how your perception towards the subject has changed. Please write complete sentences.

5. Please explain how your perception towards your competencies has changed. Please write complete sentences.

6. Please explain how you felt during the task and after solving it. What emotions did you experience? How did your personal beliefs change? Please write complete sentences.

Table A.1: Linguistic features. The coloured features are the most relevant ones, according to our analysis.



Feature	Effect size	P-value
Surface statistics		
Number of tokens	1389296.0	<0.001
Number of characters	2206504.0	0.273905
Stop words	1519336.0	<0.001
Lexical words	1584599.5	<0.001
Foreign words	2098738.5	<0.001
Negations	2042086.0	<0.001
Parts of Speech		
Number of adjectives	1945153.5	<0.001
Number of adverbs	1946744.0	<0.001
Number of prepositions	2188743.0	0.1456593
Number of demonstratives	2020975.0	<0.001
Number of numerals	2133089.5	1E-07
Number of proper nouns	2041038.0	<0.001
Number of nouns	1823595.5	<0.001
Number of pronouns	1677706.0	<0.001
Number of verbs	1754766.0	<0.001
Subordinate clauses		
Purpose	2133162.5	1.3e-06
Length of purpose	2143140.0	8.6e-06
Reason	1961005.5	<0.001
Length of reason	2084323.0	<0.001
Condition	2080748.5	<0.001
Consecutive	2225781.0	0.2218441
Temporal	2194652.5	0.0291472
Modal	2200269.5	0.0009057
Relative	2069188.5	3E-07
Consession	2187280.5	1.28e-05
Adversation	2226559.0	0.2559304
General Syntax		
Coordination conjunctions	1825060.5	<0.001
Subordination conjunctions	1631024.0	<0.001
Complex sentences	1644350.0	<0.001
Simple sentences	1868503.5	<0.001
Moods		
Modal verbs	2134704.0	<0.001
Subjunctive	1995376.5	<0.001
High modality words	1906942.0	<0.001
Patterns		
I+ finite verb	2006681.5	<0.001
To be + adjective	1968976.5	<0.001
Justification words		
Claims	1909487.0	<0.001
Supports	1800356.5	<0.001
Miscellaneous		
Discourse markers	1952871.0	<0.001
Personalizing	1687035.5	<0.001
Distansing	2181423.0	0.0001956
Tenses		
Present	2108015.5	0.0003568
Future	2175966.0	1.06e-05
Past	2138633.0	0.0051358

A.7 Chapter 8

Table A.2: Metrics mentioned in the paper.

Metric	Definition
F1-score	A harmonic mean of the precision and recall calculated per class. Can range from 0 to 1.
F1-score macro	The metric is computed independently for each class, and then the average is taken.
F1-score micro	The metric aggregates the contributions of all classes to compute the average metric.
Cohen's kappa	The metric is used to measure inter-annotator reliability for categorical items. 0.41–0.60 is interpreted as moderate agreement, 0.61–0.80 as substantial, and 0.81–1.00 as perfect agreement.
QWK	Quadratic Weighted Kappa measures the agreement between two outcomes ranging from -1 (complete disagreement) to 1 (complete agreement).
Hamming score	The metric is often used for multi-label classification, calculating the fraction of wrong labels to the total number of labels. The values higher than 0.9 are excellent scores, higher than 0.7 are good scores, and lower than 0.7 may be considered poor.

Table A.3: Topics clusters from Bertopic.

Clustering 1	Clustering 2
Lectures and editing	Teamwork and Tasks
Classroom Management	Teacher, school, teaching
Pedagogy and Educational Diagnostics	Algorithms, Computer Science, Digital Technology
Reading and Literature	Self-promotion
Conflict Analysis	Music
Feedback	Math and numeracy
Your Subject Area	Science and Experiments
Diagnostics and diagnostic procedures	
Intervention measure	
Motivation	
Portfolio	
Lecture material and video	
Psychology	

Table A.4: Emotion detection labels.

Emotions & Feelings
information
annoyance
appreciation
disapproval/critique
interest
anticipation
excitement
challenged
confidence
disappointment
insecurity
motivation
optimism
responsibility
satisfaction
surprise
uncertainty
wariness

Table A.5: Defenitions of reflective labels.

Level	Definition
Description	It is the lowest level, where the person only describes the circumstances and may include an evaluation of their own feelings.
Reflective description	Here one's own perspective analysis and superficial justifications are present.
Dialogic Reflection	It includes analysis of various perspectives as if in the form of an internal dialogue with oneself.
Transformative Reflection	It should include the plan for the next steps or what one would do next time in such a situation.
Critical Reflection	The highest level of reflection encompasses a wider context (social, political, historical).

A.8 Additional tools

When editing this thesis, I used Grammarly[1] for spellchecking and ChatGPT[2] to rephrase the wording of certain sentences, but it was not used to produce the actual content of the thesis. I also used Bibtex[3] online citation converter to produce many of the referenced citations.

[1] https://app.grammarly.com

[2] https://chat.openai.com

[3] https://www.bibtex.com

References

Abarna, S., Sheeba, J., Jayasrilakshmi, S., & Devaneyan, S. P. (2022). Identification of cyber harassment and intention of target users on social media platforms. *Engineering Applications of Artificial Intelligence*, 115, 105283.

Adriani, R. (2019). The evolution of fake news and the abuse of emerging technologies. *European Journal of Social Sciences*, 2, 32–38.

Ahirwar, R., Ajay, M., Sathyabalan, N., & Lakshmi, K. (2022). Online harassment detection using machine learning. In *2022 International Conference on Inventive Computation Technologies (ICICT)* (pp. 1222–1224).

Alam, F., Mubarak, H., Zaghouani, W., Da San Martino, G., & Nakov, P. (2022). Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)* (pp. 108–118). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Allas, R., Äli Leijen, & Toom, A. (2020). Guided reflection procedure as a method to facilitate student teachers' perception of their teaching to support the construction of practical knowledge. *Teachers and Teaching*, 26(2), 166–192.

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14), eaay3539.

Almuttalibi, N. A. Y. (2023). Social media and language evolution : A review of current theoretical efforts on communication and language change. *Nasaq Journal*, 39, 1355–1361.

Alsmadi, I. & O'Brien, M. (2020). How many bots in russian troll tweets? *Information Processing & Management*, 57, 102303.

Ananthakrishnan, U. M. & Tucker, C. E. (2021). The drivers and virality of hate speech online. *Available at SSRN 3793801*.

Anderson, A. D. (2018). *The Sources of Russian Information Warfare*. Technical Report AD1098323, AIR UNIV MAXWELL AFB AL MAXWELL AFB United States. Pagination or Media Count: 103.

Annuš, N. (2023). Chatbots in education: The impact of artificial intelligence based chatgpt on teachers and students. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(4), 366–370.

Antoun, W., Baly, F., Achour, R., Hussein, A., & Hajj, H. (2020). State of the art models for fake news detection tasks. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)* (pp. 519–524).

Arango, A., Pérez, J., & Poblete, B. (2020). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, (pp. 101584).

Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4623–4637).

Asr, F. T. & Taboada, M. (2019). Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1), 2053951719843310.

Axelsson, M. & Berglund, Y. (2002). The uppsala student english corpus (use): a multi-faceted resource for research and course development. *Language and Computers*, (pp. 79–90).

Bajaj, M. & Li, J. (2020). Students, faculty express concerns about online exam invigilation amidst COVID-19 outbreak. https://www.ubyssey.ca/news/Students-express-concerns-about-online-exams/. Accessed: June 11, 2023.

Baker, R. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, 26.

Barendt, E. (2019). What is the harm of hate speech? *Ethical Theory and Moral Practice*, 22(3), 539–553.

Barnes, R., Cooper, A., Kolkman, O., Thaler, D., & Nordmark, E. (2016). Technical considerations for internet service blocking and filtering. *Request for Comments (RFC)*, 7754.

Barrón-Cedeño, A., Martino, G. D. S., Jaradat, I., & Nakov, P. (2019). Proppy: A system to unmask propaganda in online news. In *AAAI Conference on Artificial Intelligence*.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation* Stroudsburg, PA, USA: Association for Computational Linguistics.

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*.

Batbaatar, E., Li, M., & Ryu, K. H. (2019). Semantic-emotion neural network for emotion recognition from text. *IEEE Access*, 7, 111866–111878.

Bauer, N., Holla, K., Westhues, S., & Wiemer, P. (2016). Streiten 2.0 im Shitstorm-Eine exemplarische analyse sprachlicher profilierungsmuster im sozialen netzwerk facebook. *Reihe XII*.

Becker, A. (2021). 83 Prozent der Studenten brechen Lehramts-Studium ab. *Nordkurier*.

Bellhäuser, H., Dignath, C., & Theobald, M. (2023). Daily automated feedback enhances self-regulated learning: a longitudinal randomized field experiment. *Frontiers in Psychology*, 14.

Bendel, K., Menger, N., & Eick ehem. Skottke, E.-M. (2016). *Analyse der Wahrnehmung von Shit- und Candystorms mittels Eye-tracking: Über die Grenzen der Kommunikation in Social Media*, (pp. 85–108). Nomos.

Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. *Commun. ACM*, 64(7), 58–65.

Benotti, L., Martnez, M. C., & Schapachnik, F. (2017). Atool for introducing computer science with automatic formative assessment. *IEEE Transactions on Learning Technologies*, 11(2), 179–192.

Bergengruen, V. (2022). How telegram became the digital battlefield in the russia-ukraine war. https://time.com/6158437/telegram-russia-ukraine-information-war. The Time. Accessed: March 25, 2022.

Bernabeu, P. & Vogt, P. (2015). Language evolution: Current status and future directions. In *Language at the University of Essex (LangUE)*.

Bernard, T. (2023). The evolving trust and safety vendor ecosystem. https://techpolicy.press/the-evolving-trust-and-safety-vendor-ecosystem/. Accessed: September 11, 2023.

Beskow, D. & Carley, K. (2020). *Disinformation, Misinformation, and Fake News in Social Media (ISBN 978-3-030-42698-9). Characterization and Comparison of Russian and Chinese Disinformation Campaigns*, (pp. 63–81). Springer.

Bhatt, P. & Rios, A. (2021). Detecting bot-generated text by characterizing linguistic accommodation in human-bot interactions.

Biggs, J. B. & Collis, K. F. (1982). The psychological structure of creative writing. *Australian Journal of Education*, 26, 59 – 70.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7–74.

Blank, S. (2022). Russia, china, and information war against ukraine. *The Journal of East Asian Affairs*, 35(2), 39–72.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res.*, 13(6), 4–16.

Bodaghi, A., Oliveira, J., & Zhu, J. J. (2021). The fake news graph analyzer: An open-source software for characterizing spreaders in large diffusion graphs. *Software Impacts*, 10, 100182.

Boerboom, C. (2020). Cambridge analytica: The scandal on data privacy. *Augustana Center for the Study of Ethics Essay Contest*.

Bokša, M. (2019). Russian information warfare in central and eastern europe: Strategies, impact, countermeasures. Rethink.CEE Fellowship, The German Marshall Fund of the United States. Policy Paper No. 15.

Boldyreva, E. (2018). Cambridge analytica: Ethics and online manipulation with decision-making process. In *18th PCSF 2018 - Professional Culture of the Specialist of the Future* (pp. 91–102).

Bozarth, L. & Budak, C. (2020). Toward a better performance evaluation framework for fake news classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 60–71.

Braun, J. A. & Eklund, J. L. (2019). Fake news, real money: Ad tech platforms, profit-driven hoaxes, and the business of journalism. *Digital Journalism*, 7(1), 1–21.

Bredeweg, B. & Kragten, M. (2022). Requirements and challenges for hybrid intelligence: A case-study in education. *Frontiers in Artificial Intelligence*, 5.

Breiman, L. (2004). Random forests. *Machine Learning*, 45, 5–32.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners.

Bubnys, R. (2019). A journey of self-reflection in students' perception of practice and roles in the profession. *Sustainability*, 11, 194.

Bundesministerium des Justiz und für Verbraucherschutz (2017). Gesetz zur verbesserung der rechtsdurchsetzung in sozialen netzwerken (netzwerkdurchsetzungsgesetz - netzdg).

Bundesministerium für Digitalisierung und Wirtschaftsstandort (2020). Bundesgesetz über maßnahmen zum schutz der nutzer auf kommunikationsplattformen.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1–15.

Burnap, P. & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy Internet*, 7(2), 223–242.

Burovova, K. (2023). The 4th stage of genocide: Computational analysis of dehumanization of ukrainians on russian telegram. https://er.ucu.edu.ua/handle/1/3941. Electronic repository of The Ukrainian Catholic University.

Butler, D. & Winne, P. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research - REV EDUC RES*, 65, 245–281.

Carroll, J. (2017). Image and imitation the visual rhetoric of pro-russian propaganda. *Ideology and Politics Journal*, 8, 36–79.

Cevher, D., Zepf, S., & Klinger, R. (2019). Towards multimodal emotion recognition in german speech events in cars using transfer learning.

Chaiprasurt, C., Amornchewin, R., & Kunpitak, P. (2022). Using motivation to improve learning achievement with a chatbot in blended learning. *World Journal on Educational Technology: Current Issues*, 14, 1133–1151.

Chan, M. (2016). This YouTube star got sued, raised $130,000, and wants to change the site forever. *Time*.

Chan, R. (2019). The cambridge analytica whistleblower explains how the firm used facebook data to sway elections. *Business Insider*.

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2.

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial intelligence trends in education: a narrative overview. *Procedia Computer Science*, 136, 16–24.

Chee, F. Y. (2017). Nato says it sees sharp rise in russian disinformation since crimea seizure. *Reuters*.

Chen, A. (2017). The human toll of protecting the internet from the worst of humanity. *New Yorker*.

Chen, H., Ciborowska, A., & Damevski, K. (2019). Using automated prompts for student reflection on computer security concepts. In *ITiCSE '19: Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 506–512).

Chen, L., Zaharia, M., & Zou, J. (2023). How is chatgpt's behavior changing over time?

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16 (pp. 1–5). New York, NY, USA: Association for Computing Machinery.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* New York, NY, USA: ACM.

Chiorrini, A., Diamantini, C., Mircoli, A., & Potena, D. (2021). Emotion and sentiment analysis of tweets using bert. In *EDBT/ICDT Workshops*.

Chiu, T. K., Moorhouse, B., Chai, C., & Ismailov, M. (2023). Teacher support and student motivation to learn with artificial intelligence (ai) based chatbot. *Interactive Learning Environments*.

Chudinov, A., Koshkarova, N., & Ruzhentseva, N. (2019). Linguistic interpretation of russian political agenda through fake, deepfake, post-truth. *Journal of Siberian Federal University. Humanities & Social Sciences*, (pp. 1840–1853).

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*: OpenReview.net.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, (pp. 37–46).

Combe, A. & Kolbe, F.-U. (2004). Lehrerprofessionalität: Wissen, können, handeln. In *Handbuch der Schulforschung* (pp. 833–851). Springer.

Comission, E. (2021). Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. Accessed: April 21, 2021.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Cosentino, G. (2020a). Polarize and conquer: Russian influence operations in the united states. in social media and the post-truth world order. *Springer*, (pp. 33–57).

Coster, H. (2022). More people are avoiding the news, and trusting it less, report says. https://www.reuters.com/business/media-telecom/more-people-are-avoiding-news-trusting-it-less-report-says-2022-06-14/. Reuters. Accessed: August 28, 2023.

Council of European Union (2008). Council framework decision (EU) no 913/jha.

Crawford, K. & Calo, R. (2016). There is a blind spot in AI research. *Nature*, 538(7625), 311–313.

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., & Nakov, P. (2020a). SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1377–1414). Barcelona (online): International Committee for Computational Linguistics.

Da San Martino, G., Shaar, S., Zhang, Y., Yu, S., Barrón-Cedeño, A., & Nakov, P. (2020b). Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 287–293). Online: Association for Computational Linguistics.

Dadu, T., Pant, K., & Mamidi, R. (2020). Towards detection of subjective bias using contextualized word embeddings. *CoRR*, abs/2002.06644.

Daniel Boffey, D. S. & Borger, J. (2022). Mariupol theatre bombing killed 300, ukrainian officials say. www.theguardian.com/world/2022/mar/25/mariupol-theatre-bombing-killed-300-ukrainian\-officials-say. The Guardian. Accessed: March 25, 2022.

Das, S., Saha, S., & Srihari, R. (2022). Diving deep into modes of fact hallucinations in dialogue systems. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 684–699). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512–515.

De Cao, N., Schlichtkrull, M., Aziz, W., & Titov, I. (2020). How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3243–3255).

De Gregorio, G. (2021). The rise of digital constitutionalism in the european union. *International Journal of Constitutional Law*, 19(1), 41–70.

De Lin, O., Gottipati, S., Ling, L. S., & Shankararaman, V. (2021). Mining informal & short student self-reflections for detecting challenging topics – a learning outcomes insight dashboard. In *2021 IEEE Frontiers in Education Conference (FIE)* (pp. 1–9).

Delcker, J. (2022). Russian disinformation looms large over german winter. https://www.dw.com/en/russian-disinformation-threat-looms-large-over-cold-german-winter/a-63096336. Deutsche Welle. Accessed: September 12, 2022.

Deremetz, A. & Scheffler, T. (2020). Die retribalisierung der gesellschaft? *Z. Kult.- Kollekt.*, 6(2), 171–216.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dibbell, J. (1994). A rape in cyberspace; or, how an evil clown, a haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. In *Flame Wars* (pp. 237–262). Duke University Press.

Donald, A. (1983). *The reflective practitioner: How professionals think in action*. Basic books.

Duarte, N., Llanso, E., & Loup, A. (2017). *Mixed messages? The limits of automated social media content analysis*. Center for Democracy & Technology.

Dyment, J. E. & O'connell, T. S. (2010). The quality of reflection in student journals: A review of limiting and enabling factors. *Innovative Higher Education*, 35, 233–244.

Ebel, F. (2022). Putin admits attacks on civilian infrastructure, asking: 'who started it?'. *The Washington Post*.

Ekman, P. (2023). Basic emotions. *andbook of cognition and emotion*, 98, 16.

Elands, P., Huizing, A., Kester, J., Peeters, M. M. M., & Oggero, S. (2019). Governing ethical and effective behaviour of intelligent systems: A novel framework for meaningful human control in a military context. *Militaire Spectator*, 188(6), 302–313.

Elswah, M. & Howard, P. N. (2020). "Anything that Causes Chaos": The Organizational Behavior of Russia Today (RT). *Journal of Communication*, 70(5), 623–645.

Ende, L., Reinhard, M.-A., & Göritz, L. (2023). Detecting greenwashing! the influence of product colour and product price on consumers' detection accuracy of faked bio-fashion. *J. Consumer Policy*, 46(2), 155–189.

Fabbri, A., Lai, A., Grundy, Q., & Bero, L. A. (2018). The influence of industry sponsorship on the research agenda: A scoping review. *Am. J. Public Health*, 108(11), e9–e16.

Faulconbridge, G. (2022a). Putin escalates ukraine war, issues nuclear threat to west. *Reuters*.

Faulconbridge, G. (2022b). Russia fights back in information war with jail warning. *Reuters*.

Fichten, C., Jorgensen, M., Havel, A., Vo, C., & Libman, E. (2022). Ai-based and mobile apps: Eight studies based on post-secondary students' experiences. *The Journal on Technology and Persons with Disabilities*.

Flanagin, A. J. & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journal. Mass Commun. Q.*, 77(3), 515–540.

Fleck, R. & Fitzpatrick, G. (2010a). Reflecting on reflection: Framing a design landscape. *ACM International Conference Proceeding Series*, (pp. 216–223).

Fleck, R. & Fitzpatrick, G. (2010b). Reflecting on reflection: Framing a design landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, OZCHI '10 (pp. 216–223). New York, NY, USA: Association for Computing Machinery.

Flew, T., Dulleck, U., Fisher, C., Park, S., & Isler, O. (2020). Trust and mistrust in australian news media.

Florence Davey-Attlee, I. S. (2021). The fake news machine. https://money.cnn.com/interactive/media/the-macedonia-story/. Accessed: June 4, 2021.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford.

Floridi, L. (2021). The european legislation on ai: a brief analysis of its philosophical approach. *Philosophy & Technology*, (pp. 1–8).

Foltz, P., Laham, D., & Landauer, T. (1999). Automated essay scoring: Applications to educational technology. *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 1.

Fortuin, E. (2022). "ukraine commits genocide on russians": the term "genocide" in russian propaganda. *Russ Linguist*, (pp. 313–347).

Frenda, S., Ghanem, B., Montes, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36, 4743–4752.

Frenkel, S. & Alba, D. (2021). In india, facebook grapples with an amplified version of its problems. *NY Times*.

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. *Comput. Human Behav.*, 75, 461–468.

Gadd, D. (2009). Aggravating racism and elusive motivation. *British Journal of Criminology*, 49.

Gaderer, R. (2018). Shitstorm. das eigentliche übel der vernetzten gesellschaft. *ZMK Zeitschrift für Medien- und Kulturforschung. Alternative Fakten*, 9(2), 27–42.

Gallistel, C. R. & King, A. P. (2010). *Memory and the computational brain: Why cognitive science will transform neuroscience*. John Wiley & Sons.

Gao, T., Yen, H.-C., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. *ArXiv*, abs/2305.14627.

Gaufman, E. (2016). *Security Threats and Public Perception: Digital Russia and the Ukraine Crisis*. New Security Challenges.

Gaziano, C. (1988). How credible is the credibility crisis? *Journalism Quarterly*, 65(2), 267–278.

Geden, M., Emerson, A., Carpenter, D., Rowe, J. P., Azevedo, R., & Lester, J. C. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *Int. J. Artif. Intell. Educ.*, 31, 1–23.

Geissler, D., Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). Russian propaganda on social media during the 2022 invasion of ukraine.

Gessen, M. (2022). The war that russians do not see. www.newyorker.com/news/dispatch/03/14/the-war-that-russians-do-not-see. The Guardian. Accessed: March 12, 2022.

Gibbs, G. & Unit, G. B. F. E. (1988). *Learning by Doing: A Guide to Teaching and Learning Methods. FEU.* Oxford Brookes University, Oxford.

Gibbs, G. R. (1988). Learning by doing: A guide to teaching and learning methods. *Further Education Unit*.

Gillespie, T. (2020). Content moderation, ai, and the question of scale. *Big Data & Society*, 7, 205395172094323.

Glauner, P. (2021). An assessment of the ai regulation proposed by the european commission. *arXiv preprint arXiv:2105.15133*.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7, 205395171989794.

Goujard, C. (2021). Hate speech & hate crime – inclusion on list of eu crimes. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12872-Hate-speech-hate-crime-inclusion-on-list-of-EU-crimes_en. Accessed: May 29, 2021.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *ArXiv*, abs/2008.05756.

Gröndahl, T., Pajola, L., & Juuti, M. (2018). All you need is 'love': Evading hate speech detection. In *AISec '18: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (pp. 2–12). New York: Association for Computing Machinery.

Grootendorst, M. R. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*.

Grover, K., Angara, S. M. P., Akhtar, M. S., & Chakraborty, T. (2022). Public wisdom matters! discourse-aware hyperbolic fourier co-attention for social-text classification. *ArXiv*, abs/2209.13017.

Grynszpan, E. (2022). Russian missiles target ukraine civilians and infrastructure. *Le Monde*.

Guhr, O., Schumann, A.-K., Bahrmann, F., & Böhme, H. J. (2020). Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1627–1632). Marseille, France: European Language Resources Association.

Guo, M., Chen, X., Li, J., Zhao, D., & Yan, R. (2021). How does truth evolve into fake news? an empirical study of fake news evolution. *arXiv*.

Haarkötter, H. (2016). Empörungskaskaden und rhetorische strategien in shitstorms. In *Shitstorms und andere Nettigkeiten* (pp. 17–50). Nomos Verlagsgesellschaft mbH & Co. KG.

Han, S.-H. & Brazeal, L. M. (2015). Playing nice: Modeling civility in online political discussions. *Commun. Res. Rep.*, 32(1), 20–28.

Hänsel, D. (1996). *Lehrerbildung neu denken und gestalten*. Beltz.

Harding, L. (2022). How ukrainian defiance has derailed putin's plans. www.theguardian.com/world/2022/feb/26/how-ukrainian-defiance-has-derailed-putins-\plans. The Guardian. Accessed: February 26, 2022.

Hatamian, M. (2020). Engineering privacy in smartphone apps: A technical guideline catalog for app developers. *IEEE Access*, 8, 35429–35445.

He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., & Tian, T. (2022). Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: Single-blind, three-arm randomized controlled trial. *J. Med. Internet Res.*, 24(11), e40719.

Heritage, T. (2014). Russia launches new media to lead "propaganda war" with west. www.reuters.com/article/russia-media-idUSL6N0T04MB20141110. The Guardian. Accessed: November 10, 2022.

Heylighen, F. & Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7, 293–340.

Hill, S. (2019). Empire and the megamachine: comparing two controversies over social media content. *Internet Pol. Rev.*, 8(1).

Himmelreich, S. & Einwiller, S. (2015). *Wenn der „Shitstorm" überschwappt – Eine Analyse digitaler Spillover in der deutschen Print- und Onlineberichterstattung*, (pp. 183–205). Springer Fachmedien Wiesbaden: Wiesbaden.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303.

Horne, B. D. (2023). Is automated content moderation going to solve our misinformation problems? *SSRN Electron. J.*

Horne, B. D., Nørregaard, J., & Adali, S. (2020). Robust fake news detection over time and attack. *ACM Trans. Intell. Syst. Technol.*, 11(1), 1–23.

Horton, J., Sardarizadeh, S., Schraer, R., Robinson, O., Coleman, A., Palumbo, D., & Cheetham, J. (2022). Bucha killings: Satellite image of bodies site contradicts russian claims. www.bbc.com/news/60981238. Reality Check and BBC Monitoring, BBC News. Accessed: April 5, 2022.

Hostiadi, D. P. & Ahmad, T. (2022). Hybrid model for bot group activity detection using similarity and correlation approaches based on network traffic flows analysis. *Journal of King Saud University - Computer and Information Sciences*, 34(7), 4219–4232.

Hostiadi, D. P., Ahmad, T., & Wibisono, W. (2020). A new approach of botnet activity detection model based on time periodic analysis. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)* (pp. 315–320).

Hu, M. (2020). Cambridge analytica's black box. *Big Data Soc.*, 7(2), 205395172093809.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Iris Berent, Vered Vaknin, G. F. M. (2007). Roots, stems, and the universality of lexical representations: Evidence from hebrew. *Cognition*, 104(2), 254–286.

Jang, S. M., Geng, T., Queenie Li, J.-Y., Xia, R., Huang, C.-T., Kim, H., & Tang, J. (2018). A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior*, 84, 103–113.

Janizek, J. D., Sturmfels, P., & Lee, S.-I. (2021). Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104), 1–54.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651–3657). Florence, Italy: Association for Computational Linguistics.

Jena, R. K. (2019). Sentiment mining in a collaborative learning environment: capitalising on big data. *Behaviour & Information Technology*, 38(9), 986–1001.

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.

Jung, Y. & Wise, A. F. (2020). How and how well do students reflect?: multi-dimensional automated reflection assessment in health professions education. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*.

Kabiljo, M., Vidas-Bubanja, M., Matic, R., & Zivkovic, M. (2020). Education system in the republic of serbia under COVID-19 conditions: Chatbot-acadimic digital assistant of the belgrade business and arts academy of applied studies. *Knowledge-International Journal*, 43(1), 25–30.

Kalogeropoulos, A., Suiter, J., Udris, L., & Eisenegger, M. (2019). News media trust and news consumption: factors related to trust in news in 35 countries. *International Journal of Communication*, 13.

Kapoor, P., Agrawal, P., & Ahmad, Z. (2021). Therapy chatbot: A relief from mental stress and problems. *International Journal of Scientific & Engineering Research*.

Katerynchuk, P. (2017). Russia media policy as an instrumentality of political pressure in central-eastern european countries. *Історико-політичні проблеми сучасного світу*, (pp. 283).

Kendall, B. (2014). Russian propaganda machine 'worse than soviet union. *BBC*.

Khrebtan-Hörhager, J. & Pyatovskaya, E. (2022). Putin's propaganda is rooted in russian history – and that's why it works. *The Conversation*.

Khvostunova, O. (2022). Do russians really "long for war" in ukraine? www.fpri.org/article/2022/03/do-russians-really-long-for-war-in-ukraine/. Foreign Policy Research Institute. Accessed: March 31, 2022.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

Kim, J. S., Sim, J. B., Kim, Y. J., Park, M. K., Oh, S. J., & Doo, I. C. (2023). Establishment of NLP-based greenwashing pattern detection service. In *Advances in Computer Science and Ubiquitous Computing*, Lecture notes in electrical engineering (pp. 253–259). Singapore: Springer Nature Singapore.

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *J. Commun.*, 71(6), 922–946.

Klamm, C., Rehbein, I., & Ponzetto, S. (2022). Frameast: A framework for second-level agenda setting in parliamentary debates through the lense of comparative agenda topics. *ParlaCLARIN III at LREC2022*.

Klemm, K. & Zorn, D. (2019). *Steigende Schülerzahlen im Primarbereich: Lehrkräftemangel deutlich stärker als von der KMK erwartet*. Bertelsmann Stiftung.

Kloo, I. & Carley, K. M. (2023). Social cybersecurity analysis of the telegram information environment during the 2022 invasion of ukraine. In R. Thomson, S. Al-khateeb, A. Burger, P. Park, & A. A. Pyke (Eds.), *Social, Cultural, and Behavioral Modeling* (pp. 23–32). Cham: Springer Nature Switzerland.

Knight, S., Vijay Mogarkar, R., Liu, M., Kitto, K., Sandor, A., Lucas, C., Wight, R., Sutton, N., Ryan, P., Gibson, A., Abel, S., Shibani, A., & Buckingham Shum, S. (2020). Acawriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research*, 12(1), 141–186.

Kolisko, S. & Anderson, C. J. (2023). Exploring social biases of large language models in a college artificial intelligence course. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13), 15825–15833.

Konecki, M., Konecki, M., & Biškupić, I. (2023). Using artificial intelligence in higher education. In *Proceedings of the 15th International Conference on Computer Supported Education*.

Korenyuk, M. & Goodman, J. (2022). Ukraine war: 'my city's being shelled, but mum won't believe me'. *BBC*.

Korthagen, F. & Vasalos, A. (2005). Levels in reflection: Core reflection as a means to enhance professional growth. *Teachers and teaching*, 11(1), 47–71.

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-reflections through learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, LAK '18 (pp. 389–398). New York, NY, USA: Association for Computing Machinery.

Kraemer, C. (2022). Russian bombings of civilian infrastructure raise cost of ukraine's recovery: Imf. *Reuters*.

Kruikemeier, S., Sezgin, M., & Boerman, S. (2016). Political microtargeting: Relationship between personalized advertising on facebook and voters' responses. *Cyberpsychology, behavior and social networking*, 19, 367–372.

Kruikemeier, S., Vermeer, S., Metoui, N., Dobber, T., & Zarouali, B. (2022). (tar)getting you: The use of online political targeted messages on facebook. *Big Data & Society*, 9(2), 20539517221089626.

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Educ. Inf. Technol.*, 28(1), 973–1018.

Kuhlhüser, S. (2016). Shitstorm gleich shitstorm? - eine empirische untersuchung des netzphänomens exemplarisch dargestellt am Amazon-Shitstorm 2013. In *Shitstorms und andere Nettigkeiten* (pp. 51–84). Nomos Verlagsgesellschaft mbH & Co. KG.

Kumar, S., Spezzano, F., & Subrahmanian, V. S. (2014). Accurately detecting trolls in slashdot zoo via decluttering. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*: IEEE.

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *Int. J. Educ. Technol. High. Educ.*, 20(1).

Lamberty, P. & Frühwirth, L. (2023). Pro-russian disinformation and propaganda in germany: Russia's full-scale invasion of ukraine. *Cemas*.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 1, 159–74.

Lange-Ionatamišvili, E. (2015). Analysis of russia's information campaign against ukraine: Examining non-military aspects of the crisis in ukraine from a strategic communications perspectives. *NATO Strategic Communications Centre of Excellence*.

Le, T., Wang, S., & Lee, D. (2020). Malcom: Generating malicious comments to attack neural fake news detection models. *arXiv preprint arXiv:2009.01048*.

Lee, S. (2020). Proctorio CEO releases student's chat logs, sparking renewed privacy concerns. https://www.ubyssey.ca/news/proctorio-chat-logs/. Accessed: June 11, 2023.

Li, J., Cheng, X., Zhao, W. X., Nie, J., & rong Wen, J. (2023). Halueval: A large-scale hallucination evaluation benchmark for large language models. *ArXiv*, abs/2305.11747.

Li, X., Xia, Y., Long, X., Li, Z., & Li, S. (2021). Exploring text-transformers in aaai 2021 shared task: Covid-19 fake news detection in english. In T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, & M. S. Akhtar (Eds.), *Combating Online Hostile Posts in Regional Languages during Emergency Situation* (pp. 106–115). Cham: Springer International Publishing.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Dublin, Ireland: Association for Computational Linguistics.

Link, D., Hellingrath, B., & Ling, J. (2016). A human-is-the-loop approach for semi-automated content moderation. In *International Conference on Information Systems for Crisis Response and Management*.

Liu, M., Buckingham Shum, S., Mantzourani, E., & Lucas, C. (2019a). *Evaluating Machine Learning Approaches to Classify Pharmacy Students' Reflective Statements*, (pp. 220–230). Springer.

Liu, M., Kitto, K., & Buckingham Shum, S. (2021). Combining factor analysis with writing analytics for the formative assessment of written reflection. *Computers in Human Behavior*, 120, 106733.

Liu, M., Shum, S. B., Mantzourani, E., & Lucas, C. (2019b). Evaluating machine learning approaches to classify pharmacy students' reflective statements. In S. Isotani, E. Millán, A. Ogan, P. M. Hastings, B. M. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education - 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I*, volume 11625 of *Lecture Notes in Computer Science* (pp. 220–230).: Springer.

Liu, M., Shum, S. B., Mantzourani, E., & Lucas, C. (2019c). Evaluating machine learning approaches to classify pharmacy students' reflective statements. In *AIED*.

Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., & Dolan, B. (2022). A token-level reference-free hallucination detection benchmark for free-form text generation. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6723–6737). Dublin, Ireland: Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019d). Roberta: A robustly optimized bert pretraining approach. *ArXiv*.

Llansó, E. (2019). Platforms want centralized censorship. that should scare you. *Wired*.

Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nat. Hum. Behav.*, 1(3), 0028.

Lui, M. & Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 25–30). Jeju Island, Korea: Association for Computational Linguistics.

Luke Harding, D. S. & Koshiw, I. (2022). Russia targets ukraine energy and water infrastructure in missile attacks. *The Guardian*.

Maarouf, A., Bär, D., Geissler, D., & Feuerriegel, S. (2023). Hqp: A human-annotated dataset for detecting online propaganda.

Mahyoob, M., Algaraady, J., & Alrahaili, M. (2020). Linguistic-based detection of fake news in social media. *International Journal of English Linguistics*, 11, 99.

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Marcus, G. (2020). The next decade in ai: Four steps towards robust artificial intelligence.

Markov, I., Kharitonova, K., & Grigorenko, E. L. (2023). Language: Its origin and ongoing evolution. *J. Intell.*, 11(4).

Marples, D. R. (2006). Stepan bandera: The resurrection of a ukrainian national hero. *Europe-Asia Studies*, 58(4), 555–566.

Marsden, C. & Meyer, T. (2019). *Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism*. European Parliament.

Marwa, T., Salima, O., & Souham, M. (2018). *Deep learning for online harassment detection in tweets*. 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS).

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). Hatexplain: A benchmark dataset for explainable hate speech detection.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Meister, S. (2016). *The roots and instruments of Russia's disinformation campaign*, chapter 3. Transatlantic Academy, 2015-16 Paper Series.

Meister, S. (2022). Germany: Target of russian disinformation. *DGAP External Publications*, WP/22/144, 211719.

Menendez-Alarcon, A. (2012). Newspapers coverage of spain and the united states: A comparative analysis. *Sociology Mind*, 2, 67–74.

Menn, J. (2022). Russians boasted that just 1 *The Washington Post*.

Michael Meyer-Resende, R. G. & Helena Schwertheim, N. H. (2020). Tackling disinformation and online hate speech: Eu and member state approaches, so far. *Democracy Reporting International*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, volume 26: Curran Associates, Inc.

Mohammad, S. M. & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436–465.

Moniz, A., Krings, B., & Frey, P. (2021). Technology as enabler of the automation of work? current societal challenges for a future perspective of work. *Revista Brasileira de Sociologia*, 9, 206–229.

Mont'Alverne, C., Arguedas, A. R., Toff, B., Fletcher, R., & Nielsen, R. K. (2022). The trust gap: how and why news on digital platforms is viewed more sceptically versus news in general. https://www.reuters.com/business/media-telecom/more-people-are-avoiding-news-trusting-it-less-report-says-2022-06-14/. The Reuters Institute for the Study of Journalism. Accessed: August 18, 2023.

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Morris, L. & Oremus, W. (2022). Russian disinformation is demonizing ukrainian refugees. https://www.washingtonpost.com/technology/2022/12/08/russian-disinfo-ukrainian-refugees-germany/. The Washington Post. Accessed: August 12, 2022.

Murgia, M. (2023). Openai's red team: the experts hired to 'break' chatgpt. https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8. Accessed: April 14, 2023.

Murphy, B. (2015). A corpus-based investigation of critical reflective practice and context in early career teacher settings. *Classroom Discourse*, 6, 107 – 123.

Murphy, R. F. (2019). *Artificial Intelligence Applications to Support K&ndash;12 Teachers and Teaching: A Review of Promising Applications, Challenges, and Risks*. Santa Monica, CA: RAND Corporation.

n. OED Online. Oxford University Press (2022). propaganda. March 2022. Web. 11 May 2022.

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., & Da San Martino, G. (2021). Automated fact-checking for assisting human fact-checkers. In Z.-H. Zhou (Ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (pp. 4551–4558).: International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Napanoy, J., Gayagay, G., & Tuazon, J. (2021). Difficulties encountered by pre-service teachers: Basis of a pre-service training program. *Universal Journal of Educational Research*, 9, 342–349.

Narayanan, V. P. (2017). Russian involvement and junk news during brexit. In *Comprop Data Memo 2017.10*.

NATO Strategic Communications Center of Excellence (2016). *The manipulative techniques of Russian's Information Campaign, Euro-Atlantic values and Russia's Strategic Communication in Euro-Atlantic Space*. Technical report, NATO, Latvia.

Neji, W., Boughattas, N., & Ziadi, F. (2023). Exploring new AI-based technologies to enhance students' motivation. *Issues Informing Sci. Inf. Technol.*, 20, 095–110.

Nesi, H. & Gardner, S. (2012). *Genres Across the Disciplines: Student Writing in Higher Education*. Cambridge University Press.

Nesi, H. & Gardner, S. (2013). A classification of genre families in university student writing. *Applied Linguistics*, 34, 25–52.

Nicholas, G. & Bhatia, A. (2023). Toward better automated content moderation in Low-Resource languages. *Journal of Online Trust and Safety*, 2(1).

O'Brien, P. (2022). How telegram became the digital battlefield in the russia-ukraine war. www.france24.com/en/tv-shows/tech-24/20220318-russian-invasion-of-ukraine-\telegram-finds-itself-on-frontline-\of-information-war. France 24. Accessed: March 18, 2022.

Okonkwo, C. W. & Ade-Ibijola, A. (2021a). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2(100033), 100033.

Okonkwo, C. W. & Ade-Ibijola, A. (2021b). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2, 100033.

Olex, A., DiazGranados, D., Mcinnes, B., & Goldberg, S. (2020). Local topic mining for reflective medical writing. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2020, 459–468.

Oliinyk, V.-A., Vysotska, V., Burov, Y., Mykich, K., & Fernandes, V. B. (2020). Propaganda detection in text data based on nlp and machine learning. In *MoMLeT+DS*.

on Fake News, H. L. E. G. & Disinformation, O. (2018). *Report to the European Commission on A Multi-Dimensional Approach to Disinformation*. Technical report, European Commission.

OpenAI (2023). *GPT-4 System Card*. Technical report, OpenAI.

OpenAI (2023). Gpt-4 technical report.

OpenAI, Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W., & Zhang, L. (2019). Solving rubik's cube with a robot hand. *arXiv*.

Opora, C. N. (2022). War speeches. 190 days of propaganda, or "evolution" of statements by russian politicians. *Ukrainska Pravda*.

Ostheimer, J., Chowdhury, S., & Iqbal, S. (2021). An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles. *Technology in Society*, 66, 101647.

Parlapiano, A. & Lee, J. C. (2018). The propaganda tools used by russians to influence the 2016 election. *The New York Times*.

Parliament, T. E. (2021). Consolidated version of the treaty on the functioning of the european union,part three: Union policies and internal actions, title v: Area of freedom, security and justice, chapter 4: Judicial cooperation in criminal matters, article 83. http://data.europa.eu/eli/treaty/tfeu_2008/art_83/oj. Accessed: June 4, 2021.

Patterson, L., Allan, A., & Cross, D. (2017). Adolescent bystander behavior in the school and online environments and the implications for interventions targeting cyberbullying. *Journal of School Violence*, 16(4), 361–375.

Paul, K. (2022). Flood of russian misinformation puts tech companies in the hot seat. www.theguardian.com/media/2022/feb/28/facebook-twitter-ukraine-russia-misinformation. The Guardian. Accessed: March 25, 2022.

Pavlick, E. & Tetreault, J. (2016). An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4, 61–74.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., & Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Pennebaker, J., Francis, M., & Booth, R. (1999). *Linguistic inquiry and word count (LIWC)*. Technical report, Erlbaum Publishers, Mahwah, NJ.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Comput. Human Behav.*, 140(107600), 107600.

Pérez-Mayos, L., Carlini, R., Ballesteros, M., & Wanner, L. (2021). On the evolution of syntactic information encoded by BERT's contextualized representations. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2243–2258). Online: Association for Computational Linguistics.

Plutchik, R. (1982). A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5), 529–553.

Popa-Wyatt, M. (2022a). Online hate: Is hate an infectious disease? is social media a promoter? *Journal of Applied Philosophy*, n/a(n/a).

Popa-Wyatt, M. (2022b). Social media: a viral promoter of social ills? https://blogs.cardiff.ac.uk/openfordebate/social-media-a-viral-promoter-of-social-ills/. Accessed: September, 25, 2022.

Popa-Wyatt, M. (2023). *Norm-Shifting through Oppressive Acts*, (pp. 1–15). Oxford University Press: United Kingdom.

Powers, D. (2008). Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *Mach. Learn. Technol.*, 2.

Rahat Ibn Rafiq, H., Hosseinmardi, R., Han, Q., Lv, S., & Mishra, S. A. (2015). Careful what you share in six seconds: Detect- ing cyberbullying instances in vine. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE.

Raj, M., Singh, S., Solanki, K., & Selvanambi, R. (2022). An application to detect cyberbullying using machine learning and deep learning techniques. *SN Comput. Sci.*, 3(5), 401.

Ramesh, D. & Sanampudi, S. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55, 2495–2527.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2931–2937). Copenhagen, Denmark: Association for Computational Linguistics.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 2 (pp. 241–244).

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, Jr, W. (2018). Characterizing and detecting hateful users on twitter. *arXiv*.

Ribeiro, M. H., Cheng, J., & West, R. (2023). Automated content moderation increases adherence to community guidelines.

Rillig, M., Ågerstrand, M., Bi, M., Gould, K., & Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental science & technology*, 57.

Roberts, S. T. (2017). Content moderation. In *Encyclopedia of Big Data* (pp. 1–4). Cham: Springer International Publishing.

Romanenko, V. (2022). Russia issues new guidelines on how to support mobilisation campaign. *Ukrainska Pravda*.

Roozenbeek, J., Traberg, C. S., & van der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, 9(5), 211719.

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Comput. Human Behav.*, 58, 461–470.

Rosulek, P. (2019). The post-truth age, the fake news industry, the russian federation and the central european area. *Trendy v podnikání*, 9, 46–53.

Roth, A. (2022). Kremlin reverts to type in denial of alleged war crimes in ukraine's buch. www.theguardian.com/world/2022/apr/04/. The Guardian. Accessed: April 4, 2022.

Ruthotto, I., Kreth, Q., Stevens, J., Trively, C., & Melkers, J. (2020). Lurking and participation in the virtual classroom: The effects of gender, race, and age among graduate students in computer science. *Comput. Educ.*, 151(103854), 103854.

Römer, U. & Swales, J. (2010). The michigan corpus of upper-level student papers (micusp). *Journal of English for Academic Purposes*, 9, 249–249.

Sadigov, R. (2022). Rapid growth of the world population and its socioeconomic results. *ScientificWorldJournal*, 2022, 8110229.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Sarker, M. K., Xie, N., Doran, D., Raymer, M., & Hitzler, P. (2017). Explaining trained neural networks with semantic web technologies: First steps. *arXiv*.

Scheffler, T., Solopova, V., & Popa-Wyatt, M. (2021). The telegram chronicles of online harm. *Journal of Open Humanities Data*.

Scheffler, T., Solopova, V., & Popa-Wyatt, M. (2022). Verbreitungsmechanismen schädigender sprache im netz: Anatomie zweier shitstorms. *Hassrede, Shitstorm und Darstellungspolitiken virtueller Affekt Workshop*.

Schmid, H. & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08* Morristown, NJ, USA: Association for Computational Linguistics.

Schmidt, A. & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *SocialNLP@EACL*.

Schmitt, V., Solopova, V., Woloszyn, V., & de Jesus de Pinho Pinhal, J. (2021). Implications of the New Regulation Proposed by the European Commission on Automatic Content Moderation. In *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication* (pp. 47–51).

Seering, J., Kraut, R., & Dabbish, L. (2017). Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 (pp. 111–125). New York, NY, USA: Association for Computing Machinery.

Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner–instructor interaction in online learning. *International Journal of Educational Technology in Higher Education*, 18.

Shafiei, H. & Dadlani, A. (2022). Detection of fickle trolls in large-scale online social networks. *J. Big Data*, 9(1), 22.

Shandomo, H. M. (2010). The role of critical reflection in teacher education. In *ERIC*, volume 4 (pp. 101–113).: School-University Partnerships.

Shashkov, A., Gold, R., Hemberg, E., Kong, B., Bell, A., & O'Reilly, U.-M. (2021). Analyzing student reflection sentiments and problem-solving procedures in moocs. In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, L@S '21 (pp. 247–250). New York, NY, USA: Association for Computing Machinery.

Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 312–320).

Shum, S. B., Sándor, Á., Goldsmith, R., Bass, R., & McWilliams, M. (2017). Towards reflective writing analytics: Rationale, methodology and preliminary results. *Journal of learning Analytics*, 4, 58–84.

Siatitsa, I. (2020). Freedom of assembly under attack: General and indiscriminate surveillance and interference with internet communications. *International Review of the Red Cross*, 102(913), 181–198.

Sidarenka, U. (2016). PotTS: The Potsdam Twitter sentiment corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1133–1141). Portorož, Slovenia: European Language Resources Association (ELRA).

Sinnreich, A. (2018). Four crises in algorithmic governance. *Annual Review of Law and Ethics*, 26, 181–190.

Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., Choshen, L., Cohen-Karlik, E., Dankin, L., Edelstein, L., Ein-Dor, L., Friedman-Melamed, R., Gavron, A., Gera, A., Gleize, M., Gretz, S., Gutfreund, D., Halfon, A., Hershcovich, D., Hoory, R., Hou, Y., Hummel, S., Jacovi, M., Jochim, C., Kantor, Y., Katz, Y., Konopnicki, D., Kons, Z., Kotlerman, L., Krieger, D., Lahav, D., Lavee, T., Levy, R., Liberman, N., Mass, Y., Menczel, A., Mirkin, S., Moshkowich, G., Ofek-Koifman, S., Orbach, M., Rabinovich, E., Rinott, R., Shechtman, S., Sheinwald, D., Shnarch, E., Shnayderman, I., Soffer, A., Spector, A., Sznajder, B., Toledo, A., Toledo-Ronen, O., Venezian, E., & Aharonov, R. (2021). An autonomous debating system. *Nature*, 591(7850), 379–384.

Sly, J. (2017). Timothy snyder. on tyranny: Twenty lessons from the twentieth century. new york: Tim duggan books, 2017. 126p. paper, (isbn 978-0-8041-9011-4). *College and Research Libraries*, 78(6), 868.

Smart, C. (2022). How the russian media spread false claims about ukrainian nazis. https://www.nytimes.com/interactive/2022/07/02/world/europe/ukraine-nazis-russia-media.html. The New York Times. Accessed: August 12, 2022.

Smith, B. L. (2022). propaganda. *Encyclopedia Britannica*, 24.

Snyder, T. (2018). *The road to unfreedom : Russia, Europe, America*. Tim Duggan Books New York, first edition. edition.

Solopova, V. (2023). Automated content moderation using transparent solutions and linguistic expertise. In E. Elkind (Ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23* (pp. 7097–7098).: International Joint Conferences on Artificial Intelligence Organization. Doctoral Consortium.

Solopova, V., Benzmüller, C., & Landgraf, T. (2023a). The evolution of pro-kremlin propaganda from a machine learning and linguistics perspective. In M. Romanyshyn (Ed.), *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)* (pp. 40–48). Dubrovnik, Croatia: Association for Computational Linguistics.

Solopova, V., Popescu, O.-I., Benzmüller, C., & Landgraf, T. (2023b). Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts. *Datenbank Spektrum*, 23(1), 5–14.

Solopova, V., Popescu, O.-I., Chikobava, M., Romeike, R., Landgraf, T., & Benzmüller, C. (2021). A German corpus of reflective sentences. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)* (pp. 593–600). National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI).

Solopova, V., Rostom, E., Cremer, F., Gruszczynski, A., Witte, S., Zhang, C., López, F. R., Plößl, L., Hofmann, F., Romeike, R., Gläser-Zikuda, M., Benzmüller, C., & Landgraf, T. (2023c). Papagai: Automated feedback for reflective essays. In D. Seipel & A. Steen (Eds.), *KI 2023: Advances in Artificial Intelligence* (pp. 198–206). Cham: Springer Nature Switzerland.

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44, 136–146.

Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L., & Tan, C. (2019). Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community. *arXiv*.

Stefanowitsch, A. (2020). *Der Shitstorm im Medium Twitter: Eine Fallstudie*, (pp. 185–214). De Gruyter: Berlin, Boston.

Steiger, M., Bharucha, T., Venkatagiri, S., Riedl, M., & Lease, M. (2021). The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

Stemler, S. E. & Tsai, J. W. (2008). Best practices in interrater reliability three common approaches. In *Best Practices in Quantitative Methods*.

Stone, B. (2010). Policing the web's lurid precincts. *NY Times*.

Stoop, W. & Kunneman, F. (2019). Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics.

Sturgill, A. & Motley, P. (2014). Methods of reflection about service learning: Guided vs. free, dialogic vs. expressive, and public vs. private. *Teaching and Learning Inquiry: The ISSOTL Journal*, 2, 81–93.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 7, 321–6.

Sundararajan, M., Taly, A., & Yan, Q. (2017a). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17 (pp. 3319–3328).: JMLR.org.

Sundararajan, M., Taly, A., & Yan, Q. (2017b). Axiomatic attribution for deep networks.

supérieur de l'audiovisuel, L. C. (2020). Décision n° 2020-435 du 8 juillet 2020 relative à la composition et aux missions de l'observatoire de la haine en ligne.

Sweney, M. (2022). Telegram: the app at the heart of ukraine's propaganda battle. www.theguardian.com/business/2022/mar/05/telegram-app-ukraine-rides-high-thirst-\trustworthy-news. The Guardian. Accessed: March 5, 2022.

TheResponsible AI Collaborative (2021). Ai incident database. https://incidentdatabase.ai/. Accessed: June 5, 2021.

Ting Lan, G. S. & Zhou, J. (2022). The economic impacts on germany of a potential russian gas shutoff. *International Monetary Fund*, WP/22/144, 211719.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Troianovski, A. & Safronova, V. (2022). Russia takes censorship to new extremes, stifling war coverage. *The New York Times*.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Tundis, A., Mukherjee, G., & Mühlhäuser, M. (2020). Mixed-code text analysis for the detection of online hidden propaganda. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, ARES '20 New York, NY, USA: Association for Computing Machinery.

Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. *Big Data & Society*, 10(1), 20539517231172424.

UkraineWorld.org (2022). «Страна»: популярне і проросійське медіа в Україні («strana: a popular and pro-russian media outlet in ukraine»). www.radiosvoboda.org/a/prorosiyske-media-v-ukrayini/31280240.html. Radio Svoboda. Accessed: May 30, 2022.

Ullmann, T. (2015). Keywords of written reflection - a comparison between reflective and descriptive datasets. In *ARTEL@EC-TEL*.

Ullmann, T. (2017). Reflective writing analytics: empirically determined keywords of written reflection. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*.

Ullmann, T. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, 29.

U.S. Department of State (2020). *Pillars of Russia's Disinformation and Propaganda Ecosystem*. Technical report, U.S. Department of State.

Valderrama, R., Cruz, A., Menchaca, A., & Peredo, I. (2011). Intelligent web-based education system for adaptive learning. *Expert Syst. Appl.*, 38, 14690–14702.

Valeriya Safronova, Neil MacFarquhar, A. S. (2022). Where russians turn for uncensored news on ukraine. www.france24.com/en/tv-shows/tech-24/20220318-russian-invasion-of-ukraine-\telegram-finds-itself-on-frontline-\of-information-war. The New Your Times. Accessed: April 16.

van Bekkum, M., de Boer, M., van Harmelen, F., Meyer-Vitali, A., & ten Teije, A. (2021). Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases.

Vanetik, N., Litvak, M., Reviakin, E., & Tiamanova, M. (2023). Propaganda detection in Russian telegram posts in the scope of the Russian invasion of Ukraine. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing* (pp. 1162–1170). Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.

Vorakitphan, V., Cabrio, E., & Villata, S. (2022). PROTECT – a pipeline for propaganda detection and classification. In *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-it 2021* (pp. 352–358). Accademia University Press.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.

VoxCheck (2020). Vox check propaganda diary. data retrieved from Propaganda Diary, https://rusdisinfo.voxukraine.org.

Wambsganss, T., Kueng, T., Soellner, M., & Leimeister, J. M. (2021). ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* New York, NY, USA: ACM.

Wang, H., Huang, Z., Dou, Y., & Hong, Y. (2020a). Argumentation mining on essays at multi scales. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5480–5493). Barcelona, Spain (Online): International Committee on Computational Linguistics.

Wang, Q., Jing, S., Camacho, I., Joyner, D., & Goel, A. (2020b). Jill watson SA: Design and evaluation of a virtual agent to build communities among online learners. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* New York, NY, USA: ACM.

Wardle, C. & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, 27, 1–107.

Weber, T. & Van Mol, C. (2023). The student migration transition: an empirical investigation into the nexus between development and international student migration. *Comp. Migr. Stud.*, 11(1), 5.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9, 36 – 45.

Weld, D. S. & Bansal, G. (2018). *Intelligible artificial intelligence.* arXiv,CoRR.

West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20, 146144481877305.

Wilbur, D. (2021). Propaganda or not: Examining the claims of extensive russian information operations within the united states. *Journal of Information Warfare*, 20, 146–156.

Willems, A. S., Dreiling, K., & Eckert, M. (2020). *Skalendokumentation des Projekts FeeHe: Feedback im Kontext von Heterogenität*. Universitätsverlag Göttingen.

Winkler, R. & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. *Academy of Management Proceedings*, 2018, 15903.

Wogu, I. A. P., Misra, S., Olu-Owolabi, E. F., Assibong, P. A., Udoh, O. D., Ogiri, S. O., & Damasevicius, R. (2018). Artificial intelligence, artificial teachers and the fate of learners in the 21st century education sector: Implications for theory and practice. *International Journal of Pure and Applied Mathematics*, 119(16), 2245–2259.

Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., & Biemann, C. (2017). GermEval 2017: Shared task on aspect-based sentiment in social media customer feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback* (pp. 1–12). Berlin, Germany.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, 4.

Wulff, P., Buschhüter, D., Westphal, A., Nowak, A. I., Becker, L., Robalino, H., Stede, M., & Borowski, A. (2020). Computer-based classification of preservice physics teachers' written reflections. *Journal of Science Education and Technology*, 30, 1–15.

Xu, C., Guo, D., Duan, N., & McAuley, J. (2023). Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Yang, K.-C. & Menczer, F. (2023). Anatomy of an ai-powered malicious social botnet.

Ye, M., Sikka, K., Atwell, K., Hassan, S., Divakaran, A., & Alikhani, M. (2023). Multilingual content moderation: A case study on Reddit. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 3828–3844). Dubrovnik, Croatia: Association for Computational Linguistics.

Yu, S., Martino, G. D. S., Mohtarami, M., Glass, J. R., & Nakov, P. (2021). Interpretable propaganda detection in news articles. *CoRR*, abs/2108.12802.

Yuskiv, K. (2023). Social media content moderation: The case of russia's war against ukraine. *Visnyk of the Lviv University*, 46, 373–379.

Yuzefyk, K. (2022). Потенційно неприйнятний: як онлайн платформи обмежують контент українців. https://ukrainer.net/pravo-na-kontent/. Accessed: September 11, 2023.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., & Farra, N. (2019a). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *SemEval-2019*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). Predicting the type and target of offensive posts in social media.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019a). Defending against neural fake news.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019b). Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023a). Sentiment analysis in the era of large language models: A reality check.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023b). Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

Zhou, X. & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5), 1–40.

Zhou, Z., Guan, H., Bhat, M. M., & Hsu, J. (2019). Fake news detection via nlp is vulnerable to adversarial attacks. *arXiv preprint arXiv:1901.09657*.

Zimmerman, B. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41, 64–70.

Áine MacDermott, Motylinski, M., Iqbal, F., Stamp, K., Hussain, M., & Marrington, A. (2022). Using deep learning to detect social media 'trolls'. *Forensic Science International: Digital Investigation*, 43, 301446.

Фіялка, С. В. (2022). Проблематика дотримання стандартів спільноти в соцмережі «Фейсбук» в умовах збройної агресії російської федерації («problems of compliance with community standards on facebook in the context of the armed aggression of the russian federation»). *Обрії друкарства*, 2(12), 5–17.