To make someone do something: mining alert-style directives in Bulgarian social media for low-resource language modelling

Ruslana Margova
GATE Institute
SU St. Kliment Ohridski
ruslana.margova@gate-ai.eu

Stanislav Penkov
GATE Institute
SU St. Kliment Ohridski
stanislav.penkov@gate-ai.eu

Abstract

The work demonstrates how meaningful rhetorical signals can be isolated from a social media dataset even without pre-labelled data or predefined lexicons. By combining unsupervised mining with linguistic theory and interpretable machine learning, the research offers a scalable approach to understanding how language can shape political perception and behaviour in digital spaces. The study focuses on Bulgarian, a morphologically rich, relatively low-resource language, and produces reusable resources—alert constructions, post-level features, and trained classifiers—that are explicitly designed to support low-resource language modelling, including the training and evaluation of neural language models and LLMs for tasks such as content moderation and propaganda-alert detection. The finding that rhetorical salience, not just topical content, drives engagement has implications beyond Bulgarian: it suggests that how something is said may matter as much as what is said in determining a message's viral potential and persuasive impact.

1 The age of post-truth and the alert markers in language

In the age of post-truth[1], where debate is shaped by emotional appeals and the repetition of semi-truths and falsehoods, with factual rebuttals disregarded and the actual truth deemed of secondary importance, rhetorical signals of urgency, threat, or emotional escalation often serve as a specific language of alert. Such language is not a direct synonym of disinformation in the sense of the definition of European Commission, where disinformation is false or misleading content that is spread with an intention to deceive or secure economic or political gain, and which may cause public harm, while misinformation is false or misleading content shared without harmful intent though the effects can be still harmful[2]. Nor is it simply a language of propaganda, defined as the 'systematic dissemination of information', especially in a 'biased or misleading way, to promote a political cause or point of view'[3]. Rather, such linguistic acts function as a rhetorical intensifier: a set of constructions that frame the communicative message as urgent, hidden, or morally charged. We use the term hate speech to refer to any kind of communication in speech, writing or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor[4]. Scholars and policymakers have recognised that hate speech, misinformation and disinformation can have very severe, immediate impacts and that "low-level" examples of all three types of communicative acts could cause severe harm over very long periods of time (Wardle, 2024). Although previous studies have focused on the presence of these kinds of speech in crisis narratives (Kreis, 2017), vaccine debates (Memon and Carley, 2020), or populist communication (De Vreese et al., 2018), the systematic detection, especially of the *alert-style persuasive language* that intensifies such narratives in low-resource languages,

[1] https://dictionary.cambridge.org/dictionary/english/post-truth

[2] https://digital-strategy.ec.europa.eu/en/policies/online-disinformation

[3] https://www.europarl.europa.eu/RegData/etudes/ATAG/2015/571332/EPRS_ATA(2015)571332_EN.pdf

[4] https://peacekeeping.un.org/sites/default/files/report_-_a_conceptual_analysis_of_the_overlaps_and_differences_between_hate_speech_misinformation_and_disinformation_june_2024_qrupdate.pdf

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 62–72

remains underexplored (Leite et al., 2025). In this paper, we focus on such alert markers as a distinctive layer of rhetorical escalation, rather than on hate speech or propaganda as legal categories.

2 Methodology and hypothesis

To investigate these persuasive techniques, we analyse a Bulgarian dataset from the most widely used social media in the country, Facebook[5], all containing the word *пропаганда* ("propaganda"), collected via CrowdTangle between 2014 and 2024. Crucially, because the keyword "propaganda" is present in every entry in the dataset, the task is not to determine whether a post is *propagandistic* in a narrow sense, but to identify which rhetorical and structural features differentiate emotionally resonant, high-engagement posts from less reactive ones. Neither do not comment on the notion of propaganda itself.

Theoretically, we observe *directives*, following Searle's classification of illocutionary speech acts (assertives, directives, commissives, expressives, and declarations, where as a part of directives are orders, requests, advice), which are intended to influence listeners to perform some future actions (Searle, 1975). Based on Austin's theory (Austin, 1975), this category focuses on behaviour modification and, today, it is necessary to employ behavioural interventions against such illocutionary speech acts used as misinformation (Konstantinou and Karapanos, 2025). Our central hypothesis is that specific linguistic constructions work as linguistic stimuli, *as directives*, that frame public reception and provoke action. The linguistic markers in these *directives* could be lexical, morphological or syntactical: e.g., modal verbs ("трябва да" 'must/should'), negation ("няма да" 'will not'), accusatory phrases such as ("те искат" 'they want'), and national identity framings such as ("българите сме" 'we, Bulgarians, are ...').

Additionally, we suppose that these markers, derived from the highlighted alert constructions, will systematically co-occur with every specific topic in the dataset. Such a pattern would indicate that alert-style rhetoric is thematically structured and preferentially deployed within certain rhetorical contexts. We therefore focus on *directive* language-lexico-syntactic patterns that elevate salience or imply concealed urgency by mining naturally occurring constructions from real-world Bulgarian Facebook posts. This fills a gap in the literature by proposing an unsupervised framework for extracting, clustering, and evaluating such constructions in a low-resource setting, grounded in both linguistic theory and empirical feature analysis. Within this context, our work builds on efforts to generalise propaganda detection by identifying language-independent rhetorical devices, such as modality, negation, and repetition.

Methodologically, we conducted a series of NLP experiments:

- dataset-level exploration of Bulgarian Facebook posts containing the keyword "пропаганда", including the construction of total-engagement and reaction–entropy measures and a combined *linguistic weight* to prioritise highly reactive content;
- unsupervised mining of lexico-syntactic constructions (2–5 token n-grams and skip-grams) from a high-engagement subset (top 20% by linguistic weight), using sentence segmentation and POS tagging to filter out purely functional patterns;
- computation of an alert enrichment score and z-normalised alert_z value for each construction, followed by clustering of enriched patterns into rhetorical families (modality, negation, prediction, identity framing, accusatory tone);
- enrichment of each post with alert-sensitive features (e.g., presence and counts of alert constructions, cluster coverage, and aggregated alert scores), stored together with engagement signals for downstream modelling;
- training and inspection of an interpretable linear model (logistic regression) on these alert features to predict whether a post is alert-rich, complemented by an exploratory text-based baseline (TF–IDF + Linear SVM) on the same proxy labels;
- topic–alert overlap analysis using LDA topic modelling to examine how alert con-

[5] https://gs.statcounter.com/social-media-stats/all/bulgaria

structions distribute across major thematic clusters in the corpus.

3 Related work

Research on propaganda detection has largely centred on English-language corpora, with benchmark datasets such as the Propaganda Techniques Corpus (Danziger, 2025) and CoCo (Barrón-Cedeño et al., 2020) offering fine-grained annotation of rhetorical techniques. These studies emphasise the importance of identifying implicit markers such as loaded language, fear appeals, and causal oversimplification. A related strand of work focuses on framing—how linguistic choices influence perception (Entman, 1993)—particularly in political and crisis communication (Rashkin et al., 2017; Mattei et al., 2021). Within these broader phenomena, strategies of incitement in hate speech are widely used to amplify engagement and polarise audiences (Wodak, 2015); (Starbird and Wilson, 2020). Inciting speech, a combination of alert and emotion, is an interpretive construct rather than a conventionalised speech act and needs to be seen in discourse (Danziger, 2025). The definition of inciting is not straightforward, but here we follow the Austinian distinction between illocution and perlocution: hate speech may be an illocutionary act tied to the recognition of a speaker's intention to incite discriminatory hatred, yet it can only be fully characterised if one takes into account the intended perlocutionary effects-that is, the intention of the speaker to trigger a particular kind of response from some audience (Assimakopoulos, 2020). Inciting speech seeks to instil hostility or anger in readers or motivate them to take action against a target group, whether in political (Sweeny, 2019), terrorist (Macdonald and Lorenzo-Dus, 2020), or religious context (Garg et al., 2025). At the computational level, a variety of methods have been used to detect propaganda or emotionally charged narratives. These include traditional classifiers based on surface and lexical features (Barrón-Cedeño et al., 2020), neural models with attention over rhetorical spans (Martino et al., 2020), and hybrid approaches that combine linguistically informed rules with BERT-style transformers (Nakov et al., 2021). Some scholars also identify language-independent rhetorical techniques that can be transferred across languages and domains (Kiesel et al., 2022). Many NLP approaches target large-scale, unmonitored social media data, seeking to identify hate speech and expressions that incite violence (Khan, 2025; Veeramani et al., 2023). The propaganda detection in social media in Slavic languages with prelabeled data by fine-tuning for classification on LLMs and focusing on F1 scores is made by Loginova (Loginova, 2025). Here, we are providing linguistically motivated features for the multi-label classification task. Most of the previous systems are trained on English, with limited attention to morphologically rich or low-resource languages such as Bulgarian. For Bulgarian, existing work has focused primarily on deception and disinformation, creating an extended hierarchical classification of linguistic markers signalling deception, lists of Bulgarian expressions for recognising some of these markers, and several large social media datasets on deception-related topics that are automatically annotated with such markers (Temnikova et al., 2023), others add specific morphological features for disinformation detection (Margova, 2023).

4 Data

The analysed data set consists of 13,989 Bulgarian-language Facebook posts collected from CrowdTangle between 2014 and 2024. The accounts are public and anonymised, and we will not share the raw data, in accordance with the new Meta policy[6]. The collection used the keyword *пропаганда*, which introduces topical bias toward posts that explicitly reference propaganda, but offers a unique opportunity to examine how alarmist and persuasive rhetorical devices are deployed in this narrative context. The cleaned and anonymised data set (`propaganda_bg_media.json`) contains a unified text field for each post (aggregating the available message and title fields), together with metadata such as page name, page category, creation date, and content type, as well as engagement indicators (reactions, comments, shares). Very short and duplicate posts and posts without usable text are removed during preprocessing. Following exploratory analysis,

[6] https://transparency.meta.com/researchtools/meta-content-library/

we compute two interaction-based signals for each post: (i) a total engagement score, defined as the sum of all reactions, comments, and shares; and (ii) a reaction-entropy measure capturing how dispersed the audience reactions are across reaction types. We then define a *linguistic weight* as the product of total engagement and reaction entropy. This quantity favours posts that are both widely visible and emotionally polarising, and it is later used to select a high-engagement subset and to weight linguistic constructions during alert-pattern mining.

5 Experiments and results

5.1 Identifying alert-style constructions

To identify patterns of heightened emotional or rhetorical activation—our operationalisation of alert-style directive language—we first select a high-engagement subset corresponding to the top 20% of posts ranked by linguistic weight. These 2,798 posts are parsed with Stanza (Qi et al., 2020) using the Bulgarian UD-BTB treebank models (Simov et al., 2005) for sentence segmentation and part-of-speech (POS) tagging. From the tagged sentences, we extract several million 2–5 token spans (n-grams and skip-grams) and filter out punctuation-only strings, frequent purely functional sequences, and spans consisting entirely of stopwords. This process yields 18,837 distinct multiword constructions.

Let $f_{\text{high}}(c)$ denote the number of times a construction c appears in the high-engagement subset, and $f_{\text{all}}(c)$ its count in the full dataset. Each construction is assigned an alert score:

$$\text{alert\_score}(c) = \frac{f_{\text{high}}(c)}{f_{\text{all}}(c) + 1}, \qquad (1)$$

that is, the proportion of its occurrences that come from highly engaged, high-entropy posts, with add-one smoothing for rare patterns. Scores are then z-normalised across all constructions to yield alert_z. High alert_z values indicate constructions that are disproportionately associated with highly engaged, emotionally reactive content. Example constructions are shown in Table 1.

5.2 Clustering

To organise the enriched constructions into broader rhetorical families, we embed each multiword pattern in a lexical feature space and cluster them with K-means. We take the 18,837 multiword constructions (2–5 token spans) and represent them using a TF–IDF-based vectorisation over lexical forms, then run K-means with $k = 20$ clusters. This yields compact families of constructions sharing similar lexical and syntactic profiles.

Cluster-level summaries reveal that nine of these twenty clusters exhibit clearly elevated mean alert_z scores. Qualitative inspection shows that these enriched clusters align with distinct rhetorical modalities:

- **Modality (obligation/authority)** – e.g., constructions with “трябва да” (‘must/should’) and related modal verbs;
- **Negation (refusal/resistance)** – e.g., “няма да” (‘will not’), “не искаме” (‘we do not want’);
- **Future prediction and inevitability** – constructions describing upcoming or unavoidable events;
- **National identity framing** – references to collective identity (e.g., “ние, българите” ’we, Bulgarians, “в нашата страна” ’in our country’);
- **Accusatory tone** – patterns attributing malign intent to a vaguely defined outgroup (e.g., “те искат” ‘they want’).

These data-driven families correspond closely to categories discussed in political discourse analysis (Fairclough, 2013), suggesting that theoretically grounded notions such as modality, negation, and identity framing can emerge from unlabeled data through construction-level clustering. The clustering stage produces several intermediate resources, including the full list of scored constructions (`alert_language_constructions.csv`), ranked clusters (`alert_clusters_ranked.csv`), cluster representatives (`cluster_representatives.csv`), and POS-based profiles (`cluster_pos_summary.csv`). Detailed cluster tables and additional examples are provided in the Appendix.

To facilitate visual inspection of the structural organisation of these construction families, we project representative constructions from

Construction	POS pattern	Alert_z	Example
„пропаганда в“	NOUN+ADP	4.23	„пропаганда в България“(propaganda in Bulgaria)
„трябва да“	VERB+PART	3.81	„трябва да се вземе решение“(a decision has to be taken)
„няма да“	ADV+PART	3.40	„няма да позволим това“(we would not alow it)
„в България“	ADP+PROPN	2.97	„в България се случва нещо важно“(smth important happens in Bulgaria)

Table 1: Example constructions with their POS patterns and alert_z scores.

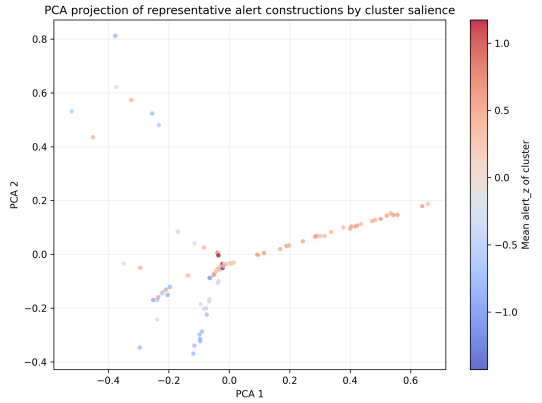

Figure 1: PCA projection of representative multiword constructions derived from K-means clustering ($k = 20$) of 18,837 POS-filtered constructions. Each point corresponds to a representative construction from a construction family. Colour indicates the mean alert_z of constructions within the corresponding cluster, highlighting rhetorical families enriched in high-engagement, high-entropy posts.

each cluster into a two-dimensional PCA space (Figure 1). Rather than plotting all 18,837 constructions, which would produce a dense and visually opaque projection, we visualise a set of representative multiword patterns per cluster. Each point in Figure 1 corresponds to one such representative construction. The colouring reflects the mean alert_z score of the cluster to which the construction belongs, allowing rhetorical families enriched in high-engagement, high-entropy discourse to be distinguished from less salient clusters.

5.3 Alert features ranking

Using the ranked constructions, we next compute a set of post-level alert features by mapping the constructions back onto every post in the corpus. For each post, we derive:

- a binary flag indicating whether the post contains at least one alert construction;
- an *alert term count*: the total number of matched alert spans;
- summary alert scores (`alert_score_mean` and `alert_score_max`) over all matched constructions in the post;
- a `cluster_count` feature capturing how many distinct alert clusters are represented in the post.

These alert-specific features are stored, together with the original text and the engagement signals, in `posts_with_alert_features.jsonl` and exported as a tabular feature matrix (`model_features.csv`) for downstream modelling. In the present work, our main linear model uses only the alert-specific features, which keep the feature space compact and directly interpretable in terms of the mined constructions.

5.4 Modelling and evaluation

We analyse how informative the unsupervised alert features are for distinguishing alert-rich posts from the rest of the corpus. Since no human labels are available, we derive a binary label from the alert scores themselves: posts whose `alert_score_max` lies in the top 20% are labelled as alert (1), and all others as non-alert (0). This yields 2,798 positives and 11,191 negatives. We randomly split the data into 80% training and 20% test sets.

5.4.1 Logistic regression on alert features

We first train a logistic regression model in scikit-learn using the alert-specific features `alert_count`, `alert_score_mean`, `alert_score_max`, and (when available) `cluster_count`. In the held-out test set, the model achieves an overall accuracy of 0.651. For the alert class (1), precision is 0.296, recall is 0.539, and F1 is 0.382. These scores confirm that the automatically derived alert features carry a non-trivial predictive signal, but also highlight the limitations of using self-derived labels as a ground truth.

Inspection of the learned coefficients shows that `alert_score_max`—capturing the single most rhetorically charged construction in each post—has by far the highest positive weight, while `cluster_count` contributes a smaller positive effect, and the remaining features play a more moderate or slightly negative role. Table 2 summarises the relative importance of the main features.

Table 2: Top predictive features from logistic regression model

Feature Name	Relative Importance
alert_score_max	0.49
cluster_count	0.08
alert_count	0.02
alert_score_mean	-0.04

Taken together, these results suggest that the concentrated rhetorical intensity—one or a few strongly enriched alert constructions — is more characteristic of alert-rich posts than a diffuse accumulation of many moderately alert expressions. We return to this point in the Discussion, where we interpret it in terms of directive illocutionary force and mobilising stimuli.

5.4.2 Exploratory baselines

For completeness, we also experimented with a stronger but less interpretable text-based baseline: a TF–IDF + Linear SVM classifier trained on the full Bulgarian post text. This model achieves substantially higher performance on the same automatically derived labels (see Appendix), confirming that the label signal is easily recoverable from the raw text. However, it does not address the inherent circularity of the labels and offers limited insight into which specific alert constructions drive the predictions. For this reason, we treat this experiment as an exploratory sanity check and focus our analysis on the interpretable construction-based model.

5.5 Topic–alert construction overlap

To understand how alert-style constructions interact with broader narrative themes, we apply Latent Dirichlet Allocation (LDA) with $k = 10$ topics to the post texts, using the same preprocessed dataset as in the exploratory analysis. Each topic yields a coherent set of high-frequency terms, covering domains such as national politics, war, energy and economics, identity, and institutional trust.

Using the topic assignments from LDA, we then examine, for each topic, which alert constructions are overrepresented. Concretely, we compute topic-wise frequencies of the mined alert constructions and identify those whose relative frequency within a topic is higher than in the dataset overall. This analysis reveals that obligations and inevitability patterns (e.g., “трябва да” ‘must/should’, “може да” ‘may/can’, “няма да” ‘will not’) are especially concentrated in topics centred on domestic politics and policy debates, while Russia/war-related constructions (e.g., “пропаганда на”, references to some Russian actors and locations) cluster in topics associated with geopolitical conflict. Normalised profiles show that several high-alert constructions, such as “в България” ‘in Bulgaria’, “няма да” ‘will not’, “руската пропаганда” ‘russian propaganda’, “трябва да” ‘must/should’, and “част от” ‘part of’, appear across multiple topics with elevated preference.

Semantically, these topics are related to the concept of Bulgaria, differing mainly in what **should** and **should not** happen in the country. These findings support our interpretation of alert language as a cross-cutting rhetorical layer: rather than belonging to a single thematic domain, the same directive, predictive, and identity-framing constructions are reused to intensify salience and emotional charge across issues. This matters both for discourse analysis and for practical monitoring: if the same constructions recur across topics, systems that track only topical keywords risk missing stylistically alert but thematically novel content. From an NLP perspective, the topic–alert overlap suggests that alert-aware representations can be reused as domain-general features, rather than re-engineered for each new issue or crisis. Additional descriptions of the topics and their associated alert constructions are summarised in Appendix C.

5.6 Contribution to low-resource NLP

This work contributes new resources and methodology for low-resource NLP, particularly for morphologically rich languages such as Bulgarian. Specifically, we provide:

1. A corpus of 13.9K Facebook posts in Bul-

garian, enriched with automatically derived alert features and proxy alert labels;

2. Over 18K ranked multiword constructions with z-normalised alert enrichment scores;

3. A full post-level feature matrix (`model_features.csv`) based on alert counts, scores, and cluster coverage;

4. An interpretable alert-feature set that supports lightweight linear models and facilitates qualitative analyses of rhetorical escalation.

By combining engagement-based proxies, unsupervised construction mining, and simple interpretable models, the approach enables resource-efficient analysis of directive and alert-style rhetoric in political discourse, and can be adapted to other low-resource languages with minimal manual annotation effort. These artefacts are intended to serve directly as training data, auxiliary supervision, and evaluation benchmarks for neural language models and large language models, with a particular focus on persuasion-aware low-resource language modelling in Bulgarian.

6 Discussion

The findings of this study, despite the constraints outlined in the Limitations section, suggest that alert-style language is not merely an artefact of lexical frequency or topic bias, but a cross-cutting rhetorical strategy observable across a wide range of Bulgarian Facebook posts tagged with the term "propaganda." The construction clusters identified align with established rhetorical categories such as modality ("трябва да" 'must/should', "може би" 'maybe'), negation ("няма да" 'will not', "не искаме" 'we do not want'), prediction and urgency ("ще се случи" 'it will happen', "нещо голямо идва" 'something big is coming'), and identity appeals ("българите сме" 'we, Bulgarians, are', "в нашата страна" 'in our country'). These patterns resonate with classic theories of persuasion and political discourse. For instance, modal constructions are known to encode obligation and authority (Fairclough, 2003), while negation and future-tense framing are key tools for shaping imagined threats or promises (Ulfatovna, 2025). Showing that such structures can be recovered through unsupervised methods, even in the absence of labelled data or predefined lexicons, indicates that meaningful rhetorical signals can be isolated from large social media dataset.

Importantly, these signals also appear predictive when evaluated against our automatically derived proxy labels. Post-level feature importance analysis points to `alert_score_max` and, to a lesser extent, `cluster_count`, as informative for distinguishing alert-rhetoric posts from the rest of the corpus, whereas diffuse accumulations of many moderately enriched constructions play a smaller role. This supports an interpretation of alert-style language as relying on concentrated spikes of directive and evaluative force, rather than on sheer volume alone. Given that the labels are derived from engagement-weighted alert scores, however, we interpret these results as evidence of internal coherence in the unsupervised signal, rather than as proof that the model detects manipulation or persuasion in a normative sense.

The overlap analysis with topic models further underscores that alert language is not domain-bound but stylistically pervasive, appearing in topics related to war, national politics, energy and economics, and broader ideological conflict alike. Obligations and inevitability patterns tend to cluster in topics associated with domestic policy debates, while Russia/war-related constructions concentrate in conflict-related topics. At the same time, this analysis inherits the limitations of bag-of-words LDA and of the underlying corpus: topics are coarse, can partially overlap, and are themselves shaped by the initial propaganda keyword filter. Overall, the results should therefore be seen as mapping a distinctive layer of directive and alert-style rhetoric within a specific propaganda-related discourse, rather than as a complete account of persuasive language in Bulgarian social media. We do not directly measure the effectiveness or behavioural outcomes of alert-style directives; instead, we focus on their linguistic form, engagement correlates, and suitability as structured input to low-resource language modelling pipelines.

Relation to large language models (LLMs). Modern large language models provide complementary ways of analysing per-

suasive and directive language, for example, through zero- or few-shot prompting for stance, intent, or rhetorical framing at the post level. Such approaches can capture broader discourse-level cues that are not directly tied to short surface constructions. At the same time, they typically produce less transparent and less easily reusable representations than the construction-level resources developed here. The present work, therefore, emphasises interpretable, language-specific units that can be directly inspected and integrated into lightweight models in a low-resource setting. Future work can explore hybrid approaches in which LLMs support interpretation, paraphrasing, or cross-lingual extension of alert-style constructions, while construction-level representations continue to provide a transparent structural backbone.

7 Limitations

Our analysis is constrained by several factors. First, the dataset is collected using a single keyword ("пропаганда"), which introduces meta-discourse bias and limits the generalisability of the findings beyond posts that explicitly invoke propaganda. Second, engagement-based quantities (total engagement, reaction entropy, and the derived linguistic weight) conflate visibility, controversy, and potential manipulation, and do not distinguish between fact-checking, critique, and genuinely manipulative content. Third, the proxy labels used for supervised modelling are defined directly in terms of engagement-weighted alert scores, which introduces an element of circularity: the model is trained to predict a label that is itself constructed from the same enriched constructions. Fourth, temporal variation is another important factor: the dataset spans 2014–2024, a period during which political events, crises, and media narratives evolved substantially. Alert constructions associated with particular topics may shift over time as discourse norms and public concerns change. Future work should explicitly model temporal dynamics to examine how alert-style rhetoric adapts to emerging events and whether certain constructions gain or lose salience in different periods. Fifth, we lack independent human annotation of alert-style rhetoric, and we do not observe behavioural outcomes; our conclusions are therefore limited to internal coherence of the unsupervised signal. Finally, Bulgarian-specific morphology, discourse conventions, and political context mean that the same pipeline may require adaptation to transfer to other low-resource languages and may not recover identical rhetorical families elsewhere.

8 Ethical considerations

The dataset used in this work consists of public Facebook posts collected via CrowdTangle between 2014 and 2024, restricted to Bulgarian-language content containing the keyword "пропаганда". CrowdTangle only exposes content from public pages and other public-facing accounts (e.g., media outlets, institutions, political actors), and does not provide access to private user posts. As it has already mentioned, we will not share the raw data and our work is linked to public interest. Our analysis is conducted at the level of aggregated linguistic patterns across the corpus, and we do not attempt to profile or infer attributes of specific pages, organisations, or individuals. Nevertheless, the dataset may contain sensitive political opinions, hate speech, or personal attacks. Any future release of resources derived from this work (e.g., ranked constructions, feature matrices, trained models) would need to comply with platform terms of service and applicable data protection regulations, and should avoid exposing personally identifiable information. The alert constructions and models described here could, in principle, be used either to optimise persuasive messaging or to support mitigation; we therefore emphasise their intended use in research, education, and content moderation contexts, and recommend coupling them with transparent, human-in-the-loop workflows.

9 Conclusions and future work

This paper presents an unsupervised framework for mining alert-style directive language in Bulgarian social media and packaging it as structured resources for low-resource language modelling, while also illuminating how propagandistic content is framed. The contributions are: (1) a large-scale dataset of Facebook posts containing the term "propaganda"; (2) a pipeline for extracting and clustering alert con-

structions; (3) post-level representations that aggregate alert-specific features together with engagement signals; and (4) a demonstration of lightweight, interpretable modelling techniques over these features.

In addition to its technical contributions, this research provides a linguistically informed approach to low-resource language processing, bridging data-driven discovery with rhetorical theory. We demonstrate that even with keyword-based content filtering, meaningful linguistic variation emerges that helps explain how messages resonate and spread. Key resources contributed to Bulgarian language processing are: 13,989 Facebook posts enriched with automatically derived alert features; approximately 18,000 ranked linguistic constructions with alert enrichment scores; and a complete post-level feature matrix combining alert counts, scores, cluster coverage and engagement-based signals. In general, we provide an interpretable, lightweight modelling approach requiring minimal manual annotation. The research operationalises Searle's concept of directive illocutionary acts in digital environments, demonstrating how linguistic stimuli in social media discourse influence behaviour and perception. The findings highlight the function of language in "post-truth" environments, where emotional appeals and rhetorical urgency dominate over factual accuracy. The alert construction database facilitates the automated detection of emotionally charged, directive constructions within Bulgarian social media, that could be used for political purposes. Furthermore, the identified patterns can inform educational work on teaching persuasive language techniques and support future efforts to build more robust, human-validated systems for monitoring manipulative rhetoric. The results could be used for developing effective crisis communication and early detection of disinformation.

Acknowledgments

This research is funded by GATE project: Horizon 2020 WIDESPREAD-2018-2020, TEAMING Phase 2 Program under grant agreement no. 857155, the Program "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 Program of the European Union under contract number 101083730. Authors thanks to Ivo Emanuilov for his help in licensing the dataset.

References

Stavros Assimakopoulos. 2020. Incitement to discriminatory hatred, illocution and perlocution. *Pragmatics and Society*, 11(2):177–195.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, volume 2696, Thessaloniki, Greece. CEUR Workshop Proceedings. ArXiv:2007.07997.

Roni Danziger. 2025. Interpretive constructs: The case of incitement to violence and terror. *Journal of Pragmatics*, 245:35–49.

Claes H De Vreese, Frank Esser, Toril Aalberg, Carsten Reinemann, and James Stanyer. 2018. Populism as an expression of political communication content and style: A new perspective. *The international journal of press/politics*, 23(4):423–438.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Norman Fairclough. 2003. *Analysing Discourse: Textual Analysis for Social Research*. Routledge, London.

Norman Fairclough. 2013. *Language and power*. Routledge.

Vaibhav Garg, Ganning Xu, and Munindar P Singh. 2025. Understanding inciting speech as new malice. *IEEE Transactions on Computational Social Systems*.

Muhammad Shahid Khan. 2025. *Analysis of Violence Incitation in Social Media from Urdu Content using NLP Techniques*. Ph.D. thesis, CAPITAL UNIVERSITY.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4459–4471.

Loukas Konstantinou and Evangelos Karapanos. 2025. Behavior change interventions combating online misinformation: A scoping review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Ramona Kreis. 2017. The "tweet politics" of president trump. *Journal of Language and Politics*, 16(4):607–618.

João A Leite, Olesya Razuvayevskaya, Carolina Scarton, and Kalina Bontcheva. 2025. A cross-domain study of the use of persuasion techniques in online disinformation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1100–1103.

Ekaterina Loginova. 2025. Fine-tuned transformers for detection and classification of persuasion techniques in slavic languages. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 151–156.

Stuart Macdonald and Nuria Lorenzo-Dus. 2020. Intentional and performative persuasion: The linguistic basis for criminalizing the (direct and indirect) encouragement of terrorism. In *Criminal Law Forum*, volume 31, pages 473–512. Springer.

Ruslana Margova. 2023. *Linguistic Markers of Fake News*. Ph.D. thesis, Sofia University St. Kliment Ohridski. Doctoral dissertation. Open access.

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

Josiemer Mattei, Katherine L Tucker, Luis M Falcón, Carlos F Ríos-Bedoya, Robert M Kaplan, H June O'Neill, Martha Tamez, Sigrid Mendoza, Claudia B Díaz-Álvarez, Jonathan E Orozco, and 1 others. 2021. Design and implementation of the puerto rico observational study of psychosocial, environmental, and chronic disease trends (prospect). *American Journal of Epidemiology*, 190(5):707–717.

Shahan Ali Memon and Kathleen M. Carley. 2020. Characterizing COVID-19 misinformation communities using a novel twitter dataset. In *Proceedings of the 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN 2020), co-located with CIKM*, pages 1–9, Online. ArXiv:2008.00791 [cs.SI].

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*, Montreal, Canada (online). International Joint Conferences on Artificial Intelligence Organization. ArXiv:2103.07769.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 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 (EMNLP)*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

John R Searle. 1975. A taxonomy of illocutionary acts.

Kiril Simov, Petya Osenova, and 1 others. 2005. Bultreebank: Building a bulgarian treebank. In *Proceedings of a Treebank-related workshop*.

Kate Starbird and Thomas Wilson. 2020. Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School (HKS) Misinformation Review*.

JoAnne Sweeny. 2019. Incitement in the era of trump and charlottesville. *Cap. UL Rev.*, 47:585.

Irina Temnikova, Silvia Gargova, Ruslana Margova, Veneta Kireva, Ivo Dzhumerov, Tsvetelina Stefanova, and Hristiana Krasteva. 2023. New bulgarian resources for studying deception and detecting disinformation.

Rakhmonova Amira Ulfatovna. 2025. Linguistic characteristics of political discourse. *FRONTIERS OF KNOWLEDGE AND INTERDISCIPLINARY DISCOVERY*, 1(1):366–374.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023. Lowresourcenlu at blp-2023 task 1 & 2: Enhancing sentiment classification and violence incitement detection in bangla through aggregated language models. In *BLP 2023-1st Workshop on Bangla Language Processing, Proceedings of the Workshop*, pages 273–278. Association for Computational Linguistics.

Claire Wardle. 2024. A conceptual analysis of the overlaps and differences between hate speech, misinformation and disinformation. *Department of Peace Operations (DPO). Office of the Special Adviser on the Prevention of Genocide (OSAPG). United Nations*.

Ruth Wodak. 2015. *The Politics of Fear: What Right-Wing Populist Discourses Mean.* SAGE Publications, London, UK.

A Alert-language resources

For completeness, we briefly describe the main artefacts produced by the alert-mining pipeline. The file `propaganda_bg_media.json` contains the cleaned dataset of 13,989 Facebook posts, together with metadata and engagement counts. The file `alert_language_constructions.csv` lists all multiword constructions (2–5 token spans) extracted from the corpus, along with their counts in the full dataset and in the high-engagement subset, the corresponding alert_score values, and their z-normalised alert_z scores. The file `alert_clusters_ranked.csv` contains the K-means clusters with their mean alert_z scores and sizes, while `cluster_pos_summary.csv` summarises the dominant POS patterns per cluster. Representative examples for qualitative analysis are provided in `cluster_representatives.csv` and `highlighted_alert_constructions.csv`.

A.1 Example mined constructions with English glosses

Table 3 lists illustrative high-alert constructions (short multiword spans) produced by the mining pipeline, together with approximate English glosses. Constructions are shown as mined (2–5 token spans), rather than full sentences.

Bulgarian construction	**English gloss**
е виновен	is guilty / is to blame
не бива да се	one should not / must not
бива да се	one may / is allowed to
няма как да	there is no way to / cannot
да се случи	to happen / to occur
ще стане	it will happen / it will become
истината е	the truth is
страната ни	our country

Table 3: Examples of mined high-alert constructions (multiword spans) with approximate English glosses.

Note: Constructions are short spans and may require surrounding context for a fully grammatical sentence-level translation.

B Modelling artefacts

The post-level alert features described in the main text are stored in `posts_with_alert_features.jsonl` and exported as a tabular feature matrix `model_features.csv`. The logistic regression model trained on these features is saved as `slm_alert_model_logreg.pkl`. Its full classification report and confusion matrix on the held-out test set are available in the accompanying notebook `04_Linear_Model_Training_for_Alert_Language_Detection.ipynb`.

For exploratory comparison, we also train additional text-based baselines on the same proxy labels (e.g., a TF–IDF + Linear SVM classifier). Their implementation details and evaluation outputs (saved model files and prediction arrays) are included in the project materials but are not central to the analysis in the main paper.

C Topics and alert constructions

The topic–alert overlap analysis is based on an LDA model with $k = 10$ topics trained on the preprocessed corpus. For each topic, we compute the relative frequency of each alert construction and identify those that are overrepresented compared to the dataset as a whole. The full topic lists and topic-specific alert profiles are provided in the notebook `06_Topic_Alert_Constructions_Overlap_Analysis.ipynb` and the associated exported files.