Trustworthy AI and Multimodal Modelling

PROPer-BG: A New Publicly Accessible Package of Resources for Propaganda Detection in Bulgarian

Irina Temnikova[1,♡], **Devora Kotseva**[2,†], **Ruslana Margova**[1,‡], **Todor Kiriakov**[2,†], **Yolina Petrova**[2,3,*], **Nevena Grigorova**[1,‡], **Alexander Komarov**[1,◇], **Bogomil Katanov**[2,†], **Boryana Kostadinova**[2,†], **Stefan Minkov**[1,#]

[1]Big Data for Smart Society Institute (GATE),
[2]Identrics, [3]New Bulgarian University, Sofia, Bulgaria

[♡]irina.temnikova@gmail.com,[*]ypetrova@nbu.bg,
[‡]{ruslana.margova,nevena.grigorova}@gate-ai.eu,
[†]{devora.kotseva,todor.kiriakov,bogomil.katanov,boryana.kostadinova}
@identrics.ai,
[◇]av.kom0411@gmail.com, [#]stefanminkov2003@gmail.com

Abstract

Detecting texts containing propaganda has emerged as an important Natural Language Processing (NLP) task. While there are many models and resources for other languages, Bulgarian has scarce resources for this task. This article presents PROPer-BG – the first publicly accessible package of resources for detecting and studying propaganda in Bulgarian language. PROPer-BG is an umbrella package, which contains two sub-parts – the WASPer dataset and taxonomy (created by a team at Identrics) and the PropaGATE dataset and lexical resources (created by a team at GATE Institute). The two datasets contain texts in a different selection of topics, collected from different sources, and annotated at different levels of granularity. PropaGATE contains 1939 Large Language Models (LLM)-generated short texts, manually annotated as containing propaganda or not. Its originality consists of including propaganda examples on the same topics but with contrasting viewpoints (e.g. pro- and anti-capitalism), to prevent models from confusing topics with propaganda. The WASPer dataset was used to train the previously released propaganda detection model WASPer. It contains 1696 human-written and LLM-generated short texts on several topics, annotated for the presence of propaganda and propaganda techniques, using the WASPer taxonomy. Additionally, two lists of superlatives and intensifiers were created, and their presence in both datasets was analysed. The paper presents all the resources and the results from the analysis, as well as the link at which the package can be accessed.

Keywords: propaganda; datasets; Bulgarian; linguistic analysis

1 Introduction

Propaganda originates from Latin "to propagate" and was first used by the Vatican in 1622 for propagating the faith of the Roman Catholic Church (Jowett and O'Donnell, 2018). Later, propaganda acquired pejorative meaning. (Lee and Lee, 1939) defines propaganda as "the expression of opinions or actions carried out deliberately by individuals or groups with a view to influencing the opinions or actions of other individuals or groups for predetermined ends through psychological manipulations".

There is a lot of research on the automatic detection of propaganda in texts, including annual organised Shared Tasks, such as those at Slavic NLP workshop, Sem-Eval, and CLEF-CheckThat! (Da San Martino et al., 2020a; Nakov et al., 2020; Piskorski et al., 2025). Because of this, datasets in multiple languages also exist (Da San Martino et al., 2019; Nakov et al., 2020; Baisa et al., 2019; Abdygalym et al., 2025). The datasets contain a mixture of text types, including parliamentary speeches, news articles, and social media posts (Piskorski et al., 2025). The shortcomings of some of these datasets are that when the texts are long, they may contain a mixture of sentences on different topics and with and without propaganda, while when containing social media texts, the proper texts could not be shared in most cases due to the social media platforms' limitations. As a solution to this, as a new trend, datasets with synthetically generated texts also started to appear (Liu et al., 2025; Pakina et al., 2025; Shah et al., 2024). However, such datasets for the moment exist only for some languages (English, Russian, Arabic, Chinese, Spanish, French).

An additional issue is that often the texts are selected on the basis of the main view, separating the official beliefs from alternative opinions. Such official beliefs are sometimes universal and valid for many countries (like the statements that Covid-vaccines are beneficial). But some other beliefs are culturally and country-specific, for example, the Turkish president Erdogan is regarded as a hero in Arab countries and a villain in European countries. However, annotating only the alternative views as propaganda risks making Machine Learning (ML) models confuse topics and views with propaganda.

Stance-annotated datasets represent a kind of solution to this issue, by presenting an author's attitude toward a topic, typically modelled as pro, contra or neutral. However, rarely do datasets contain both propaganda and stance annotations. Specifically, there are no publicly accessible and synthetically generated text datasets, especially those presenting stance and multiple points of view for Bulgarian language.

As solutions to the above issues, we present a package of publicly accessible Bulgarian language resources for propaganda detection[1]. The resources include two text datasets with different text types and annotation granularity. One - partially annotated for stance (called "views"), and a lexical resource, containing superlatives, which can be used for additional analysis.

Contributions:

- We introduce **the first publicly accessible Bulgarian text datasets with propaganda annotations**. The texts partially originate from news articles' comment sections, while the rest are LLM-generated.
  - The PropaGATE dataset is the first Bulgarian texts dataset to contain propaganda presence, partial propaganda techniques, and stance annotations.
  - The WASPer dataset contains fine-grained annotations for the texts origin, presence of propaganda, propaganda categories, and propaganda techniques. Its efficiency has been confirmed as it was previously used to train the propaganda detection model WASPer (Petrova et al., 2025), achieving an F1-Score of 0.853.
- We **release a curated lexical resource with lists of superlatives and intensifiers, comprising 81 positive and 102 negative expressions**, supporting interpretable, lexicon-based analysis of framing signals in Bulgarian.
- We **present the results of a semi-automatic linguistic analysis of the presence of intensifiers and superlatives in the two datasets**.

PROPer-BG is the overall publicly accessible resource package introduced in this paper. It integrates two complementary components: (i) the WASPer component, which comprises the previously developed propaganda taxonomy, detection models (introduced in a previous article), and the WASPer dataset (released publicly for the first time in this paper); (ii) PropaGATE, a newly created synthetic dataset with propaganda, stance, and partial technique annotations, and two lexicons with superlative and intensifier expressions.

2 Related Work

Propaganda and Persuasion Detection: Propaganda and persuasion detection have been studied through the identification of rhetorical and framing devices in political and social discourse, often relying on manually annotated corpora. Early work on propaganda detection has focused on identifying rhetorical techniques such as loaded language, appeal to authority, and emotional manipulation. A widely adopted taxonomy of propaganda techniques was introduced by (Da San Martino et al., 2019), which has been used in subsequent shared tasks such as (Da San Martino et al., 2020a). These efforts distinguish between sentence-level classification and fine-grained span detection, highlighting the complexity of modelling propagandistic language.

More recent approaches rely on transformer-based models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and domain-adapted architectures, achieving strong performance on propaganda detection benchmarks. However, these models often rely heavily on dataset-specific cues and may fail to generalise across topics and domains.

Stance Detection in Political and Social Text: Stance detection aims to identify an author's attitude toward a target, typically modeled as pro, contra, or neutral. Stance detection has been widely studied in the context of political debates and misinformation analysis, often framed as a target-dependent

[1]The PROPer-BG package, containing a version of WASPer and the up-to-date versions of PropaGAT can be viewed at https://osf.io/2cr7w/overview?view_only=aeeba9a6376e4cdf8246726559dac368

classification task (Mohammad et al., 2016). Recent work explores its interaction with related tasks such as sentiment analysis and argument mining, as well as its role in detecting biased or misleading content. It is closely related to the analysis of political and social debates, including cross-topic and target-dependent settings.

Synthetic Data Generation for NLP: The use of synthetic data generated by large language models has recently gained attention as a strategy to address data scarcity in NLP. Studies such as (Shah et al., 2024) and (Liu et al., 2025) demonstrate that controlled generation can produce diverse and balanced datasets for tasks including misinformation and propaganda detection. However, concerns remain regarding distributional mismatch and the risk of models overfitting to synthetic patterns.

Lexicon-Based and Interpretable Approaches: Lexicon-based approaches have long been used for sentiment and framing analysis, offering interpretable features that complement data-driven models (Taboada et al., 2011). In the context of propaganda detection, lexical cues, such as intensifiers, superlatives, and emotionally charged expressions can signal exaggerated or manipulative framing, although they are not sufficient indicators on their own. Research on propaganda and persuasion detection has emphasized the identification of rhetorical and framing devices in political and social discourse, often relying on large, manually annotated corpora. Closely related work on stance detection focuses on identifying viewpoint polarity with respect to specific targets, a task that remains underexplored for low-resource languages such as Bulgarian. Recent advances in large language models have enabled the generation of synthetic datasets for NLP, providing controlled alternatives to real-world data, particularly in low-resource settings. At the same time, lexicon-based approaches continue to offer interpretable signals for propagandistic language, especially through intensifiers and superlatives.

Despite these advances, existing work remains limited in several respects: (i) low-resource languages such as Bulgarian are underrepresented; (ii) few datasets combine propaganda detection with stance annotations; and (iii) the role of synthetic data in this domain is still underexplored. Our work addresses these gaps by introducing a Bulgarian-language resource that integrates propaganda, stance, and lexical signals within both human-written and LLM-generated data.

2.1 Propaganda Datasets for Other Languages and for Bulgarian

Table 1 summarizes the main propaganda datasets, including those for English, Slavic languages, Romanian (as close to Bulgarian), and Bulgarian. The datasets with Bulgarian texts contain parliamentary speeches and Twitter/X posts. Parliamentary speeches are long texts with many sentences on diverse topics, some of which may contain propaganda, while others do not. Twitter/X posts are, instead not directly accessible as texts, according to Twitter/X's requirements to share only tweet ids.

In contrast, both our datasets contain only short texts of 1-3 sentences (which restricts the diversity of topics per text). Additionally, our datasets are the first Bulgarian-language propaganda datasets, containing synthetically generated examples. Also, we are not aware of other datasets which show multiple viewpoints on the same topic, which makes the contribution of PropaGATE unique.

3 PROPer-BG Description

PROPer-BG is an umbrella package comprising the WASPer solution, the PropaGATE dataset, and two lexical resources; its structure and release status are summarised in Table 2. WASPer comprises a dataset, a hierarchical propaganda taxonomy, and detection models. The taxonomy and some models were released previously, while the WASPer dataset is publicly released for the first time in this paper and is described in Section 3.1. The taxonomy is shown in Figure 1 and Appendix A. PropaGATE is described in Section 3.2, and the lexical resources in Section 3.3; both are used in the analysis in Section 4.

3.1 WASPer Dataset

The WASPer dataset contains texts in English and Bulgarian and is composed by training and testing datasets. **The training datasets** consist of a balanced set of examples that include both propaganda and non-propaganda content. These examples are collected from a variety of traditional media and social media sources, ensuring a diverse range of content. The **"Propaganda" examples** include an equal distribution of human-generated and Artificial Intelligence (AI)-generated propaganda examples. Human-generated examples were manually selected and annotated by domain experts based on a specially designed taxonomy of propaganda techniques (see Section 3.1.1). The AI-generated

Dataset	Year	Lang.	Domain	Source	Annot.	Granul.	Size	Type	Stance	Tech.
SlavicNLP 2025 (Piskorski et al., 2025)	2025	BG, PL, HR, SL, RU	News, parliament, social	Human	Manual	Sent., doc	~10k	Real	No	No
SemEval-2020 (Da San Martino et al., 2020b)	2020	EN	News	Human	Manual	Span	~500 art.	Real	No	Yes
Propaganda Techniques Corpus (Da San Martino et al., 2019)	2019	EN	News	Human	Manual	Span	~18k	Real	No	Yes
CLEF / CheckThat! (Nakov et al., 2020)	2019–24	EN, FR, IT, DE, RU, PL, ES, EL, KA, AR, PT, SL, BG	Political, media	Human	Mixed	Claim	>20k	Real	Partial	Partial
Czech Dataset (Baisa et al., 2019)	2023	CS	News	Human	Manual	Sentence	~4k	Real	No	No
Multi-mil Corpus (Abdygalym et al., 2025)	2023	EN, RU, UK, BG, SR, HR	Info ops	Human	Mixed	Doc.	~50k	Real	No	Partial
SeLeRoSa (Smădu et al., 2025)	2025	RO	News, satire	Human	Manual	Doc.	~5k	Real	No	No
COVID-19 in Bulgarian Social Media (Nakov et al., 2021)	2021	BG	Covid-19	Human	Manual	Doc.	2172 tweets	Real	No	No

Table 1: Unified comparison of propaganda datasets with stance and technique annotations.

Component	Role in PROPer-BG	Release status
PROPer-BG	Umbrella resource package introduced in this paper	New
WASPer solution	Integrated solution comprising a taxonomy, models, and a dataset	Partly previous, partly new
WASPer taxonomy	Hierarchical taxonomy of propaganda categories and techniques	Previously released
WASPer models	Models for propaganda detection	Some previously released
WASPer dataset	Annotated dataset underlying the WASPer solution	First public release in this paper
PropaGATE	Separate annotated dataset included in PROPer-BG	New
Lexical resources	Superlative and intensifier lexicons	New

Table 2: Structure and release status of the resources included in PROPer-BG.

examples were produced using LLMs and subsequently reviewed and annotated by domain experts to ensure consistency with the human-generated content. The specific approach for generating synthetic propaganda is further explored in below. The **"Non-Propaganda"** examples were sourced similarly from traditional and social media platforms, and some were AI-generated. To ensure variety and relevance, these examples were filtered and sampled through a topic modelling process aimed at capturing a wide range of topics and reducing bias in the dataset. Topic modelling is a technique used to uncover hidden themes within large collections of documents by automatically grouping texts based on their underlying topics.

Topic modelling was performed using BERTopic (Grootendorst, 2022), a state-of-the-art library for topic modelling, which leverages language model embeddings to identify coherent topics in text data. BERTopic integrates two key algorithms: UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). UMAP first reduces the dimensionality of the text data, transforming it into a lower-dimensional space that captures the relationships between different texts. HDBSCAN then identifies clusters within this reduced space, grouping similar texts into distinct topics based on density. To guarantee that the non-propaganda dataset includes content related to the same topics as the propaganda examples, zero-shot topic modelling was also applied. In this approach, the already described topics were used as a guide to locate non-propaganda examples discussing similar themes. This approach ensures that non-propaganda examples are present across the same range of topics as propaganda, allowing for balanced and comprehensive coverage of themes This overall process helps organise content and ensures that the dataset covers a wide range of topics without requiring extensive manual annotation. Yet, the sampled examples were verified later. Finally, each example in the dataset was labelled with three types of annotations:

- Whether the content is synthetic (AI-generated) or organic (human-written);
- Whether it contains propaganda;
- If propaganda is present, the specific category of the propaganda.

The total size of the two training sets was as follows: 734 examples in Bulgarian and 862 examples in English (for a more detailed distribution split,

Language	Label	Human	AI-generated
Bulgarian	Propaganda	206	161
Bulgarian	No propaganda	345	22
English	Propaganda	214	206
English	No propaganda	442	0

Table 3: Distribution of the WASPer training data by language, propaganda label, and text origin.

Language	Label	Human	AI-generated
Bulgarian	Propaganda	12	13
Bulgarian	No propaganda	12	13
English	Propaganda	12	13
English	No propaganda	10	13

Table 4: Distribution of the WASPer evaluation data by language, propaganda label, and text origin.

refer to Table 3 below). To supplement the organic propaganda examples, synthetic data was generated using LLMs. This approach was designed to extend the dataset, particularly to include AI-generated propaganda scenarios as well. The generation of synthetic data was driven by few-shot prompting, where the models were provided with carefully selected organic propaganda examples. Different examples were provided for each propaganda technique to ensure the generated content varied in vocabulary, structure, and style. These prompts included specific details related to each propaganda technique, guiding the model to generate content that adhered to the established taxonomy.

Taxonomy-driven generation was also applied. The same taxonomy utilised for annotating the organic examples was applied during the generation of synthetic content, ensuring consistency across the dataset. In addition to organic examples, the models were provided with definitions of each technique found within them. The synthetic examples were reviewed and annotated by domain experts to verify their quality and alignment with the organic examples. To manage the prompting process across different models and tasks, the LangChain framework was used. This framework helped organise the interactions with the language models and streamline the generation process by structuring the prompts and outputs in a consistent manner. The models used were BgGPT (Alexandrov et al., 2024) for Bulgarian, and Mistral (Jiang et al., 2023) and LLaMA (Touvron et al., 2023) for English.

In addition to the training dataset, a separate **evaluation dataset** was created to assess the overall performance of the inference pipeline. This evaluation dataset was sourced similarly to the training set. It includes both organic and AI-generated content from various traditional and social media platforms to ensure diverse and balanced coverage across all labels, including the specific propaganda categories (Table 4 for reference). The Bulgarian evaluation set consists of a total of 50 examples, 25 of which non-propaganda and 25 propaganda examples, 26 of which organic and 24 AI-generated examples. The English evaluation set also has a total of 50 examples and contains 25 non-propaganda examples and 25 propaganda examples, 26 of which organic and 22 AI-generated.

3.1.1 Taxonomy of Propaganda Techniques

The WASPer taxonomy is shown on Figure 1. It was designed to systematically classify and detect AI-generated propaganda in both traditional media comment sections and social media platforms. This taxonomy progresses through multiple stages, each aimed at narrowing down specific characteristics of the content, starting from the basic identification of text type (as human or artificially generated) and advancing to a detailed classification of the propaganda techniques present within the message. The taxonomy contains the following four levels:

- Level 1: Text Origin Identification (Human vs. AI-generated);
- Level 2: Presence of Propaganda (Propaganda vs. No Propaganda);
- Level 3: Propaganda Category;
- Level 4: Propaganda Technique.

The taxonomy consists of 22 separate propaganda techniques (on the lowest level) grouped into five parent categories (Figure 1 for reference). Appendix A provides more details about it. The taxonomy is open source and can be accessed online[2].

3.2 PropaGATE Dataset

The PropaGATE dataset is completely synthetically generated and contains 1939 short "social media-like" texts, generated with BARD, OpenAI 3.5 and 5.4 in different ways. The texts containing propaganda were generated by providing examples from Facebook, deemed to contain propaganda by one human annotator. Subsequently, the same number of factual non-propaganda texts were generated with the ChatGPT OpenAI model 5.4, corresponding to the number of the propaganda-containing texts.

[2] https://github.com/Identrics/wasper.

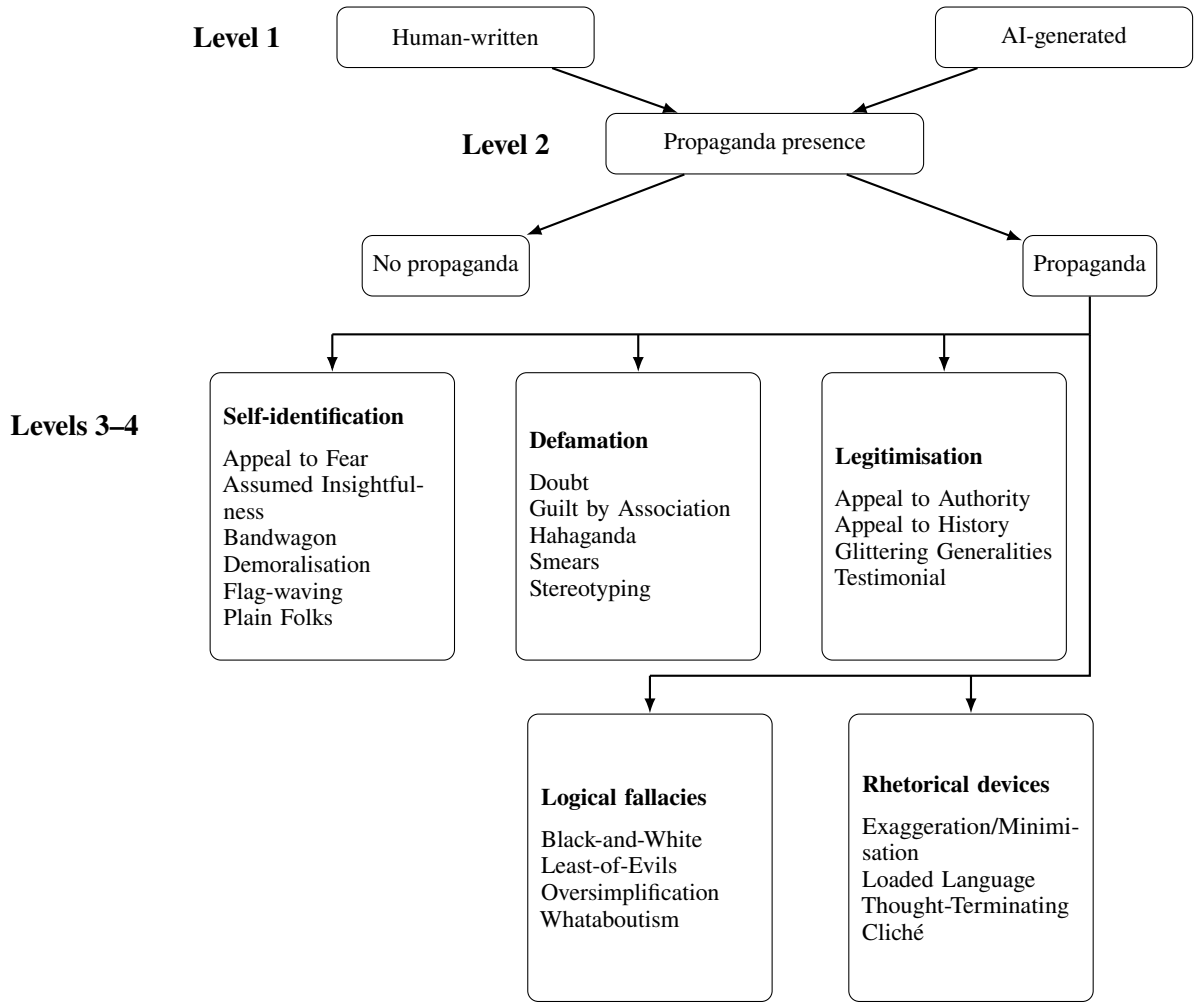

Figure 1: Overview of the four-level WASPer taxonomy. Level 1 identifies text origin, Level 2 identifies propaganda presence, and Levels 3–4 assign propaganda categories and techniques.

The recorded prompts used for generating the texts can be viewed online[3] or obtained by contacting Irina Temnikova[4].

Each text was also manually annotated by one linguist for topic and stance (called "view"). The stances partially reflected the standard pro-, contra-, and neutral ones, and on some occasions reflected the existing society's beliefs (like global warming views were: 1) pro-global-warming as a problem; 2) against global warming as a problem; 3) pro-global warming as a conspiracy; 4) neutral). Next, two human annotators (Bulgarian language linguists, specializing in disinformation and propaganda detection) reviewed each generated text for whether it was written in correct Bulgarian, and whether it contained propaganda or not. The decision of whether it contains propaganda was taken on the basis of several propaganda definitions and the description of propaganda techniques in the WASPer taxonomy and in (Da San Martino et al., 2020a). A text was considered to contain propaganda if it attempted to convince the reader that something was bad or good and if it contained at least one of the propaganda techniques described in the WASPer taxonomy. The annotators were also asked to annotate at least the first propaganda technique they noticed in each tagged as propaganda-containing text. We used two annotation tools. We first used Datasaur[5] (the annotation interface is shown in Figure 2). When the available free annotations ended, we continued the annotation in Google Spreadsheets, using fall-down menus (see Figures 3 and 4). We obtained substantial agreement with Cohen's k = 0.717, and percentage agreement of 85.35% in terms of whether the text contained propaganda or not. Due to the fact that propaganda techniques

[3] https://osf.io/2cr7w/overview?view_only=aeeba9a6376e4cdf8246726559dac368

[4] irina.temnikova@gmail.com

[5] https://datasaur.ai/data-annotation

were not comprehensively annotated, we did not calculate agreement for them.

Table 6 in Appendix B shows the number of texts, their topics, subgroups and groups in PropaGATE.

3.3 Lexical Resources

As an addition to PropaGATE, we provide two manually curated lexicons - one with 366 expressions, out of which (i) 162 (81 ADJ and 81 ADV) positive superlatives and (ii) 204 negative (102 ADJ and 102 ADV) superlatives. Additionally, we include a list of intensifiers, used in Section 4. These lexicons can be used to approximate lexical intensification as a proxy signal for exaggerated or evaluative framing. Some **Examples** are: "изключително", "ужасно", "безспорно" (English translation: "exceptionally", "terribly", "undoubtedly").

4 Linguistic Analysis for Intensifiers and Superlatives

We conducted an exploratory manual analysis of the occurrence and functions of intensifiers and superlatives from our lexical resources in both datasets. Table 5 presents representative intensifiers for several broad topic categories in PropaGATE. The analysis suggests several category-specific tendencies. For example, texts related to COVID-19 and vaccines contain lexical items emphasising authority, safety, and collective responsibility, typically supporting expert-driven and institutional framing. Climate change texts employ intensifiers expressing urgency and escalating risk, whereas ideology-related texts favour universal and exclusionary formulations. The Gender and Social Roles category relies on comparative and capability-denying expressions that construct hierarchies and discredit target groups. Finally, the Wealth and Inequality category combines blame, dehumanisation, and moral judgement, framing particular groups as responsible for systemic problems.

When analysing WASPer, we observed that intensifiers and superlative-like expressions are not distinguished by their frequency but by how they are used in context. In propaganda texts, intensifiers such as "супер" (super), "ужасно" (terribly), "страшно" (horribly), "абсолютно" (absolutely), "напълно" (completely) typically appear in structured patterns that combine emotional amplification with negative or ideologically loaded content. They are often attached to negative events or enemies, as in "супер мощна атака" (super powerful attack) or "ужасно решение" (terrible decision), or to evaluative nouns like "невероятно лицемерие" (incredible hypocrisy) and "изумително безобразие" (astonishing outrage). These constructions serve functions such as dramatization, moral judgment, and delegitimization. In addition, hyperbolic expressions like "абсолютно доказателство" (absolute proof) and "напълно ясно" (completely clear) present opinions as facts and remove doubt. Instead, the non-propaganda texts use intensifiers in a more neutral and descriptive way. Common expressions such as "много" (very), "доста" (quite), "малко" (a bit) appear in phrases like "много трудно" (very difficult), "много странно" (very strange), or "доста интересно" (quite interesting). Here, the function is descriptive rather than persuasive. An important difference is in the sentiment shift. In propaganda, intensifiers push already evaluative statements toward extremes — for example, "лошо" (bad) → "ужасно" (terrible) or "проблем" (problem) → "огромен проблем" (huge problem) — thus creating emotional polarization. In non-propaganda texts, intensifiers such as "много" (very) only slightly strengthen meaning, as in "трудно" (difficult) → "много трудно" (very difficult), without moving into extreme emotional territory. The observation suggest that the same linguistic forms — such as "супер" (super), "ужасно" (terribly), "много" (very), "невероятно" (incredibly) — can function very differently depending on context: in propaganda, they act as strategic tools for emotional manipulation and persuasion, while in non-propaganda - primarily serve descriptive and informational purposes. The comparison between the WASPer and PropaGATE datasets suggests two distinct patterns of lexical intensification in relation to propaganda. The WASPer dataset exhibits a broader and more natural use of intensifiers such as "супер" (super), "ужасно" (terribly), and "невероятно" (incredibly), which occur in both propaganda and non-propaganda texts. Their communicative function is highly context-dependent: in propaganda they tend to amplify emotional and ideological messages, whereas in non-propaganda they are used primarily for descriptive or conversational purposes. Consequently, these expressions alone constitute weak indicators of propaganda. In contrast, the PropaGATE dataset contains fewer instances of lexical intensification but relies more frequently on explicit superlative constructions such as "най-голям" (the biggest) and "най-лош" (the

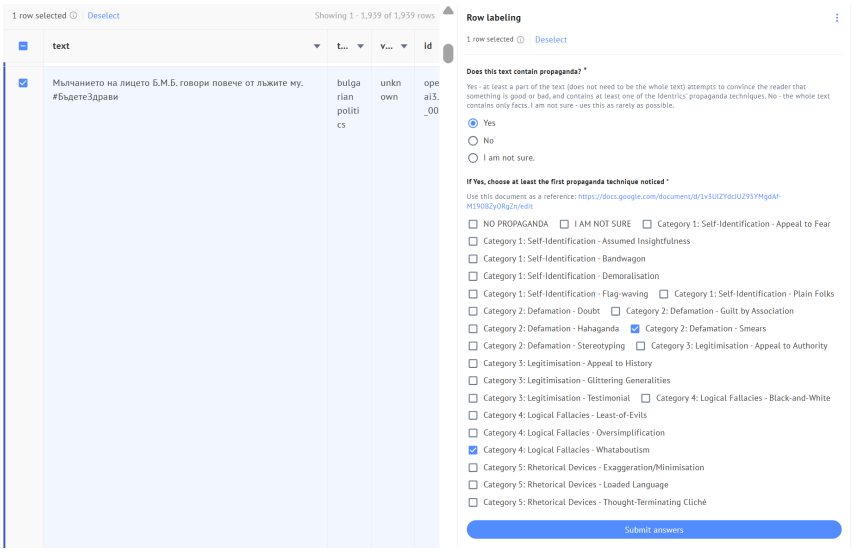

Figure 2: PropaGATE Annotation in Datasaur.

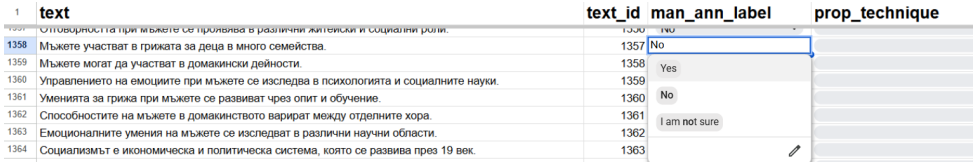

Figure 3: PropaGATE Annotation in Google Spreadsheets - fall down menu for Propaganda or Not.

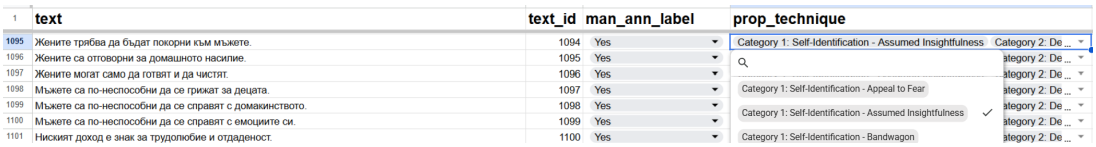

Figure 4: PropaGATE Annotation in Google Spreadsheets - fall down menu for Propaganda techniques.

worst). These forms directly express categorical and extreme evaluations, often corresponding to rhetorical strategies such as black-and-white thinking and exaggeration. A further difference concerns emotional intensity. In WASPer, intensifiers typically create a gradual shift from moderate to stronger sentiment (e.g., "лошо" (bad) → "ужасно" (terrible)), whereas in PropaGATE superlative constructions already encode maximal intensity without intermediate stages. Despite these differences, neither dataset suggests that intensifiers or superlatives alone are reliable markers of propaganda. The observations suggest that propaganda relies less on overtly extreme language such as "най-. . . " (the most. . .) and more on subtle lexical intensification combined with discourse context and other rhetorical strategies, including loaded language, fear appeals, and labelling. The identified intensifiers were compiled into a second lexical resource comprising 70 polarity-marked expressions, which is released together with the superlative lexicon as part of the PROPer-BG package.

The observations presented in this section are intended as an exploratory linguistic analysis. They are descriptive rather than statistically validated and should therefore be interpreted as qualitative evidence of different patterns of lexical intensification across the two datasets. Future work will extend this exploratory analysis with corpus-level quantitative measurements and statistical significance testing.

5 Limitations and Ethical Aspects

It should be noted that the real intent of the authors of the human posts in the WASPer dataset remain unclear. Annotating some as propaganda, thus, should not be considered as a hypothesis. In PropaGATE, because the data is fully synthetic, it is not intended to represent real actors or events and should not be used as a standalone basis for

Topic	Bulgarian Intensifiers	English Translation	Type
Vaccines / COVID-19	ключова; изключително; най-ефективните; смъртоносни; строги; безопасни; много нисък риск; социална отговорност; защита на уязвимите	key; extremely; most effective; deadly; strict; safe; very low risk; social responsibility; protection of the vulnerable	Institutional / Moral
Climate Change	най-голямата заплаха; катастрофални; тревожно бързо; спешни действия; неизбежен; опустошителни; сериозни щети; не можем да си позволим	greatest threat; catastrophic; alarmingly fast; urgent actions; inevitable; devastating; serious damage; cannot afford	Alarmist
Ideologies	винаги; единствената система; най-успешната; обречена на неуспех; няма място; трябва да бъде; в основата на всички	always; the only system; most successful; doomed to fail; no place for; must be; at the core of everything	Absolutist
Gender / Social Roles	по-малко способни; не могат; по-слаби; по-емоционални; по-малко интелигентни; не са в състояние	less capable; cannot; weaker; more emotional; less intelligent; not capable of	Discrediting
Wealth / Inequality	най-големите врагове; паразити; крадат; винаги; трябва да бъдат наказани; фалшиво щастие	greatest enemies; parasites; steal; always; must be punished; false happiness	Accusatory

Table 5: Topic-specific Bulgarian-language lexical intensifiers with English translations and functional categorisation.

real-world propaganda detection. Intensification is treated as a proxy signal rather than a definitive indicator of propaganda, as exaggerated language may also occur in persuasive but non-manipulative contexts. Finally, the lists of superlatives and intensifiers are not exhaustive.

The publicly accessible resources are supplied with legal disclaimers, listing these limitations and that these resources should not be trusted to unambiguously distinguish propaganda.

6 Conclusions

In this article, we have introduced a publicly accessible package, containing two datasets with Bulgarian texts, annotated with propaganda (and one also with stance), as well as two lexicons with superlative and intensifier expressions. We have also linguistically analysed which intensifiers and superlatives, how many, and in what function, appear in texts, annotated as containing propaganda and not in both datasets. The new package will be very useful for training propaganda detection models for Bulgarian, as well as for expanding the linguistic knowledge about propaganda in Bulgarian.

7 Acknowledgments and Authors' Contributions

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020, TEAMING Phase 2, the Programme under grant agreement no. 857155, the Programme "Research, Innovation and Digitalization for Smart Transformation" 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01 and, the project BROD (Bulgarian-Romanian Observatory of Digital Media), funded by the Digital Europe Programme of the European Union under contract number 101226153.

Authors' Contributions. From the Institute GATE's team, Irina Temnikova led the internally funded project, during which the dataset was created, wrote most parts of this article (except the WASPer- and Identrics-related paragraphs and sections), generated the non-propaganda texts and the lexical resources, annotated PropaGATE for topic and view, created the propaganda or not annotation guidelines, set the annotation tools, calculated IAA, ran the linguistic analysis, organised GATE's review response, and contributed to the final camera-ready version of the article. Alexander Komarov generated the BARD pro- and anti- texts. Stefan Minkov generated the texts with OpenAI 3.5. Nevena Grigorova checked the Bulgarian of the BARD-generated texts, and together with Ruslana Margova manually annotated all PropaGATE texts for containing propaganda or not, and with at least 1 technique. Ruslana Margova has also contributed to the camera-ready article.

From Identrics' team, Devora Kotseva and Todor Kiriakov designed the WASPer Taxonomy, annotated the human-written texts, and verified the synthetically generated texts. Bogomil Katanov developed the prompting strategy and the text gen-

eration pipeline. Boryana Kostadinova performed automated NLP-driven filtering and topic diversity selection of the datasets and contributed to the writing of the manuscript. Yolina Petrova conceived and supervised the overall research process and contributed to the writing of the manuscript. All authors reviewed and approved the final version of the manuscript.

References

Bayangali Abdygalym, Madina Sambetbayeva, Aigerim Yerimbetova, Anargul Nekessova, Nurbolat Tasbolatuly, Nurzhigit Smailov, and Aksaule Nazymkhan. 2025. NLP models for military terminology analysis and detection of information operations on social media. *Computers*, 14(11):485.

Angel Alexandrov, Veselin Raychev, Martin N. Müller, Cheng Zhang, Martin Vechev, and Kristina Toutanova. 2024. Mitigating catastrophic forgetting in language transfer via model merging. *arXiv preprint arXiv:2407.08699*.

Vít Baisa, Ondřej Heřman, and Aleš Horák. 2019. Benchmark dataset for propaganda detection in Czech newspaper texts. In *Recent Advances in Natural Language Processing*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of EMNLP*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2020b. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling. *arXiv preprint arXiv:2203.05794*.

Albert Q. Jiang et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Garth S Jowett and Victoria O'Donnell. 2018. *Propaganda & persuasion*. Sage publications.

Alfred Lee and Elizabeth Briant Lee. 1939. *The fine art of propaganda*. Harcourt, Brace.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, Yi Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2025. Propainsight: Toward deeper understanding of propaganda in terms of techniques, appeals, and intent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5607–5628.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. Association for Computational Linguistics.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. Covid-19 in bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009.

Preslav Nakov et al. 2020. Overview of the clef-2020 checkthat! lab: Automatic identification and verification of claims in social media. In *CLEF Working Notes*.

Anil Kumar Pakina, Ashwin Sharma, and Deepak Kejriwal. 2025. Ai-driven disinformation campaigns: Detecting synthetic propaganda in encrypted messaging via graph neural networks. *International Journal Science and Technology*, 4(1).

Yolina Petrova, Todor Kiryakov, Devora Kotseva, and Boryana Kostadinova. 2025. Wasper: a bulgarian-language model for detecting propaganda in social media content. In *Papers from the International Scientific Conference of the European Studies Department, Jean Monnet Centre of Excellence, Faculty of Philosophy at Sofia University "St. Kliment Ohridski"*, volume 12, pages 223–230.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michał Marcińczuk, Arkadiusz Modzelewski, Ivo Moravski, et al. 2025. Slavic-NLP 2025 shared task: Detection and classification of persuasion techniques in parliamentary debates and social media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 254–275.

Uzair Shah, Md Rafiul Biswas, Marco Agus, Mowafa Househ, and Wajdi Zaghouani. 2024. Mememind at araieval shared task: Generative augmentation and

feature fusion for multimodal propaganda detection in arabic memes through advanced language and vision models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 467–472.

Răzvan-Alexandru Smădu, Andreea Iuga, Dumitru-Clementin Cercel, and Florin Pop. 2025. Selerosa: Sentence-level romanian satire detection dataset. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6528–6533.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Hugo Touvron et al. 2023. Llama. *arXiv preprint arXiv:2302.13971*.

Appendix A First Appendix

The four-level hierarchical taxonomy provides a solid framework for the detection and classification of AI-generated propaganda. Further, each level is described in depth, detailing the specific features and techniques used to identify and categorise these components.

1. **Human vs. AI-generated**:
    - **Human/Organic examples** are comments sourced from traditional or social media that are assumed to be human-generated content, representing natural, unaltered communication by individuals or organisations.
    - **AI-generated/Synthetic examples** are AI-generated pieces of content created by large language models (LLMs). In this context, synthetic examples are used to simulate AI-driven propaganda and help the model learn to differentiate between human and AI-generated messaging.

2. **Propaganda vs. No Propaganda**: The propaganda or not level indicates whether a specific piece of content contains propaganda technique(s). If labelled as "propaganda", it means that a piece of content associated with propaganda has been detected, but this doesn't guarantee that the entire piece is propaganda — only that a manipulative method has been identified. Content labelled as "No Propaganda" lacks any propaganda content and/or identifiable techniques.

3. **Propaganda categories (Level 3)** and **Propaganda techniques (Level 4)**:
    - **Category 1: Self-Identification Techniques:**
      These techniques exploit the audience's feelings of association (or desire to be associated) with a larger group. They suggest that the audience should feel united, motivated, or threatened by the same factors that unite, motivate, or threaten that group. The propagandist may be internal or external to the group, but the audience must necessarily identify (or strive to identify) with the group.
        - Appeal to Fear
        - Assumed Insightfulness
        - Bandwagon
        - Demoralisation
        - Flag-waving
        - Plain Folks
    - **Category 2: Defamation Techniques:**
      These techniques represent direct or indirect attacks against an entity's reputation and worth.
        - Doubt
        - Guilt by Association
        - Hahaganda
        - Smears
        - Stereotyping
    - **Category 3: Legitimisation Techniques:** These techniques attempt to prove and legitimise the propagandist's statements by using arguments that cannot be falsified because they are based on moral values or personal experiences.
        - Appeal to Authority
        - Appeal to History
        - Glittering Generalities
        - Testimonial
    - **Category 4: Logical Fallacies:** These techniques appeal to the audience's reason and masquerade as objective and factual arguments, but in reality, they exploit distractions and flawed logic.
        - Black-and-White
        - Least-of-Evils
        - Oversimplification
        - Whataboutism

- **Category 5: Rhetorical Devices:** These techniques seek to influence the audience and control the conversation by using rhetorical methods of persuasion.
  - Exaggeration/Minimisation
  - Loaded Language
  - Thought-Terminating Cliché

The full definitions of the techniques can be accessed in the open-sourced taxonomy: [6]

Appendix B Second Appendix

Table 6 shows the number of rows in PropaGATE in total, per Group, Subgroup, Topic, and View.

Appendix C Third Appendix

The synthetic examples were generated using a single instruction template. For each generation, the placeholder `propaganda definitions` was replaced with the complete list of propaganda techniques and their corresponding definitions, while `example` and `technique list` were substituted with a human-written example and its associated labels to provide in-context guidance.

System instruction

You are an expert Large Language Model capable of generating high-quality synthetic data when given instructions on the desired type of data. You will be generating examples of social media messages that contain some sort of propaganda. You will be given a list of propaganda techniques that includes the names and a detailed descriptions of them. Your task will be to generate a single synthetic example in Bulgarian based on the techniques and a list of the used techniques in the synthetic example by carefully examining and using the definition.

Here are the techniques and their definitions.

Propaganda techniques and definitions `propaganda definitions`

In-context example `example technique list`

Instruction

[INST] Generate an example and a list of the techniques, without any clarifications.[/INST]

During generation, the in-context example was varied across prompts to encourage lexical and topical diversity while maintaining consistency with the annotation schema.

[6] https://github.com/Identrics/wasper.

Group	Subgroup	Topic	View	Rows
Politics & Ideology (738)	General (252)	capitalism (78)	pro_capitalism	23
			against_capitalism	16
			neutral	39
		socialism (78)	pro_socialism	20
			against_socialism	19
			neutral	39
		fascism (76)	pro_fascism	9
			against_fascism	29
			neutral	38
		democracy (6)	against_democracy	3
			neutral	3
		nationalism (13)	pro_nationalism	3
			neutral	10
		liberalism (1)	against_liberalism	1
	Bulgarian Politics (486)	bulgarian politics (446)	unknown	376
			neutral	70
		bulgaria euro (40)	pro_euro	10
			against_euro	10
			neutral	20
Society (444)	Economic Class (160)	rich_people (74)	pro_rich_people	21
			against_rich_people	16
			neutral	37
		low_paid_jobs (86)	pro_low_paid_jobs	31
			against_low_paid_jobs	11
			neutral	43
	Gender & Age Class (222)	women (124)	pro_women	32
			against_women	30
			neutral	62
		men (38)	against_men	19
			neutral	19
		lgbt (56)	against_lgbt	28
			neutral	28
		gender equality (4)	pro_gender_equality	2
			neutral	2
	Rights (62)	abortion (62)	pro_abortion	18
			against_abortion	13
			neutral	31
Religion (234)	General (38)	religion (38)	pro_religion	4
			against_religion	15
			neutral	19
	Christianity (122)	religion – christianity (122)	pro_religion_christianity	33
			against_religion_christianity	28
			neutral	61
	Islam (74)	religion – islam (74)	pro_religion_islam	19
			against_religion_islam	18
			neutral	37
Health, Science & Environment (246)	Health (66)	covid_vaccines (54)	pro_covid_vaccines	9
			against_covid_vaccines	28
			neutral	17
		vaccines (12)	pro_vaccines	2
			against_vaccines	4
			neutral	6
	Environment & Climate (180)	global_warming (122)	pro_global_warming_as_problem	34
			against_global_warming_as_a_problem	16
			pro_global_warming_as_conspiracy	11
			neutral	61
		climate_change (24)	pro_climate_change_as_problem	10
			against_climate_change	2
			neutral	12
		ecology (4)	pro_ecology	2
			neutral	2
		traditional_science_believes (30)	against_traditional_science_believes	15
			neutral	15
TOTAL SHOWN IN TABLE				**1662**

Table 6: Updated topic–view distribution of PropaGATE with original tags.