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

Veronika Solopova
Freie Universität Berlin, Germany
veronika.solopova@fu-berlin.de

Christoph Benzmüller
Universität Bamberg, Germany
Freie Universität Berlin, Germany
christoph.benzmueller@uni-bamberg.de

Tim Landgraf
Freie Universität Berlin, Germany
tim.landgraf@fu-berlin.de

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.

1 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., 2019). 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.

Inn 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 and 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 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 and 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 biolabs in Ukraine", "8 years of genocide in Donbas" and the claim that the Ukrainian government is responsible for shelling its own cities (Korenyuk and 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 and 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., 2023), we proposed two multilingual automated pro-Kremlin

Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 40–48

propaganda identification methods, based on the multilingual BERT model (Devlin et al., 2018) 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.

In Section 2, describe the experimental setup and the new data set. We present our results in comparison to those from 2022 in Section 3. In Section 4 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 5, 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 6.

2 Methods

2.1 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., 2018) 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 and false negative rate.

2.2 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 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].

3 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 makes 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 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

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

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

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

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 negatives, 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.

4 Error analysis

SVM. Regarding the SVM model, some patterns can be observed by looking into the distributions between the true positives, true negatives, false positives, and false negatives. 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 positives. In the case of relative clauses, clauses of condition and time, there is a correlation between both true positives-false positives 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 positives, 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.1

BERT. We were inspired by the attribution method (Sundararajan et al., 2017). 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 words indicating propaganda for Romanian were '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 how 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’.

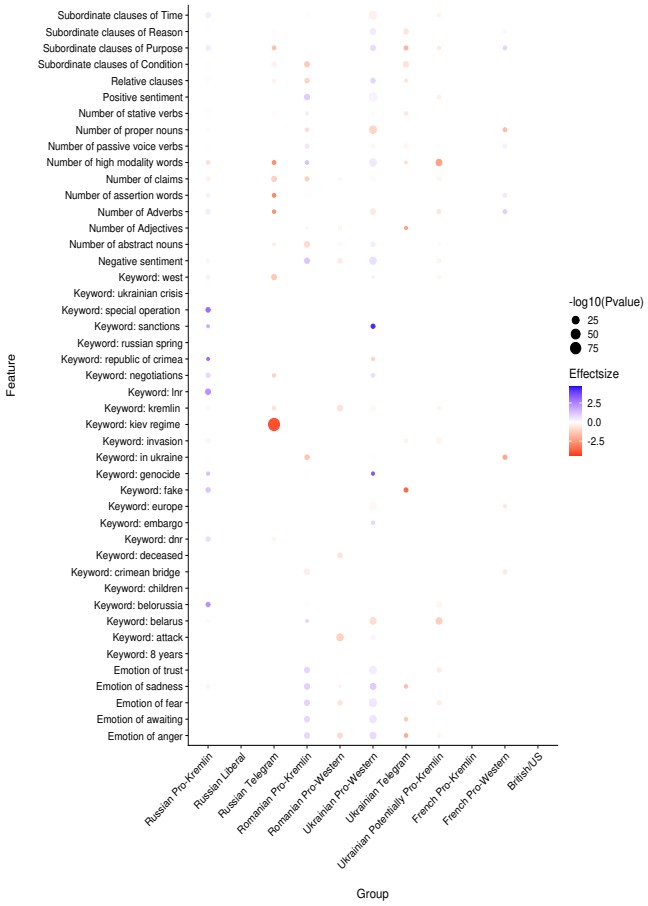

Figure 1: The dot plot shows the comparison between 2022 and 2023 subsets according to linguistic features. The dot size shows P-values while the colour shows the effect size. It represents the difference between the 2023 and 2022 averages, with red indicating growth in usage and blue meaning the drop.

5 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 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 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, 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.1.

6 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 and 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.

Acknowledgements

The author VS would like to express gratitude to fellow researcher Lev Petrov, who actively helped us with the visual component of this paper.

References

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

Foo Yun Chee. 2017. Nato says it sees sharp rise in russian disinformation since crimea seizure. *Reuters*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, pages 37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

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

Guy Faulconbridge. 2022b. Russia fights back in information war with jail warning. *Reuters*.

Emmanuel Grynszpan. 2022. Russian missiles target ukraine civilians and infrastructure. *Le Monde*.

Mingfei Guo, Xiuying Chen, Juntao Li, Dongyan Zhao, and Rui Yan. 2021. How does truth evolve into fake news? an empirical study of fake news evolution. *arXiv*.

S. Mo Jang, Tieming Geng, Jo-Yun Queenie Li, Ruofan Xia, Chin-Tser Huang, Hwalbin Kim, and Jijun Tang. 2018. A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior*, 84:103–113.

Bridget Kendall. 2014. Russian propaganda machine 'worse than soviet union. *BBC*.

Julia Khrebtan-Hörhager and Evgeniya Pyatovskaya. 2022. Putin's propaganda is rooted in russian history – and that's why it works. *The Conversation*.

Maria Korenyuk and Jack Goodman. 2022. Ukraine war: 'my city's being shelled, but mum won't believe me'. *BBC*.

Christian Kraemer. 2022. Russian bombings of civilian infrastructure raise cost of ukraine's recovery: Imf. *Reuters*.

Elina Lange-Ionatamišvili. 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*.

Dan Sabbagh Luke Harding and Isobel Koshiw. 2022. Russia targets ukraine energy and water infrastructure in missile attacks. *The Guardian*.

Civil Network Opora. 2022. War speeches. 190 days of propaganda, or "evolution" of statements by russian politicians. *Ukrainska Pravda*.

Alicia Parlapiano and Jasmine C. Lee. 2018. The propaganda tools used by russians to influence the 2016 election. *The New York Times*.

David Powers. 2008. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *Mach. Learn. Technol.*, 2.

Valentyna Romanenko. 2022. Russia issues new guidelines on how to support mobilisation campaign. *Ukrainska Pravda*.

Veronika Solopova, Oana-Iuliana Popescu, Christoph Benzmüller, and Tim Landgraf. 2023. Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts. *Datenbank Spektrum*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Anton Troianovski and Valeriya Safronova. 2022. Russia takes censorship to new extremes, stifling war coverage. *The New York Times*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news.

A Appendix

A.1 Appendix

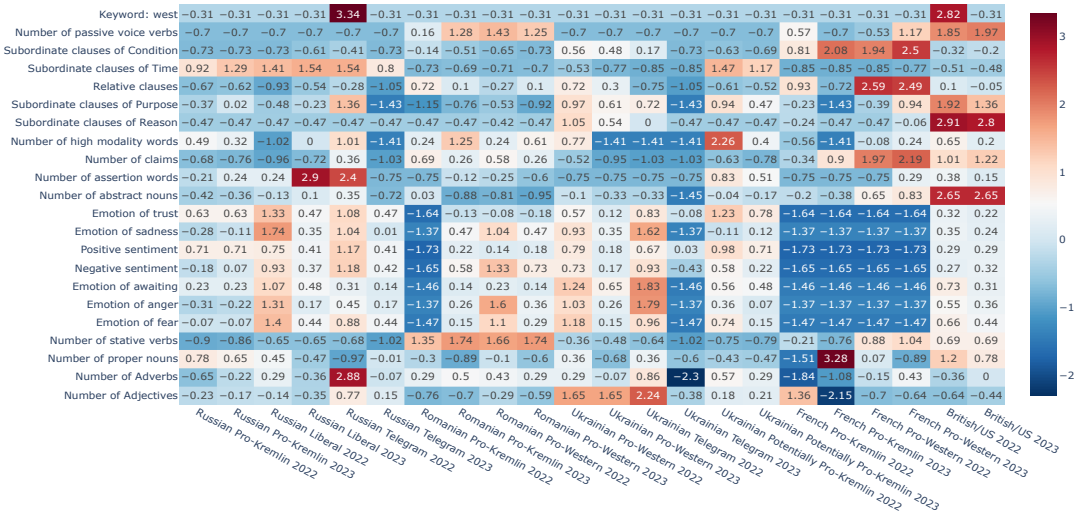

Figure 2: Normalized averages from the Comparative analysis. Linguistic features.

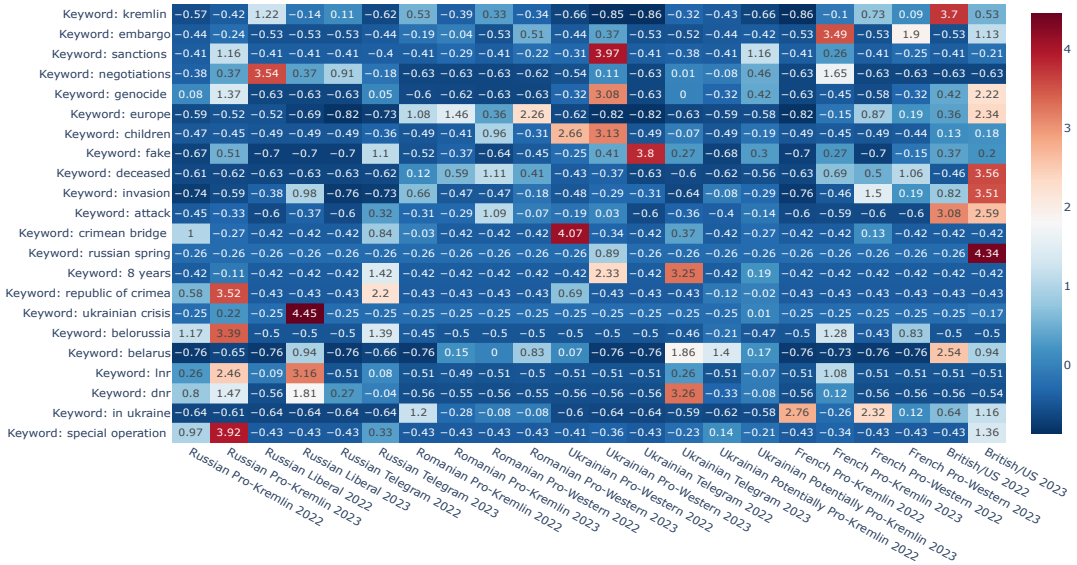

Figure 3: Normalized averages from the Comparative analysis. Keywords.

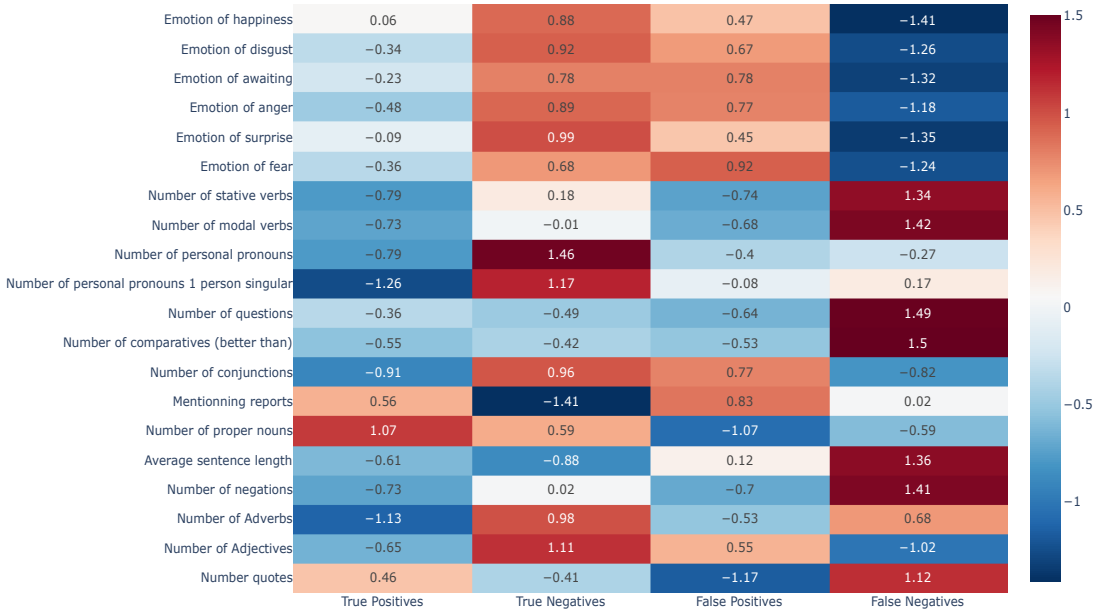

Figure 4: Error analysis. Normalized averages of linguistic features for the groups of errors.

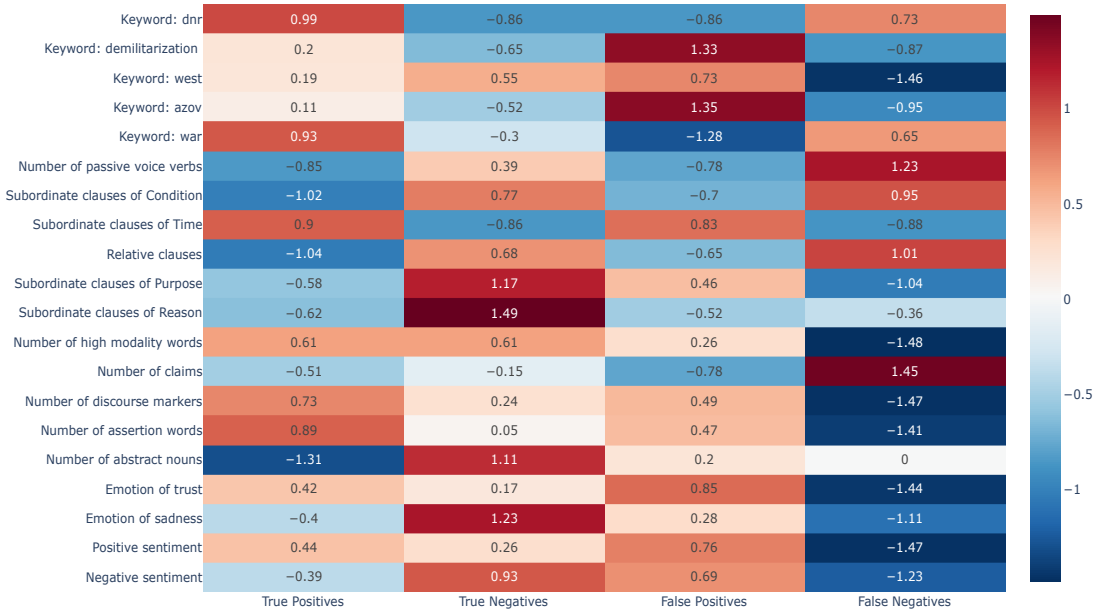

Figure 5: Error analysis. Normalized averages of keyword occurrences for the groups of errors.