Political Stance Detection in Estonian News Media

Lauri Lüüsi
University of Tartu
lauri.luusi@ut.ee

Uku Kangur
University of Tartu
uku.kangur@ut.ee

Roshni Chakraborty
ABV IIITM Gwalior
roshni@iiitm.ac.in

Rajesh Sharma
University of Tartu
rajesh.sharma@ut.ee

Abstract

Newspapers have always remained an important medium for disseminating information to the masses. With continuous access and availability of news, there is a severe competition among news media agencies to attract user attention. Therefore, ensuring fairness in news reporting, such as, politically stance neutral reporting has become more crucial than before. Although several research studies have explored and detected political stance in English news articles, there is a lack of research focusing on low-resource languages like Estonian. To address this gap, this paper examines the effectiveness of established stance-detection features that have been successful for English news media, while also proposing novel features tailored specifically for Estonian. Our study consists of 32 different features comprising of lexical, Estonian-specific, framing and sentiment-related features out of which we identify 15 features as useful for stance detection.

1 Introduction

With the rise of the internet, the information-seeking behavior has undergone a shift such that news media has pivoted away from traditional printed newspapers towards social media and online platforms (Chakraborty et al., 2017, 2019). Furthermore, in today's interconnected world, online news articles becoming readily available within minutes of an event occurring (Bucy et al., 2007; Chakraborty and Chakraborty, 2023). While users rely on the news media agencies for a fair and high quality reporting of news events, there has been several instances of deviation from journalistic news values in news reporting (Tandoc et al., 2021), such as, deliberately lying or leaving out context, not fact-checking sources, using clickbait, being biased (Spinde, 2021), using politically aligned news reporting (Park et al., 2022; Chakraborty et al., 2020), etc. However, with this massive growth in news media and shift in news consumption behavior, it has become increasingly challenging and time-consuming to manually verify bias in news articles and ensure that the news articles follows journalistic standards. This is especially true for low-resource languages, where building machine learning based solutions is often more difficult due to the lack of training data. Therefore, it is important to develop and explore techniques, which use automatically extracted text features as a way to gain insight and monitor news media.

Although there are several forms of bias in news media, in this paper, we focus on political leaning or stance in news articles. While few automated stance detection concerning political leaning has been explored for different topics in English news media, such as political elections and candidates, climate change, COVID-19, and abortion rights, more extensive research is needed to enhance understanding and accuracy (ALDayel and Magdy, 2021; Farsi et al., 2024; Mohammad et al., 2016a; Neha et al., 2022; Fisher et al., 2023; Baxi et al., 2022; Chakraborty et al., 2022). However, political stance detection in Estonian news media is a mostly unexplored topic. Estonian, spoken by about a million people, has a much smaller language corpus with around 3 billion words compared to English, which has over 1.4 billion speakers (Dyvik) and a corpus of 800 billion words (Piir, 2023). Meanwhile, the rise of online news media in Estonia is significant. For example, Delfi Meedia, a major Estonian media company, has over 700,000 monthly readers and has amassed over 100,000 paid online subscribers by 2023 (Delfi Meedia; Eesti Meediaettevõtete Liit). With such growth, the need for automated systems to verify news articles and detect political stances is essential.

Developing automated approaches for low-resource languages can be challenging, as these smaller languages are particularly affected by the

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 12–28
November 28-29, 2024 ©2024 Association for Computational Linguistics

non-availability of task and domain-specific data (Hedderich et al., 2021). Furthermore, identifying labeled data requires manual annotation, which is time and cost intensive. Compared to English, the data can be of lower quality, which can lead to poorer results and varied performance. Therefore, it is particularly challenging to build automated models, especially train large language models, for political stance detection in Estonian news media.

In this paper, political stance in Estonian news media is analyzed on the target of immigration. Immigration is a concept that encompasses the international movement of people, usually foreign nationals (22, 2019). Immigration is a suitable target for automated stance detection, as stances towards it are varied and can often veer towards extremes (Päll, 2021). Immigrants can be viewed as strong and talented workers with great potential or, conversely, burdens on society who take jobs from locals and will not integrate into the local culture (Kosho, 2016). Media coverage of immigration can influence public opinion, especially when it adopts an overly negative stance. These shifts in attitude can potentially translate to negative treatment of immigrants, fueling racism and social division, and the enactment of discriminatory policies (Vetik, 2000). We explore and identify relevant features and techniques indicative of political stance in Estonian news media. The study identifies 15 significant features out of 32 for detecting political stance in Estonian news media. Sentences opposing immigration are longer, more complex, and used more adjectives and quotes, indicating emotionally charged language. Content analysis shows that anti-immigration texts mentioned destinations like Sweden and Germany, while supportive texts focus on transit countries like Greece and Turkey, highlighting different framing strategies. Estonian-specific features like conditionals and translatives are more prevalent in both supportive and against stances. Framing analysis uncovers distinct language use based on stance: negative terms like *illegaalne immigrant (illegal immigrant)* and *neeger (nigger)* for opposition, versus more politically neutral ones like *aafrika päritolu (African origin)* for support, highlighting contrasting frames in legality and humanity. Sentiment analysis shows that the XLM-RoBERTa model outperforms others, achieving the highest F1-scores across all stances.

The organization of the paper is as follows. Section 2 gives an overview of the dataset and describes the preprocessing step followed by the proposed methodology in Section 3. We discuss the exhaustive analysis of extracted features and their usefulness in political stance detection in Section 4 and finally, conclude in Section 5.

2 Dataset

We use the dataset described by Mets et al. (2023) who collected 266 628 news articles from two Estonian news providers - Ekspress Grupp[1] (the parent company of Delfi Meedia) and Uued Uudised[2] between 2015 and 2022 on the topic of Immigration. The target is immigration, and the text is a topic-related sentence. The dataset comprises of 3261 sentences out of which 1175 sentences are of against stance, 1597 neutral stance, and 489 supportive stance towards immigration. For our study, we consider only the text of the news article. While additional meta-features, such as the title, author, publication date, and publisher, are available, they are not considered as they require prior outside knowledge about a media outlet or author and their stance on specific issues. An overview of the dataset is illustrated in Table 1. The dataset is publicly accessible on GitHub[3]. The code and implementation details are available on GitHub [4].

Stance	Number of Sentences
Against	1175
Neutral	1597
Supportive	489

Table 1: Distribution of sentences in the dataset

Preprocessing Details We employ Estonian language specific preprocessing. For example, two letters with diacritics (š, ž) and sentences in the dataset which had these letters were represented by question marks or other nonsensical symbols. In order to fix this, we used EstNLTK's SpellCheckRetagger[5], a tool that identifies misspellings and adds corrected forms (Laur et al., 2020), i.e., misrepresented letter was replaced by either *š* or *ž* and further, validated both by POS tagger and spell check. Additionally, we removed repeated symbols

[1] https://ekspress.delfi.ee/
[2] https://uueduudised.ee/
[3] https://github.com/markmets/immigration-prediction-EST
[4] https://github.com/laurilyysi/EstonianStanceDetection
[5] https://github.com/estnltk/

Dataset Sentence	English Translation
aastaga tuli Eestisse 22 000 “ajutist” töölist	in a year, 22,000 “temporary” workers came to Estonia
võttes vastu inimsmugeldajate “ohvreid” aafriklaste ja teiste migrantide näol.	by accepting “victims” of human traffickers in the form of Africans and other migrants.
“Sallivuslased” aitavad neil oma soovituste ja muu “abiga” ennast hädalistena tutvustada ja mõrvarid seavad ennast “pagulastena” Euroopas sisse.	“Tolerance advocates” help them with suggestions and other “assistance” so they could present themselves as sufferers, as murderers establish themselves as “refugees” in Europe.

Table 2: Sentences where quotation marks are used to express doubt or irony.

Estonian Sentence	English Translation
Aga selleks ju migrandipaadid **kehvakesed** ongi, ja ilmselt lastakse need mõnda laeva märgates meelega vett täis.	But that’s exactly why migrant boats are so **flimsy**, and presumably they are intentionally filled with water when spotted by a ship.

Table 3: Use of diminutives in a sentence from the dataset to express a stance.

and fixed issues with missing punctuation.

3 Methodology

In this Section, we study different features that can help in identifying the political stance of Estonian news media segregated into lexical features, features specific to the Estonian language, framing-related features, and sentiment features. A complete list of all the features used is shown in Appendix A.

3.1 Lexical features

Lexical features are related to the grammar and construction of words. We consider the following lexical features, such as word count, dependency tree height, Flesch Reading Ease Score (FRES), named entities, noun phrases, adjectives, quotes, and quoted phrases. For calculating lexical features, we use EstNLTK (Tkachenko et al., 2013; Maide, 2020; Laur et al., 2020).

In order to understand sentence complexity and readability, we consider **Dependency tree height** and **Flesch Reading Ease Score (FRES)**. Dependency trees map the grammatical relationships within a sentence and indicate complexity through their height, such as, a taller tree suggests a more complex sentence structure (Nivre, 2010). Subsequently, FRES provides a numerical indication of readability, combining average sentence length and syllable count to generate a score where a higher value signifies easier readability (Zamanian and Heydari, 2012). These metrics help discern whether sentences are structured in ways that might simplify or complicate the reader’s understanding.

Further, we study **named entities** and **noun phrases** to identify the difference in framing of sentences across different stances. These features are essential for extracting the thematic substance of texts and for understanding the emphases within a narrative (Erelt, 2013). **Adjectives** and **quotes** significantly influence the tone and suggestiveness of sentences, therefore can aid in understanding how the usage of these can impact reader perception. For example, the use of adjectives describe and modify nouns, potentially imbuing them with positive or negative connotations that can subtly influence the reader’s perception of the discussed topics. Quotes, whether marking direct speech or emphasizing irony, can alter the meaning conveyed by sentences. The use of quotes can imply skepticism or irony, potentially shifting the interpreted meaning of the text as shown in Table 2 (Schlechtweg and Härtl, 2023; van den Berg and Markert, 2020).

3.2 Estonian-specific features

We describe features that could be indicative of stance and are specific to the Estonian language. The Estonian language is morphologically complex (Mets et al., 2023) due to the abundance of verb conjugation forms and grammatical cases for nouns and adjectives (Ehala, 2009; Argus, 2009). Although these characteristics can make analyzing Estonian texts challenging, it also aids in identifying features for stance detection. Additionally, to the

Estonian Sentence	English Translation
Massimigratsiooni mahitajad aga **ujutaksid** kontinendid pigem migrantidega üle ja **segaksid** ära kogu maailma rahvastiku.	The proponents of mass migration, however, would rather **flood** continents with migrants and **mix up** the entire world population.

Table 4: Use of the conditional form in sentences from the dataset to express a stance.

Estonian Sentence	English Translation
Suurim probleem ongi see, et kogu Euroopa on sunnitud migrantidega tegelema [—]	The **biggest** problem is that the whole of Europe is forced to deal with migrants [—]

Table 5: Use of the superlative form in a sentence from the dataset to express a stance.

best of our knowledge, we could not find any existing research work which analyze Estonian-specific features in detail with respect to stance detection. We discuss the following Estonian-specific features: diminutives, superlatives, conditional form, translative case and indirect speech next.

Diminutives in Estonian are formed by adding the suffixes *-ke* or *-kene* to nouns and adjectives. This can alter the emotional tone of a word to express either affection or belittlement (Liivak, 2023; Kasik, 2015). For instance, the diminutive form of *lollike (stupid)* can imply a lack of concern, subtly shifting the stance. English does not have a consistent suffix for diminutive words, unlike Estonian where forming the diminutive is mostly uniform across nouns and adjectives. For example, Liivak (2023) highlights that out of 143 instances of diminutives, 43 were used to express a positive sentiment and 27 were used to express a negative sentiment. An example of diminutives form is shown in Table 3 and conditional form in Table 4 respectively.

In Estonian, the **superlative** is usually denoted by the suffix *-im* (*suurim* – biggest) or by the word *kõige* preceding the comparative form (*kõige kiirem* – fastest) (Erelt et al., 2020). The use of the superlative form can convey extreme opinions or positions, which can indicate stance. **Conditional forms** of verbs, ending with the suffix *-ks*, often imply that the situation being described is hypothetical or unrealistic. This form is used to express exaggerated or implausible scenarios, signaling a stance that suggests skepticism or disapproval. An example of superlative form is shown in Table 5. Similarly, nouns in the **translative case** also end with the suffix *-ks* which can suggest peculiarities or express attitudes (*Pean teda lolliks*, I think he's stupid) (Pai, 2001) on the basis of the context.

Finally, **indirect speech** in Estonian, recognizable by verbs ending in *-vat*, is used to convey statements heard from others rather than directly from the speaker. This form enables plausible deniability, introduces uncertainty, and allows the speaker to distance themselves from the information, often reflecting a stance of skepticism or disagreement with the reported statements (Teptiuk and Tuuling, 2024).

kõik, kõige (all), kunagi, eales (ever), iial (never), alati (always), igavesti (forever), tervenisti, täiesti, üleni (entirely), täitsa, täielikult (completely), üdini, läbinisti (thoroughly), läbini (through and through), absoluutne (absolute), absoluutselt (absolutely), totaalne (total), totaalselt (totally), ainult (only), ainus (sole), kogu (whole)

Table 6: List of words to detect black-and-white thinking. English translation in parenthesis.

3.3 Framing Analysis

As framing in news media can indicate towards political bias or stance (Kaukonen, 2022), we discuss next black-and-white thinking, bigram analysis and adjective-based framing that are related to framing.

Black-and-white thinking is a logical fallacy in which a complex situation is simplified into two extremes (Vleet, 2011). When authors use extreme or polarizing language, they often eliminate or do not consider alternate perspectives or possibilities. Black-and-white thinking is detected by word choice. Table 6 contains a list of hyperbolic words that could be considered polarizing. By detecting words from this list, it can be assessed whether a particular stance is being portrayed in a binary matter and lacks a middle ground. An example of a black-and-white thinking is shown in

Estonian Sentence	English Translation
Vahemere paadipõgenike ümber toimuv jätab üha enam mulje, et rändekriis hakkab **kõigile** närvidele käima, välja arvatud inimõiguslased ja teised sallivuslased, kes ei muutu **kunagi**.	The events around Mediterranean boat refugees increasingly give the impression that the migration crisis is getting on **everyone's** nerves, except for human rights activists and other tolerant individuals who **never** change.

Table 7: Black-and-white thinking in a sentence from the dataset that expresses a stance.

Table 7.

We additionally employ **bigram analysis** in order to identify any specific word pairs associated with a negative or positive stance. We also study **adjective-based framing** to understand the framing of certain concepts (Morstatter et al., 2018). For example, we observe that the concept of immigration can be referred to as illegal, uncontrollable or unlawful in sentences for the against stance and lawful or controlled are used to frame immigration in the supportive stance.

3.4 Sentiment Analysis

While sentiment analysis focuses on the polarity of the text, stance detection focuses on the viewpoint expressed towards a specific target (Mohammad et al., 2016b). We discuss next how we study and evaluate sentiment analysis on Estonian news media text in order to understand whether it can aid in political stance understanding. For **lexicon-based** sentiment analysis in Estonian, two notable corpora are available (Regita, 2023; Mohammad and Turney, 2013). While Regita (2023) developed a lexicon of 2454 sentiment-annotated words, provided by the Institute of the Estonian Language (EKI), Mohammad and Turney (2013) introduced EmoLex which comprises of 3693 words annotated for positive and negative sentiment. Subsequently, **Emotsioonidetektor** (Pajupuu et al., 2016) classifies a text negative, neutral, or positive directly being trained on Estonian Valence Corpus (Pajupuu et al., 2016). Emotsioonidetektor differs from lexicon-based approaches since it also considers context, such as cases where a positive or negative word was negated, obtaining the opposite sentiment. Inspired by the effectiveness of **BERT** (Bidirectional Encoder Representations from Transformers) in several natural language processing based tasks (Devlin et al., 2019), **EstBERT**, an Estonian-specific BERT model, was trained on the Estonian National Corpus, which contains approximately 1.34 billion words (Tanvir et al., 2021). This extensive dataset enabled EstBERT to outperform some multilingual BERT models in specific tasks (Tanvir et al., 2021). For sentiment analysis in this paper, both a fine-tuned EstBERT model and a multilingual XLM-RoBERTa model were used.

4 Results

In this Section, we discuss the results for features that provides a significant difference in observations results for the different stances. In total, we confirmed 15 features out of total of 32 features (as shown in Appendix A) to be useful for political stance detection in Estonian news media. All results were tested using a p-test and useful results were confirmed to have a p-value of under 0.01.

4.1 Lexical features

Our observations on comparison of the lexical features across different stances indicate that word count, dependency tree height, Flesch Readability Score, adjectives and quotes are useful whereas named entity counts did not show any difference. For example, our observations indicate sentences with against stance immigration had a higher word count, with a mean of 22.32 which is higher by $15 - 18\%$ compared to supportive and neutral stances, respectively (shown in Appendix B Table 14). Similarly, Dependency tree height of the sentences with against stance are higher by $4-10\%$ compared to supportive and neutral stances, respectively (shown in Appendix B Table 15) and Flesch Readability Score indicates that sentences with against stance are more complex by $10 - 15\%$ compared to supportive and neutral stances (Appendix B Table 16). However, we did not observe any difference across the number of named entities used in the sentences irrespective of the stance (Appendix B Table 17). Additionally, we observe that sentences with against stance has a higher usage of adjectives and quotes (Appendix B Tables 19 and 20), thereby highlighting anti-immigration texts use more emotionally charged language. We

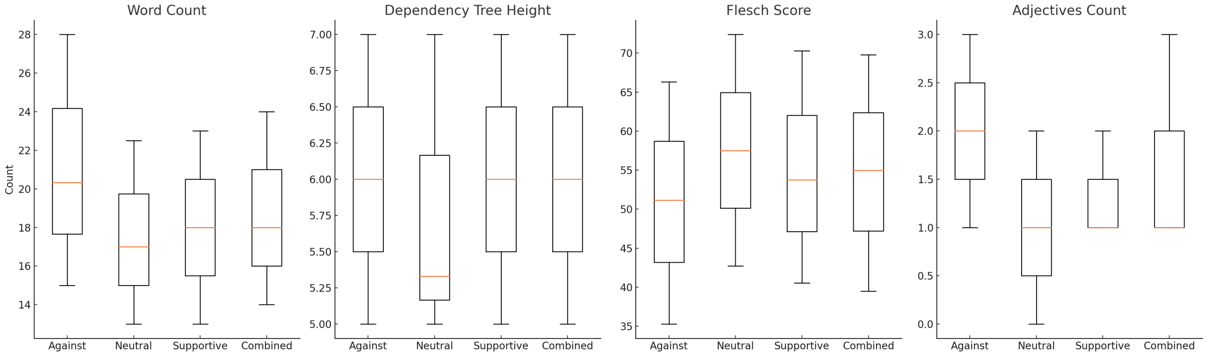

Figure 1: Box plots for Word Count, Dependency Tree Height, Flesch Score, and Adjectives Count across different stances (including a combined class, which includes all sentences).

Stance	Estonian Sentence	English Translation
Against	Rändel on suur demograafiline mõju, mis mõjutab **Rootsi** rahvuslikku ja kultuurilist identiteeti, samuti hävitav majanduslik mõju **Rootsi** heaoluriigile.	Migration has a significant demographic impact, affecting **Sweden's** national and cultural identity, as well as having a devastating economic impact on **Sweden's** welfare state.
Supportive	See, et inimesed **Lesbosel** elavad isetehtud telkides vihma ja külma käes, ei ole Euroopa Liidu vääriline [—].	The fact that people on **Lesbos** are living in makeshift tents in the rain and cold is not worthy of the European Union [—].

Table 8: Examples of countries (and regions) in sentences for different political stances

show a summary of these results in Figure 1 and the detailed results are provided in Appendix B corresponding to each of these features.

Additionally, we explore the most frequent named entities across different stances. Our observations indicate that the against sentences mentioned Sweden and Germany, which are popular immigration destinations. In contrast, supportive sentences focused more on the immigration transit destinations such as Greece and Turkey (and their associated regions). This can be due to against sentences highlighting troubles in immigration destinations, while supportive sentences focus on the troubles immigrants go through during transit. We highlight few examples in Table 8 and the most the most frequent named entities across different stances is shown in Table 9.

4.2 Estonian-specific features

Diminutives: Diminutives were infrequent in the dataset. Out of 3261 sentences, only 5 contained a word in the diminutive form. However, three of these sentences were annotated as against, and the remaining two were neutral. Although the diminutive is typically used to sound more gentle and positive, no sentence with supportive stance towards immigration were found with this feature. Two anti-immigration sentences are shown in Table 10 which contained the word *lumehelbeke* (snowflake), a derogatory term used to mock sensitive and delicate young adults who easily take offense and cannot tolerate conflict or criticism. Both sentences also included quoted words, insinuating doubt and judgement. These findings suggest that detecting and analyzing diminutives can be useful in political stance detection. However, as this feature is uncommon in this dataset, we could not make any significant conclusion of its importance as a feature.

Superlatives: Adjectives in the superlative form were uncommon among the sentences, as only 87 sentences contained them. Although the superlative form was slightly more common in the anti-immigration stance, we could not make any significant conclusion due to the lack of data. The most common superlative adjective was *suurim (biggest)*, with 24 occurrences, followed by *parim (best)* and *kõige olulisem (most important)*. However, no specific superlative adjective was typical for any stance.

	Against		Neutral		Supportive	
3.	Rootsi	51	Kreeka	75	Kreeka	22
4.	Saksamaa	51	Euroopa Liit	67	Türgi	22
5.	EKRE	47	Türgi	62	Saksamaa	20
6.	Euroopa Liit	44	Saksamaa	53	Soome	20
7.	Ungari	42	Rootsi	45	Euroopa Liit	19
8.	Itaalia	31	Itaalia	45	Süüria	16
9.	Prantsusmaa	29	Süüria	41	Prantsusmaa	12
10.	Vahemeri	28	Valgevene	34	Rootsi	12
11.	Kreeka	28	Aafrika	34	Vahemeri	12
12.	Helme	25	Ungari	34	Aafrika	10

Table 9: Top 10 most common named entities per stance, skipping the top 2 for each stance, which were *Eesti (Estonia)* and *Euroopa (Europe)*.

Stance	Estonian Sentence	English Translation
Against	Kui Ameerikas tuli võimule Trump, lubasid paljud Hollywoodi näitlejad samuti emigreeruda ja **lumehelbekesed** akendest välja viskuda, aga jäid siiski kohapeale ussitama – BLM-i suitsulõhnalised meeleavaldused lubasid ennast vabalt maha maandada, Portlandis loodi koguni oma anarhistlik “autonoomia”. [—]	When Trump came to power in America, many Hollywood actors promised to emigrate and **snowflakes** [promised] to throw themselves out of windows, but they still stuck around to nag - the BLM smoke-smelling protests allowed them to calm down, and in Portland, an anarchist ‘autonomy’ was created [—].
Against	Hiljuti lõi “progressiivses maailmas” laineid Rootsi **lumehelbeke**, kes olevat justkui väljasaadetud afgaani elu päästnud – tegu oli paraku Rootsis juba tuntud kriminaaliga.	Recently, a Swedish **snowflake** made waves in the “progressive world” for supposedly saving the life of a deported Afghan – who was unfortunately already a known criminal in Sweden.

Table 10: Examples of diminutives in sentences for different political stances

Conditionals and Translatives: The conditional form was present in 380 sentences, as seen in Appendix B Table 21 where 14% of both supportive and against stance based sentences contain a conditional form and only 8% of the neutral stance based sentences use conditionals. Similarly, we observe that while 28% of the against stance based sentences and and 26% of supportive stance based sentences use translatives while only 19% in the neutral sentences. The translative case was present in 763 sentences, as seen in Table 11. Therefore, it can be concluded that these features can aid in political stance detection and that there is a statistically significant association between stance and frequency of both conditional form ($\chi^2 = 24.78, p < 0.01$) and translative case ($\chi^2 = 27.31, p < 0.01$). However, the content of the words in both conditional form and translative case do not reveal much insight about stance. The most common word across stances in the translative case is *näiteks (for example)*. Similarly, the two most common words in the translative case are *oleks (would be)* and *peaks (should be)*. These words are not indicative of stance solely on their own.

Indirect speech: Although Indirect speech was only present in 38 sentences, it was most prevalent in the against stance based sentences towards immigration (22 occurrences) and was more than in the neutral and supportive stances combined.

4.3 Framing Analysis

On analyzing the most frequently occurring bigrams across different stances for news articles on immigration, we observe that while few bigrams are generic and has been used for both the

Sentences with feature *translatives_count* not equaling 0.								
Stance	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Against	326	1.27	0.74	0.33	1	1	1.88	6
Neutral	310	1.23	0.60	0.33	1	1	1	4
Supportive	127	1.22	0.60	0.33	1	1	1	5

Table 11: Summary of statistics for feature *translatives_count*.

Model	**Against**			**Neutral**			**Supportive**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
EKI	0.47	0.39	0.43	0.57	0.47	0.51	0.27	0.51	0.35
EmoLex	0.41	0.40	0.41	0.53	0.30	0.38	0.20	0.49	0.28
Emotsioonidetektor	0.39	0.60	0.47	0.67	0.07	0.13	0.20	**0.53**	0.29
EstBERT	0.44	**0.89**	0.59	**0.70**	0.25	0.37	0.48	0.31	0.38
XLM-RoBERTa	**0.53**	0.80	**0.64**	0.68	**0.49**	**0.57**	**0.50**	0.34	**0.40**

Table 12: Evaluation metrics for each stance class. P – precision, R – recall, F1 – F1-score.

stances, such as *Euroopa Liit (European Union)*, *eesti keel (estonian language)*, there are several bigrams which highlight distinct framing for different stances. For example, while against stance uses *illegaalne immigrant (illegal immigrant)*, *neeger (nigger)*, *araablane (Arab)*, etc., to show their stance, supportive stance uses examples, such as, *aafrika päritolu (African origin)* and *(Eesti Pagulasabi (Estonian Refugee Aid)*. This showcases a distinct difference in tone of the news articles on the basis of stance. We show the top 10 most frequently occurring bigrams for both against and supportive stance in Appendix B Table 23. Adjective-noun pair based understanding of framing similarly reveals contrasting frames (shown in Appendix B Table 24). For example, against immigration sentences focus on illegality, threat and about the massive problem this can lead to whereas supportive stance emphasize on equality and humanity.

4.4 Sentiment analysis

We show the comparative results of sentiment analysis of of the lexicon based models, EstBERT and XLM-RoBERTa model in Table 12. Our observations indicate that the XLM-RoBERTa model outperforms by ensuring highest F1-scores across all stances. Lexicon-based models and the Emotsioonidetektor model underperformed, especially on positive sentiment (Appendix B Figures 2, 3 and 4). The fine-tuned *EstBERT128_Sentiment* model, trained on the Estonian Valence Corpus, achieved an accuracy of 0.74, while the XLM-RoBERTa model slightly outperformed it with an accuracy of 0.76 (Appendix B Figure 5).

5 Conclusion

Automated political stance detection in Estonian is highly challenging due to the lack of existing datasets and Estonian specific language processing tools. In this paper, we study political stance detection with respect to Immigration in detail. Our analysis comprises of 32 features segregated between lexical features, Estonian-specific features, framing-related features and sentiment-related features. These features were exhaustively analyzed to determine their suitability for political stance detection in Estonian news media. Our observations indicate that 15 features were shown to be helpful in political stance detection. Furthermore, to the best of our knowledge, this is the first work that explores novel political stance detection features specific to the Estonian language.

As a future direction, the rich morphology of Estonian could be studied by conducting a comprehensive frequency analysis of all cases and conjugation forms. This could reveal additional features and insights related to stance or sentiment. Additionally, there is a need to develop a more extensive multi-domain dataset focused on political stance detection in Estonian news, which would support the development of automated machine learning models in this language.

Acknowledgments

This work has been funded from the EU H2020 program under the SoBigData++ project (grant agreement No. 871042), ETAg (grant No. SLTAT21096), HAMISON project (PCI2022-135026-2), and PSG grant (PSG855).

References

2019. Glossary on migration. In *International Migration Law No. 34*.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Reili Argus. 2009. Acquisition of estonian: some typologically relevant features. *Language Typology and Universals*, 62(1-2):91–108.

Manmeet Kaur Baxi, Rajesh Sharma, and Vijay Mago. 2022. Studying topic engagement and synergy among candidates for 2020 us elections. *Social Network Analysis and Mining*, 12(1):136.

Erik Bucy, W. Gantz, and Z. Wang. 2007. Media technology and the 24 hour news cycle. *Communication technology and social change*, pages 143–164.

Roshni Chakraborty, Srishti Bhandari, Nilotpal Chakraborty, and Ritwika Das. 2020. Eve2sign: Creating signed networks of news events. In *Text2Story@ ECIR*, pages 79–87.

Roshni Chakraborty, Maitry Bhavsar, Sourav Dandapat, and Joydeep Chandra. 2017. A network based stratification approach for summarizing relevant comment tweets of news articles. In *Web Information Systems Engineering–WISE 2017: 18th International Conference, Puschino, Russia, October 7-11, 2017, Proceedings, Part I 18*, pages 33–48. Springer.

Roshni Chakraborty, Maitry Bhavsar, Sourav Dandapat, and Joydeep Chandra. 2019. Tweet summarization of news articles: An objective ordering-based perspective. *IEEE Transactions on Computational Social Systems*, PP:1–17.

Roshni Chakraborty, Maitry Bhavsar, Sourav Kumar Dandapat, and Joydeep Chandra. 2022. Detecting stance in tweets : A signed network based approach. *Preprint*, arXiv:2201.07472.

Roshni Chakraborty and Nilotpal Chakraborty. 2023. Twminer: Mining relevant tweets of news articles. In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW)*, pages 1–3.

Delfi Meedia. Delfi meedia. https://delfimeedia.ee/. Accessed: July 5, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Einar H. Dyvik. The most spoken languages worldwide. https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/. Accessed: July 5, 2024.

Eesti Meediaettevõtete Liit. Tasulised digitellimused 2023. https://meedialiit.ee/statistika/statistika-2023/. Accessed: July 5, 2024.

Martin Ehala. 2009. Linguistic strategies and markedness in estonian morphology. *Language Typology and Universals*, 62(1-2):29–48.

Mati Erelt. 2013. Nimisõnafraasi sõnajärjest. *Oma Keel*, 26:56–60. (Accessed on 15.05.2024).

Mati Erelt, Tiiu Erelt, and Kristiina Ross. 2020. *Eesti keele käsiraamat*. Eesti Keele Instituut; EKSA, Tallinn.

Salman Farsi, Asrarul Hoque Eusha, and Mohammad Shamsul Arefin. 2024. CUET_Binary_Hackers at ClimateActivism 2024: A comprehensive evaluation and superior performance of transformer-based models in hate speech event detection and stance classification for climate activism. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 145–155, St. Julians, Malta. Association for Computational Linguistics.

Andrew Fisher, Rajesh Sharma, and Vijay Mago. 2023. Analyzing the trends of responses to covid-19 related tweets from news stations: an analysis of three countries. In *International Workshop on Health Intelligence*, pages 273–288. Springer.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. *Preprint*, arXiv:2010.12309.

Reet Kasik. 2015. *Sõnamoodustus*. Tartu Ülikooli Kirjastus, Tartu.

Elisabeth Kaukonen. 2022. Sooliselt markeeritud sõnad eesti spordiuudistes. *Keel ja Kirjandus*, 65(6):526–545.

Joana Kosho. 2016. Media influence on public opinion attitudes toward the migration crisis. *International Journal of Scientific & Technology Research*, 5:86–91.

Sven Laur, Siim Orasmaa, Dage Särg, and Paul Tammo. 2020. EstNLTK 1.6: Remastered Estonian NLP pipeline. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7152–7160, Marseille, France. European Language Resources Association.

Mirjam Liivak. 2023. Ke(ne)-liitelised deminutiivid eesti suulises argisuhtluses. Master's thesis, University of Tartu, Institute of Estonian and General Linguistics. (Accessed on 15.05.2024).

Rasmus Maide. 2020. Eesti keele nimeolemite märgendaja analüüs ja parandamine. (Accessed on 15.05.2024).

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2023. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *Preprint*, arXiv:2305.13047.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R. Corman, and Huan Liu. 2018. Identifying framing bias in online news. *Trans. Soc. Comput.*, 1(2).

Kumari Neha, Vibhu Agrawal, Vishwesh Kumar, Tushar Mohan, Abhishek Chopra, Arun Balaji Buduru, Rajesh Sharma, and Ponnurangam Kumaraguru. 2022. A tale of two sides: Study of protesters and counter-protesters on# citizenshipamendmentact campaign on twitter. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 279–289.

Joakim Nivre. 2010. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152.

Kristina Pai. 2001. Translatiivne ja essiivne predikatiivadverbiaal eesti kirjakeeles. Master's thesis, University of Tartu, Institute of Philosophy.

Hille Pajupuu, Rene Altrov, and Jaan Pajupuu. 2016. Identifying polarity in different text types. *Folklore: Electronic Journal of Folklore*, 64:125–142.

Jinkyung Park, Rahul Ellezhuthil, Ramanathan Arunachalam, Lauren Feldman, and Vivek Singh. 2022. Toward fairness in misinformation detection algorithms. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media. Retrieved from https://doi. org/10.36190*.

Rait Piir. 2023. Finland's chatgpt equivalent begins to think in estonian as well. *ERR News*. Accessed: July 5, 2024.

Richard Päll. 2021. SisserÄndevastaste hoiakute mÕju poliitilisele usaldusele euroopa rÄndekriisi valguses: Ungari ja poola nÄitel.

Luukas Regita. 2023. Tartu Ülikooli õppeainete tagasiside meelsusanalüüs.

Marcel Schlechtweg and Holden Härtl. 2023. Quotation marks and the processing of irony in english: evidence from a reading time study. *Linguistics*, 61(2):355–390.

Timo Spinde. 2021. An interdisciplinary approach for the automated detection and visualization of media bias in news articles. *Preprint*, arXiv:2112.13352.

E Tandoc, RJ Thomas, and L Bishop. 2021. What is (fake) news? analyzing news values (and more) in fake stories. media and communication, 9 (1), 110-119.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A pretrained language-specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Denys Teptiuk and Eda-Riin Tuuling. 2024. Manner expressions in finnish and estonian: their use in quotative constructions and beyond. *Linguistics*, 62(3):577–616.

Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. Named entity recognition in Estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83, Sofia, Bulgaria. Association for Computational Linguistics.

Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Raivo Vetik. 2000. Eesti elanike hoiakud poliitilise integratsiooniga seoses. *Integratsioon Eesti ühiskonnas. Monitooring 2000.*

Jacob E. Van Vleet. 2011. *Informal Logical Fallacies: A Brief Guide*. University Press of America, Lanham.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2:43–53.

A Summary of features

Feature Name		Description
LEXICAL	*word_count*	The number of words in a sentence.
	dependency_tree_height	The height of a dependency tree as calculated by EstNLTK's Maltparser model.
	flesch_score	The Flesch Reading Ease Score as calculated by EstNLTK's *SentenceFleschScoreRetagger*.
	named_entities	A list of named entities extracted by EstNLTK's named entity tagger.
	named_entities_count	Number of named entities in a sentence.
	noun_phrases	A list of noun phrases extracted by EstNLTK's experimental noun phrase chunker.
	noun_phrases_count	Number of noun phrases in a sentence.
	adjectives	Lemmas of adjectives used in a sentence.
	adjectives_count	Number of adjectives used in a sentence.
	quotes_count	Number of quotes in a sentence.
	quoted_words	A list of words and short phrases that are between quotes in a sentence.
	quoted_words_count	Number of quoted words and short phrases.
ESTONIAN-SPECIFIC	*diminutives*	A list of words that are in the diminutive form, noted by the ending *-ke* or *-kene*.
	diminutives_count	Number of words in the diminutive form.
	superlatives	A list of adjectives in the superlative form.
	superlatives_count	The number of adjectives in the superlative form.
	conditionals	A list of verbs that are in the conditional form, noted by the suffix *-ks*.
	conditionals_count	Number of words in the conditional form.
	translatives	A list of nouns that are in the translative case, noted by the suffix *-ks*.
	translatives_count	Number of words in the translative case.
	indirects	A list of verbs that are indirect, noted by the suffix *-vat*.
	indirects_count	Number of indirect words.
FRAMING	*bw_count*	Number of words that insinuate black and white thinking.
	has_against_bigram	A categorical variable indicating whether an against or supportive stance bigram or adjective used for framing was present in the sentence or not.
	has_support_bigram	
	framing_against	
	framing_supportive	
SENTIMENT	*ekilex_sentiment*	A sentiment classification of either negative, neutral, or positive, as determined by the respective model.
	emolex_sentiment	
	eki_emotion	
	estbert_sentiment	
	xlmroberta_sentiment	

Table 13: Summary of features.

B Tables and figures of results

Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	22.32	10.19	3	15	20.33	28	94
Neutral	1597	18.31	8.04	3	13	17	22.5	60
Supportive	489	18.95	7.95	4	13	18	23	52
Combined	3261	19.85	9.06	3	14	18	24	94

Table 14: Summary of statistics for feature *word_count*.

Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	6.33	1.94	2	5	6	7	18
Neutral	1597	5.72	1.71	2	5	5.33	7	17
Supportive	489	6.06	1.73	2	5	6	7	13
Combined	3261	5.99	1.82	2	5	6	7	18

Table 15: Summary of statistics for feature *dependency_tree_height*.

Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	49.46	25.37	-91.73	35.29	51.13	66.30	123.93
Neutral	1597	56.76	22.74	-36.52	42.73	57.50	72.38	134.12
Supportive	489	54.56	23.11	-29.21	40.53	53.76	70.30	129.57
Combined	3621	53.80	24.00	-91.73	39.49	54.96	69.79	134.12

Table 16: Summary of statistics for feature *flesch_score*.

named_entities_count across all sentences								
Stance	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Against	1175	1.54	1.43	0	0	1	2	10
Neutral	1597	1.60	1.56	0	0	1	2	11
Supportive	489	1.55	1.60	0	0	1	2	11
Combined	3621	1.57	1.53	0	0	1	2	11
Sentences with feature *named_entities_count* not equaling 0.								
Against	897	2.01	1.33	0.33	1	2	3	10
Neutral	1185	2.16	1.45	0.25	1	2	3	11
Supportive	359	2.11	1.52	0.50	1	2	3	11
Combined	2441	2.10	1.42	0.25	1	2	3	11

Table 17: Summary of statistics for feature *named_entities_count*.

noun_phrases_count across all sentences								
Stance	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Against	1175	3.11	1.87	0	2	3	4	17
Neutral	1597	2.83	1.69	0	2	3	4	15
Supportive	489	2.82	1.68	0	2	3	4	11
Combined	3621	2.93	1.76	0	2	3	4	17
Sentences with feature *noun_phrases_count* not equaling 0.								
Against	1132	3.23	1.80	0.50	2	3	4	17
Neutral	1529	2.95	1.62	0.50	2	3	4	15
Supportive	456	3.02	1.56	1	2	3	4	11
Combined	3117	3.06	1.67	0.50	2	3	4	17

Table 18: Summary of statistics for feature *noun_phrases_count*.

adjectives_count across all sentences								
Stance	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Against	1175	2.18	1.81	0	1	2	3	11
Neutral	1597	1.50	1.47	0	0	1	2	13
Supportive	489	1.64	1.60	0	1	1	2	12
Combined	3621	1.77	1.66	0	1	1	3	13
Sentences with feature *adjectives_count* not equaling 0.								
Against	1132	3.23	1.80	0.50	2	3	4	17
Neutral	1529	2.95	1.62	0.50	2	3	4	15
Supportive	456	3.02	1.56	1	2	3	4	11
Combined	3117	3.06	1.67	0.50	2	3	4	17

Table 19: Summary of statistics for feature *adjectives_count*.

Stance	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Against	104	1.05	0.60	0.25	0.63	1	1	4
Neutral	53	1.06	0.55	0.25	1	1	1	3
Supportive	27	1.18	0.67	0.25	1	1	1	3
Combined	184	1.07	0.60	0.25	1	1	1	4

Table 20: Summary of statistics for feature *quoted_words_count*.

Sentences with feature *conditionals_count* not equaling 0.								
Stance	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Against	167	1.20	0.72	0.33	1	1	1	4
Neutral	140	1.15	0.55	0.50	1	1	1	5
Supportive	73	1.14	0.49	0.33	1	1	1	3
Combined	380	1.17	0.62	0.33	1	1	1	5

Table 21: Summary of statistics for feature *conditionals_count*.

Stance	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Against	233	0.99	0.41	0.25	1	1	1	3
Neutral	175	0.99	0.33	0.33	1	1	1	3
Supportive	62	1.02	0.29	0.33	1	1	1	2
Combined	470	0.99	0.37	0.25	1	1	1	3

Table 22: Summary of statistics for feature *bw_count*.

	Against		Supportive	
1.	(euroopa, liit) (european, union)	38	(euroopa, liit) (european, union)	29
2.	(mart, helme) (mart, helme)	18	(eesti, keel) (estonian, language)	11
3.	(eesti, keel) (estonian, language)	18	(euroopa, komisjon) (european, commission)	8
4.	**(konservatiivne, rahvaerakond)** **(conservative, peoples party)**	14	(miljon, euro) (million, euro)	7
5.	(illegaalne, immigrant) (illegal, immigrant)	13	(välismaalane, seadus) (foreigner, law)	6
6.	(kogu, euroopa) (whole, [of] europe)	12	**(süüria, põgenik)** **(syrian, refugee)**	5
7.	(martin, helme) (martin, helme)	11	**(eesti, pagulasabi)** **(estonian, refugee aid)**	5
8.	(tooma, kaasa) (bring, along)	11	**(aafrika, päritolu)** **(african, origin)**	5
9.	**(eesti, konservatiivne)** **(estonian, conservative)**	10	**(sisseränne, piirarv)** **(immigration, limit)**	5
10.	**(neeger, araablane)** **(negro, arab)**	10	**(globaalne, ränderaamistik)** **(global, migration framework)**	5

Table 23: Top 10 most common bigrams in the against and supportive stances with English translations. Bigrams of interest are bolded.

	Against		Supportive	
1.	**(illegaalsete, immigrantide)** **(illegal, immigrants)**	8	(ebaseadusliku, rände) (unlawful, migration)	3
2.	**(odava, tööjõu)** **(cheap, labour)**	8	**(rahvusvahelist, kaitset)** **(international, defense)**	3
3.	**(konservatiivne, rahvaerakond)** **(conservative, peoples party)**	7	**(rahvusvahelise, rändekava)** **(international, migration plan**)	2
4.	**(massilise, sisserände)** **(massive, immigration)**	4	**(soolise, võrdõiguslikkuse)** **(gender, equality)**	2
5.	(uute, uudiste) (new, news)	4	**(avatud, algus)** **(open, beginning**)	2
6.	**(illegaalse, immigratsiooni)** **(illegal, immigration)**	4	**(salliva, õpikeskkonna)** **(tolerant, learning environment)**	2
7.	(uus, valitsus) (new, government)	3	(kogu, maailmas) ([in the] entire, world)	2
8.	(uued, uudised) (new, news)	3	**(suure, panuse)** **(big, contribution)**	2
9.	**(illegaalseid, immigrante)** **(illegal, immigrants)**	3	(globaalses, ränderaamistikus) (global, migration framework)	2
10.	**(suur, probleem)** **(big, problem)**	3	(rassilise, diskrimineerimise) (racial, discrimination)	2

Table 24: Top 10 most common adjective-noun pairs in the against and supportive stances with English translations. Pairs of interest are bolded.

Against	Terms in both	Supportive
agressiivne, allaheitlik, avantüristlik, efektiivne, elama, isiklik, islamiusuline, jahtiv, järgmine, jätkuv, kahjulik, kogu, konservatiivne, kriminaalne, kuritahtlik, käiv, kõrge, lõtv, ohtlik, paarituhandeline, potentsiaalne, range, rekordkõrge, riiklik, salakaval, sarnane, sealne, senine, seotud, suunduv, suvaline, tark, teisene, toimuv, tugevnev, tuntud, tülikas, valimatu, ähvardav, ühine, üksik, üleeuroopaline	lähtuv, piiramatu, kasvav	esitatud, hiiglaslik, inimlik, laiahaardeline, lubatud, noor, oluline, seaduslik, tõstatatud, vaba, väärikas, üleilmne
aggressive, submissive, adventurous, efficient, living, personal, muslim, hunting, next, ongoing, harmful, entire, conservative, criminal, malicious, ongoing, high, relaxed, dangerous, a few thousand, potential, strict, record high, national, cunning, similar, local, previous, related, heading, arbitrary, smart, secondary, occurring, strengthening, known, troublesome, indiscriminate, threatening, common, single, pan-european	originating, unlimited, growing	submitted, gigantic, humane, extensive, permitted, young, important, legal, raised, free, dignified, global

Table 25: Lemmatized list of unique adjectives in Estonian used to frame immigration, that preceded the stems *immigra* and *rän*. Translation in English is added.

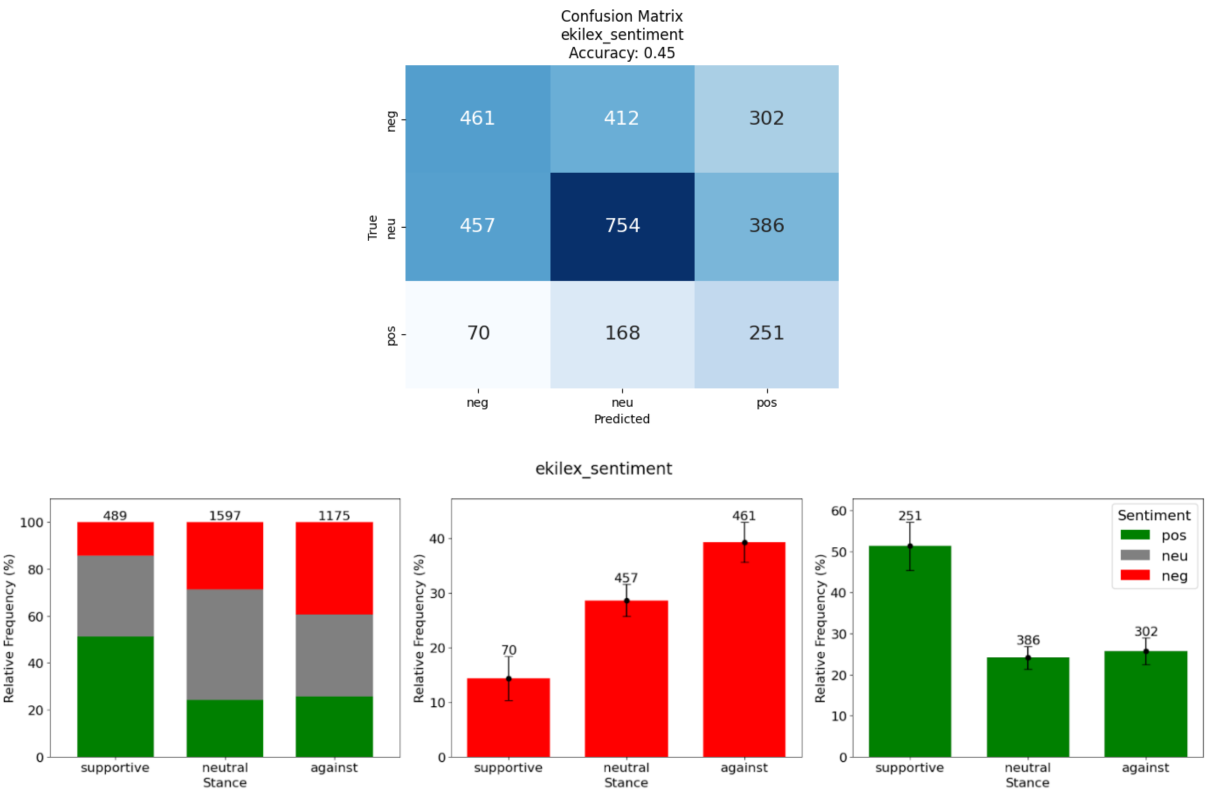

Figure 2: Confusion matrix and relative frequency graphs for sentiment predictions using the lexicon provided by the Institute of the Estonian Language (EKI).

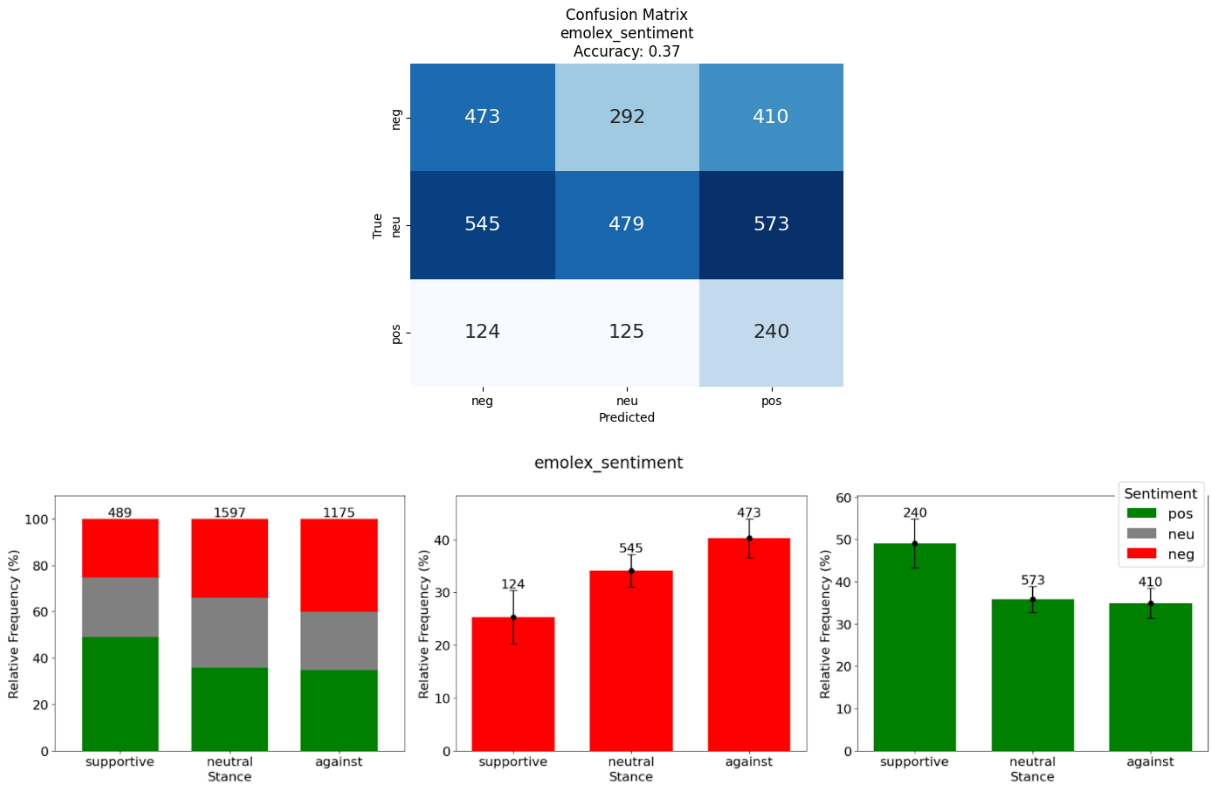

Figure 3: Confusion matrix and relative frequency graphs for sentiment predictions using the EmoLex lexicon.

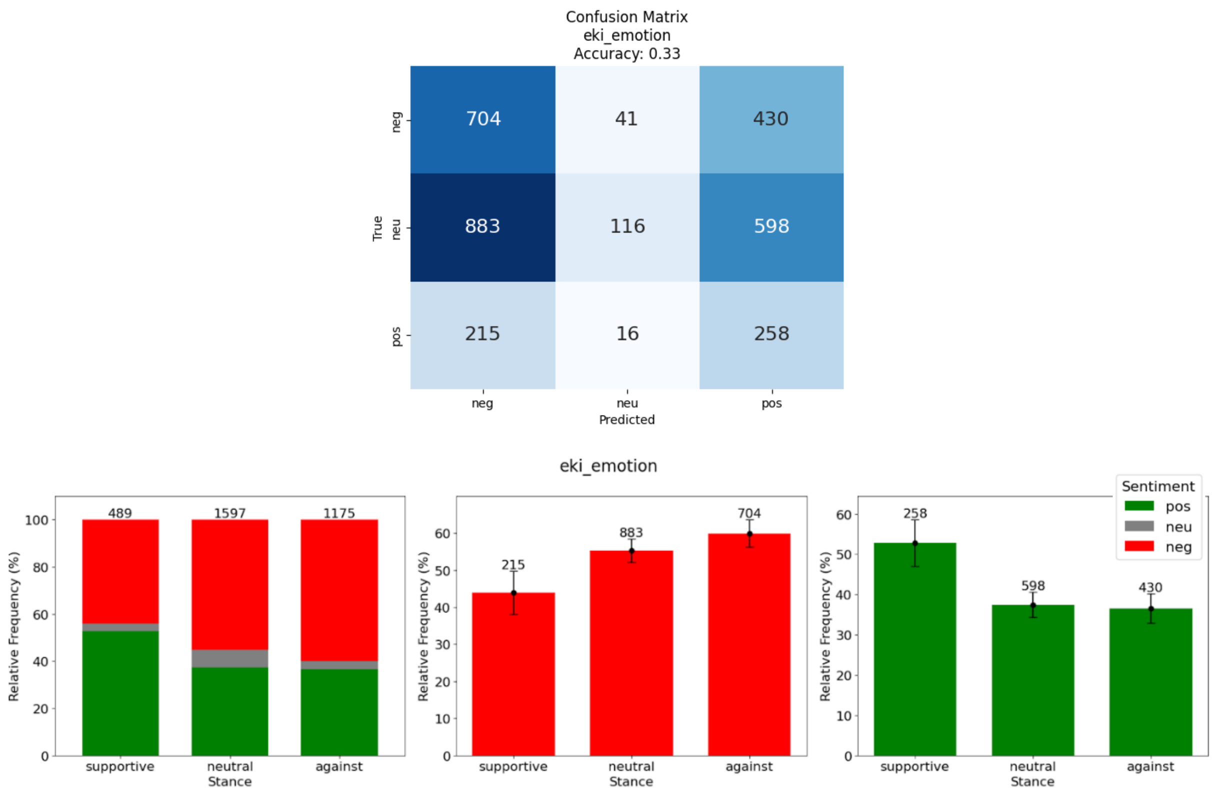

Figure 4: Confusion matrix and relative frequency graphs displaying the results of Emotsioonidetektor predictions.

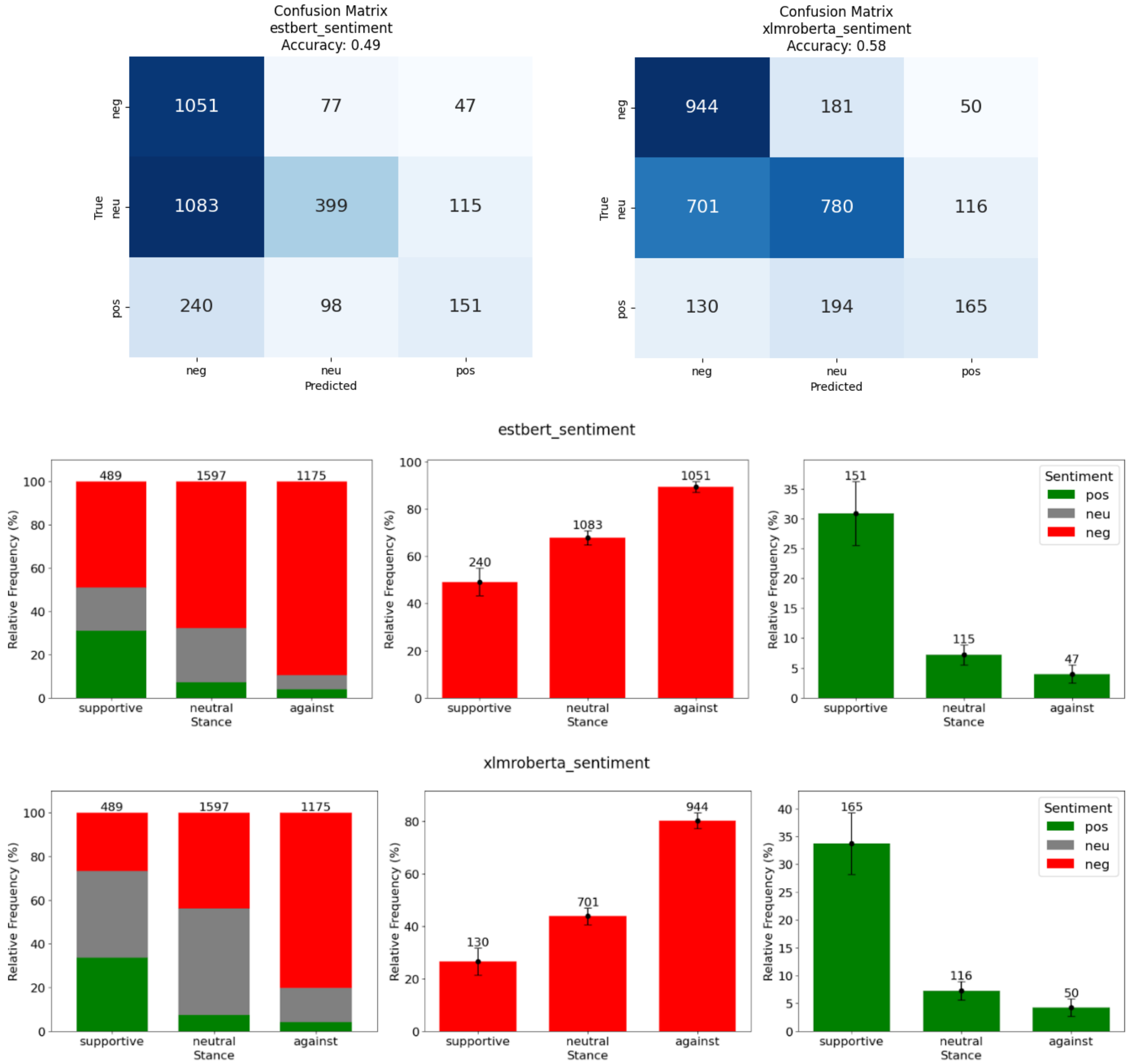

Figure 5: Confusion matrices and relative frequency graphs displaying the results of BERT sentiment model predictions.