Datenbank Spektrum (2023) 23:5–14
https://doi.org/10.1007/s13222-023-00437-2

SCHWERPUNKTBEITRAG

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

Veronika Solopova[1] · Oana-Iuliana Popescu[2] · Christoph Benzmüller[1] · Tim Landgraf[1]

Received: 3 February 2023 / Accepted: 8 February 2023 / Published online: 3 March 2023

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, French and English. We propose and compare two methods of multilingual automated pro-Kremlin propaganda identification, 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

1 Introduction

Propaganda influences an audience to support a political agenda. [34, 43]. Propaganda has been shown to play a vital role in the Russian invasion of Ukraine, shaping the war approval rate [24] by e.g. fabricating explanations for war crimes [40]. As a result, fake news also spreads through Ukrainian, Central European and Western media [21], seeding mistrust and confusion [36].

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 fully placing the responsibility on human moderators [45]. The objective of this study is 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 [32], it also has a number of uniform strategies and approaches which are recurrently mentioned in research throughout the whole of Russian-Ukrainian war by international bodies and independent researchers [8, 16, 28, 49]. 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 intransparent neural networks. Here, we propose a linguistics-based approach for detecting war propaganda in the context of the ongo-

✉ Veronika Solopova
veronika.solopova@fu-berlin.de

Oana-Iuliana Popescu
oana.iuliana.popescu@gmail.com

Christoph Benzmüller
c.benzmueller@fu-berlin.de

Tim Landgraf
tim.landgraf@fu-berlin.de

[1] Dahlem Center for Machine Learning and Robotics, Freie Universität Berlin, Berlin, Germany

[2] German Aerospace Center, Jena, Germany

ing 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 from 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 Sect. 2 we present previous work related to our research. In Sect. 3 we introduce our training setup for each experiment, describing the data and model configurations. In Sect. 4 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 Sect. 5 we present the results for each setting, while in Sect. 6 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 Sect. 7 we consider the main findings and limitations of our work. We lay the ground for future work opportunities and delve into ethical dangers in terms of its potential usage for automated content moderation. In Sect. 8 we summarize the main contributions of our study.

2 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 [30].

Our work is motivated by the fact that despite the assumed involvement of Russia in Brexit [31] and the 2016 US presidential elections [11], there is still only a small number of peer-reviewed research publications investigating Russian state-sponsored fake news [5, 15]. Furthermore, existing publications are not always neutral, with some using accusations and emotional lexicon [39], while others accuse Western media of propaganda used to discredit Russia [9].

Wilbur [50] 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 [42, 44] 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 and O'Brien [1] 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 and Carley [5] 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 [31] 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 and Taboada [3] show-cased that significant performance can be achieved even with n-grams; Antoun et al. [2], Mahyoob et al. [26], Rashkin et al. [38] implemented different linguistic feature-based solutions, while Li et al. [25] 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 [51]. Research on propaganda detection is less frequent. Although existing works proposed using both feature-based and contextual embedding approaches [12, 35, 47, 51], 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?

3 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 [14] 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 Fig. 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.

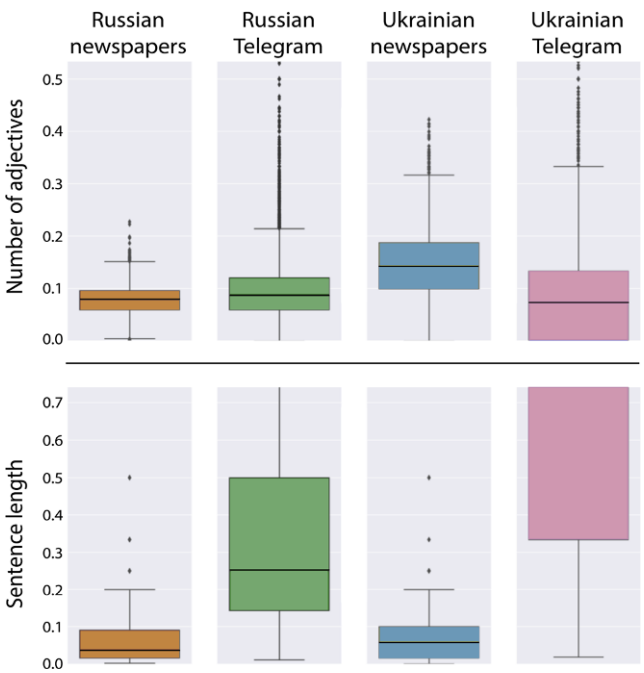

Fig. 1 Examples of genre-related differences between Newspapers and Telegram subsets. The boxplot represents 25% around the median, the whiskers show the first and last quartiles. Ukrainian news has more adjectives than Telegram posts, while this is vice-versa for Russian news. Sentences are longer in Telegram for both languages

4 Data

4.1 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 that support pro-Kremlin narratives in Ukraine that have been confirmed by journalistic investigations to directly copy pieces of news from Russian news outlets [48]. 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 [39], 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 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 1.

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

4.2 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 [4, 33, 41]. However, it is a source without moderation and fact-checking, hence fake news and emotional lexicon, including profanity and propaganda, are not unusual [46]. Therefore, we included both Russian and Ukrainian Telegram news in our data collection.

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

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

Table 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'	4049
English	'The Guardian', 'BBC', 'The New York Times', 'Reuters')	1060
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 'Dc-news'.	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

4.3 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 [26] 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 [3], 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 [29]. 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. [38], 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 synonyms search and can be found in 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.

4.4 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 and 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 lexi-

[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.

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

con being avoided by Russian officials (e.g. 'war', 'victims', 'children', 'casualties') [18].

5 Results

We evaluate the performance of our models using Cohen's κ [10] and F-measure [37]. While 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 [19]. 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. 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, added our extracted keywords to the dataset, but found no improvement.

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.

Table 2 Comparative results achieved on best folds

Algorithm	Cohen's κ	F1	False positives	False negatives
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

6 Additional Analysis

6.1 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 [7]. The analysis is illustrated in Fig. 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.

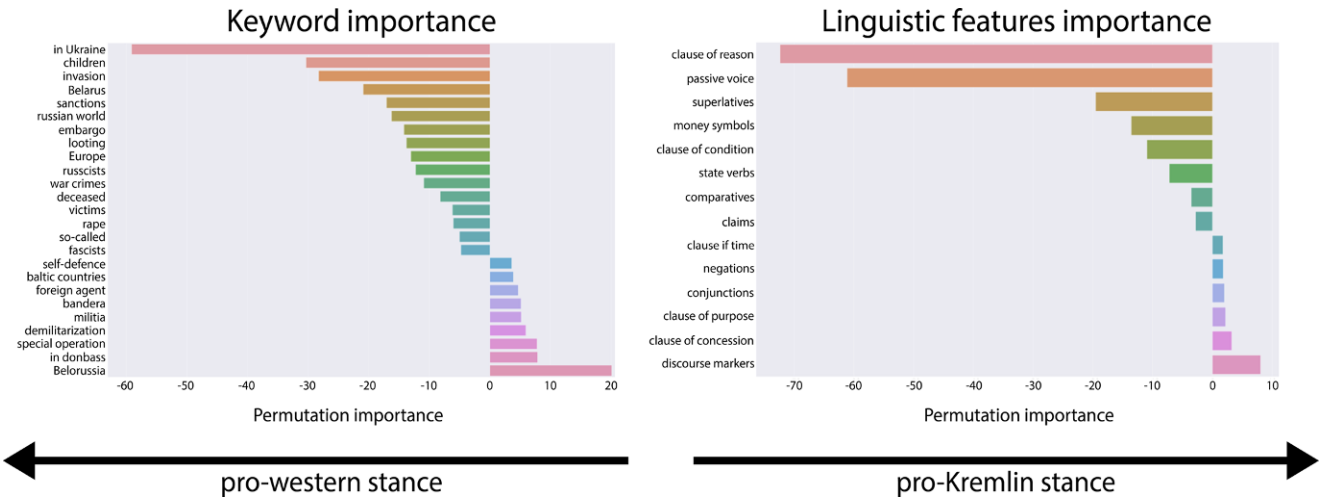

Fig. 2 Permutation importance (drop of F1 score in %) for an SVM with linear kernel. Keyword importance is on the left side and the importance of linguistic features is on the right. Negative bars indicate features that are important for classifying a data point as pro-Western, while positive bars represent features indicative of pro-Kremlin propaganda

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 will 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 [17].

6.2 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 to explain frequent model errors and observe 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 [6]: 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.

[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 [27].

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 [20].

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 [22]. Pro-Western Romanian data also did not change considerably, with an 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' [44] 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.

7 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 [13] of such models, we leave this for future work. As BERT models can capture syntax [23], 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

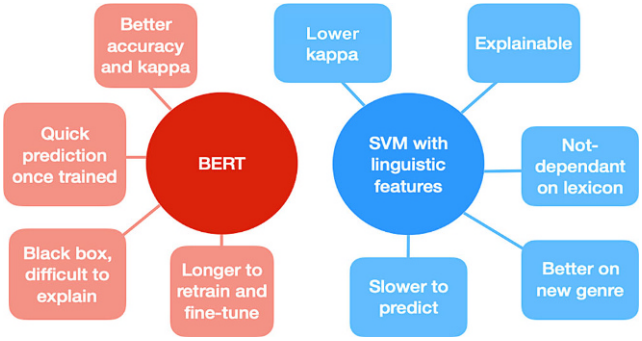

Fig. 3 Advantages and disadvantages of the presented methods

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 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 Fig. 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.

8 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.

9 Appendix

9.1 Hyper-parameters

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.

For the neural network used with linguistic features, our best setup was achieved with 2 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, 4 epochs and batch size 16.

Acknowledgements The research was completed while the second author was still affiliated with Freie Universität Berlin.

Funding Not applicable

Competing interests The corresponding author VS is of Ukrainian nationality but is not affiliated with any of the Ukrainian governmental or private institutions. Other authors have no competing interests to declare.

Ethics approval Not applicable

Consent to participate Not applicable

Consent for publication Not applicable

Availability of data and materials All data is available in a GitHub depository.

Code availability All code is available in the same GitHub repository.

Authors' contributions Conceptualization: [Veronika Solopova]; Methodology: [Veronika Solopova, Oana-Iuliana Popescu]; Formal analysis and investigation: [Veronika Solopova, Oana-Iuliana Popescu]; Writing – original draft preparation: [Veronika Solopova]; Writing – review and editing: [Veronika Solopova, Tim Landgraf, Oana-Iuliana Popescu]; Supervision: [Tim Landgraf, Christoph Benzmüller].

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Alsmadi I, O'Brien M (2020) How many bots in russian troll tweets? Inf Process Manag 57:102,303. https://doi.org/10.1016/j.ipm.2020.102303
2. 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 https://doi.org/10.1109/ICIoT48696.2020.9089487
3. Asr FT, Taboada M (2019) Big data and quality data for fake news and misinformation detection. Big Data Soc 6(1): 2053951719843310. https://doi.org/10.1177/2053951719843310
4. Bergengruen V (2022) How telegram became the digital battlefield in the russia-ukraine war. The Time March 21. https://time.com/6158437/telegram-russia-ukraine-information-war/. Accessed 11.05.2022
5. Beskow D, Carley K (2020) Characterization and comparison of Russian and Chinese disinformation campaigns. In: Disinformation, misinformation, and fake news in social media, pp 63–81 https://doi.org/10.1007/978-3-030-42699-6_4
6. Boffey DDS, Borger J (2022) Mariupol theatre bombing killed 300, ukrainian officials say. The Guardian March 25. https://www.theguardian.com/world/2022/mar/25/mariupol-theatre-bombing-killed-300-ukrainian-officials-say. Accessed 11.05.2022
7. Breiman L (2004) Random forests. Mach Learn 45:5–32
8. Carroll J (2017) Image and imitation the visual rhetoric of pro-russian propaganda. Ideol Polit J 8:36–79
9. Chudinov A, Koshkarova N, Ruzhentseva N (2019) Linguistic interpretation of russian political agenda through fake, deepfake, post-truth. J Siberian Fed Univ Humanit Soc Sci. https://doi.org/10.17516/1997-1370-0492
10. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104
11. Cosentino G (2020a) Polarize and conquer: Russian influence operations in the united states. In: Social media and the post-truth world order. Springer, Cham, pp 33–57
12. Dadu T, Pant K, Mamidi R (2020) Towards detection of subjective bias using contextualized word embeddings. CoRR, vol abs/2002.06644
13. De Cao N, Schlichtkrull M, Aziz W, Titov I (2020) How do decisions emerge across layers in neural models? interpretation with differentiable masking https://doi.org/10.18653/v1/2020.emnlp-main.262
14. Devlin J, Chang MW, 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). Association for Computational Linguistics, Minneapolis, pp 4171–4186 https://doi.org/10.18653/v1/N19-1423
15. Elswah M, Howard PN (2020) "Anything that causes chaos": the organizational behavior of Russia Today (RT). J Commun 70(5):623–645. https://doi.org/10.1093/joc/jqaa027
16. Fortuin E (2022) "Ukraine commits genocide on russians": the term "genocide" in russian propaganda. Russ Linguist 46:313–347. https://doi.org/10.1007/s11185-022-09258-5
17. Gaufman E (2016) Security threats and public perception: digital russia and the Ukraine crisis. New security challenges
18. Gessen M (2022) The war that russians do not see. The New Yorker March 12. https://www.newyorker.com/news/dispatch/03/14/the-war-that-russians-do-not-see. Accessed 11.05.2022
19. Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview (ArXiv abs/2008.05756)
20. Harding L (2022) How ukrainian defiance has derailed putin's plans. The Guardian February 26. https://www.theguardian.com/world/2022/feb/26/how-ukrainian-defiance-has-derailed-putins-plans. Accessed 11.05.2022
21. Heritage T (2014) Russia launches new media to lead "propaganda war" with west. Reuters November 10. https://www.reuters.com/article/russia-media-idUSL6N0T04MB20141110. Accessed 11.05.2022
22. 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. Reality Check and BBC Monitoring, BBC News April 5. https://www.bbc.com/news/60981238. Accessed 11.05.2022
23. 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, Association for Computational Linguistics, Florence, Italy, pp 3651–3657 https://doi.org/10.18653/v1/P19-1356
24. Khvostunova O (2022) Do russians really "long for war" in ukraine? Foreign Policy Research Institute March 31. https://www.fpri.org/article/2022/03/do-russians-really-long-for-war-in-ukraine/. Accessed 10.05.2022
25. 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. Springer, Cham https://doi.org/10.48550/ARXIV.2101.02359
26. Mahyoob M, Algaraady J, Alrahaili M (2020) Linguistic-based detection of fake news in social media. IJEL 11:99
27. Marples DR (2006) Stepan bandera: The resurrection of a ukrainian national hero. Eur Asia Stud 58(4):555–566
28. Meister S (2016) The roots and instruments of Russia's disinformation campaign (Transatlantic Academy, 2015-16 paper series, chap 3)
29. Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436–465
30. 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: Zhou ZH (ed) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, pp 4551–4558 https://doi.org/10.24963/ijcai.2021/619
31. Narayanan VP (2017) Russian involvement and junk news during brexit comprop data memo 2017.10
32. 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 (Tech. rep.)
33. O'Brien P (2022) How telegram became the digital battlefield in the russia-ukraine war. France 24 March 18. https://www.france24.com/en/tv-shows/tech-24/20220318-russian-invasion-of-ukraine-

telegram-finds-itself-on-frontline-of-information-war. Accessed 11.05.2022
34. OED Online (2022) Propaganda vol 11. Oxford University Press, Oxford
35. Oliinyk VA, Vysotska V, Burov Y, Mykich K, Fernandes VB (2020) Propaganda detection in text data based on nlp and machine learning. In: MoMLeT+DS
36. Paul K (2022) Flood of russian misinformation puts tech companies in the hot seat. The Guardian March 1. www.theguardian.com/media/2022/feb/28/facebook-twitter-ukraine-russia-misinformation. Accessed 10.05.2022
37. Powers D (2008) Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. Mach Learn Technol 2. https://doi.org/10.48550/arXiv.2010.16061
38. Rashkin H, Choi E, Jang JY, 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, Association for Computational Linguistics, Copenhagen, Denmark, pp 2931–2937 https://doi.org/10.18653/v1/D17-1317
39. Rosulek P (2019) The post-truth age, the fake news industry, the russian federation and the central european area. Trendy V Podnikání9:46–53. https://doi.org/10.24132/jtb.2019.9.3.46_53
40. Roth A (2022) Kremlin reverts to type in denial of alleged war crimes in ukraine's bucha. The Guardian April 4. https://www.theguardian.com/world/2022/apr/04/. Accessed 10.05.2022
41. Safronova V, MacFarquhar ASN (2022) Where russians turn for uncensored news on ukraine. The New Your Times April 16. https://www.nytimes.com/2022/04/16/world/europe/russian-propaganda-telegram-ukraine.html. Accessed 11.05.2022
42. 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). Coll Res Libr 78(6):868–16744. https://doi.org/10.5860/crl.78.6.868
43. Smith BL (2022) Propaganda. Encyclopedia Britannica, p 24
44. Snyder T (2018) The road to unfreedom : Russia, Europe, America, 1st edn. Tim Duggan, New York
45. Steiger M, Bharucha TJ, Venkatagiri S, Riedl MJ, Lease M (2021) The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. Association for Computing Machinery, New York https://doi.org/10.1145/3411764.3445092
46. Sweney M (2022) Telegram: the app at the heart of ukraine's propaganda battle. The Guardian March 5. www.theguardian.com/business/2022/mar/05/telegram-app-ukraine-rides-high-thirst-trustworthy-news. Accessed 11.05.2022
47. 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, Association for Computing Machinery, New York, NY, USA, ARES '20 https://doi.org/10.1145/3407023.3409211
48. UkraineWorldorg (2022) «Страна»: популярне і проросійське медіа в Україні Radio Svoboda May 30. https://www.radiosvoboda.org/a/prorosiyske-media-v-ukrayini/31280240.html. Accessed 11.05.2022
49. US Department of State (2020) Pillars of russia's disinformation and propaganda ecosystem (Tech. rep.)
50. Wilbur D (2021) Propaganda or not: Examining the claims of extensive russian information operations within the united states. J Inf Warf 20:146–156
51. Yu S, Martino GDS, Mohtarami M, Glass JR, Nakov P (2021) Interpretable propaganda detection in news articles (CoRR abs/2108.12802)