ORIGINAL PAPER

"Approaches to sentiment analysis of Hungarian political news at the sentence level"

Orsolya Ring[1] · Martina Katalin Szabó[2,3] · Csenge Guba[1,4] ·
Bendegúz Váradi[5] · István Üveges[1]

Abstract

Automated sentiment analysis of textual data is one of the central and most challenging tasks in political communication studies. However, the toolkits available are primarily for English texts and require contextual adaptation to produce valid results—especially concerning morphologically rich languages such as Hungarian. This study introduces (1) a new sentiment and emotion annotation framework that uses inductive approaches to identify emotions in the corpus and aggregate these emotions into positive, negative, and mixed sentiment categories, (2) a manually annotated sentiment data set with 5700 political news sentences, (3) a new Hungarian sentiment dictionary for political text analysis created via word embeddings, whose performance was compared with other available sentiment dictionaries. (4) Because of the limitations of sentiment analysis using dictionaries we have also applied various machine learning algorithms to analyze our dataset, (5) Last but not least to move towards state-of-the-art approaches, we have fine-tuned the Hungarian BERT-base model for sentiment analysis. Meanwhile, we have also tested how different pre-processing steps could affect the performance of machine-learning algorithms in the case of Hungarian texts.

Keywords Hungarian political news · Sentiment analysis · Sentence-level analysis · Dictionary-based methods · Machine learning approaches · BERT model

✉ Orsolya Ring
  ring.orsolya@tk.hu

1 HUN-REN Centre for Social Sciences, Institute for Political Science, Tóth Kálmán Utca 4, 1097 Budapest, Hungary

2 HUN-REN Centre for Social Sciences, CSS-RECENS Research Group, Budapest, Hungary

3 Institute of Informatics, University of Szeged, Szeged, Hungary

4 Doctoral School in Linguistics, University of Szeged, Szeged, Hungary

5 Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary

🙂 Springer

1 Introduction

An increasing number of studies deal with the issue of sentiment analysis or opinion mining of political texts due to the vast amount of information available through the internet and the development of various natural language processing algorithms (Boukes et al., 2020; Haselmayer & Jenny, 2017; Mohammad, 2016; Mullen & Malouf, 2006; Pang & Lee, 2008; Pang et al., 2002; Rauh, 2018; Van Atteveldt et al., 2008; Young & Soroka, 2012). Political discourse cannot be considered a simple statement of facts: the tone of a given text is at least as important as the facts. Moreover, the tone can be the central component of individual decision-making and political judgment (Young & Soroka, 2012). The analysis of political news is particularly challenging, as it must refer to the direct content and the message perceived by the reader. Journalists may emphasize some facts more or quote other people, so news articles indirectly express their opinions. Automated sentiment scoring offers a good opportunity to measure the tone of political texts (Boukes et al., 2020). However, it can only work if it adequately reflects term usage in the political context (Rauh, 2018).

Sentiment analysis refers to the opinion we form of an object subjectively while interpreting texts. The identification framework allows us to interpret words, phrases, expressions, text fragments, or extended text sections into three collective sentiment types: positive, negative, and neutral (Hu & Liu, 2004). Due to the rising tendency of online sources of political news, sentiment analysis of political texts has become incredibly beneficial as empirical information in modern political science; therefore, the need for a new, computer-assisted method for large-scale analysis appeared. In addition, sentiment analysis in the social sciences suffers from a lack of agreed-upon conceptualization and operationalization (Lengauer et al., 2012; Van Atteveldt et al., 2008). Computational approaches to emotion analysis can potentially address the problems of scalability and repeatability inherent in manual coding (Boumans & Trilling, 2018; Van Atteveldt et al., 2008; Welbers et al., 2017). The most cost-effective of these is the dictionary-based approach, but its efficiency is significantly behind that of machine learning, in particular, the performance of state-of-the-art large-scale language models such as BERT (Devlin et al., 2018).

Implementing a dictionary-based approach with a domain-specific lexicon creates an inexpensive and unquestionably rapid way for researchers to gain valuable material for further studies. However, as languages vary, they carry their specific sentiment value, and the means are generally unavailable in most languages (Mohammad, 2016).

The toolkits for sentiment analysis are available primarily for English texts and require contextual adaptation to produce valid results—especially concerning morphologically rich languages such as Hungarian, hampering comparative communication research (Haselmayer & Jenny, 2017).

In Hungarian, grammatical information that is expressed by, e.g., prepositions in other languages (such as English) are encoded by inflections, which means that it uses various affixes, mainly suffixes, to change the meaning of words and their grammatical function so a word can have many different forms, given the possible

inversion and derivation morphemes. Which is why Hungarian is called a morphologically rich language.

For nouns, there are about 20 different cases, e.g., 'ház'- house (nominative), 'házban'—in the house (inessive), 'háznál'—at the house (adesssive), etc. The word order is free, i.e., the position of subject, subordinate, and object is not fixed within the sentence but refers to previously known information (new information is placed first within the sentence). There is also conjugation in verbs, which are matched with the object in number, person, and definiteness (e.g., 'mondhattam'—I could have said, where the verb itself is 'mond'—said). For all these reasons, in Hungarian, it is essential to reduce words to a common lexical form when processing texts, which in our case, is provided by lemmatization.

In the process of lemmatization, for example, the word 'autók' (cars), 'autóban' (in the car), and autóhoz (to the car) will be mapped back to the common dictionary form 'autó' (car), so only this dictionary form will appear in a list. We also used this later on when testing machine learning solutions.

As far as the Hungarian language is concerned, there are only two available sentiment corpora. The OpinHuBank corpus (Miháltz, 2013) is a sentiment corpus freely available for research and development. The manually annotated part of the corpus contains 10,000 sentences from 500 domestic online sources of different domains (news sites, blogs, forums) from 2009 to 2012. In the course of annotation, the sentiment value of each sentence was annotated on a three-point scale (positive, negative, or neutral). The other Hungarian sentiment corpus consists of opinion texts written about different types of products (M. K. Szabó, 2015). The corpus is annotated on entity and aspect levels. The corpus comprises 154 opinion texts approximately 17 thousand sentences and 251 thousand tokens. However, we are unaware of any Hungarian sentiment corpus of political texts.

Recognizing these research gaps, we have induced semi-automatically a Hungarian political domain-specific sentiment lexicon from an unlabeled corpus of 31,376 political news from small sets of seed words via a word-embedding method following the process detailed in (Hamilton et al., 2016). In addition, we have constructed a manually annotated sentiment dataset containing 5700 sentences from Hungarian news portals to validate our sentiment dictionary and test various machine-learning approaches to identify positive, negative, neutral, and mixed sentiment categories. Because of the complexity of news texts, the units of our analysis were sentences. Based on previous research on sentence-level analysis, we can draw more accurate information this way than text-level (Liu, 2010; Lutz et al., 2018; Yang et al., 2007).

During the annotation, we experienced that human coders had difficulty determining only sentences' positive and negative polarity. To solve this challenge, we have identified 12 different emotions in our corpus via the inductive approach. Then our annotators labelled the sentences into such emotion categories. In our experience, this system made the annotation for human coders much more accessible and the intercoder agreement higher (the average Cohen's Kappa from 0,4127 to 0,5828). Later we aggregated these emotions into positive, negative, and mixed sentiment categories.

The main contributions of our paper are summarized as follows:

1. We have designed a new sentiment and emotion annotation framework that uses inductive approaches to identify emotions in the corpus and aggregate these emotions into sentiment categories.
2. We have presented a manually annotated validation set with 5700 political news sentences
3. We have introduced a new Hungarian sentiment dictionary for political text analysis developed via word embeddings, whose performance was compared with the Hungarian general-purpose sentiment lexicon, with NRC Word-Emotion Association Lexicon and the Bing-Liu sentiment lexicon (Hu & Liu, 2004), translated automatically with Google Translate API from English to Hungarian.
4. We have fine-tuned the Hungarian BERT-based model (huBERT) for sentiment analysis (HunMediBERT).
5. We have compared the performance of different machine learning algorithms to analyze our dataset.

2 Related works

Sentiment and emotion analysis is an important research topic. However, there are often the terms "sentiment analysis" and "emotion analysis" interchangeably. Nevertheless, while sentiments and emotions are related, these two concepts have different meanings. Sentiment analysis or opinion mining concerns the concept of opinions expressed in texts that can be positive, negative, or neutral. In contrast, emotion analysis studies emotions (e.g., anger, sadness, or joy) reflected in a text. Hence, we should discriminate them from each other.

In this study, we use sentiment analysis only to determine whether the text expresses a positive, a negative, mixed, or a neutral opinion. By the task of emotion analysis, we mean emotion detection and emotion classification, which means both the task of detecting if a text conveys any type of emotion or not and the task of classification of existing emotion in a text into a set of defined emotions.

2.1 Sentiment analysis based on dictionary-methods

Sentiment analysis based on dictionaries is much less costly than applying more complex machine learning methods. In addition, dictionaries can be sources of features in the machine-learning framework (Mohammad, 2016). Sentiment lexicons are compiled of so-called sentiment words or phrases, in which each word usually carries a positive or negative tone (Liu, 2010).

The sentiment analysis approach can be identified as a document-, sentence-, word-, or target-level classification. The first two applications focus on the sentiment of the whole document or sentence. In these cases, researchers assume that each unit possesses positive, negative, mixed, or neutral sentiments. With target-level classification, the exact relation to each sentiment is more accurately identifiable, as we can link sentiment to specific objects (Gao et al., 2019; Liu, 2010; Song et al., 2020).

Over the general sentiment analysis, emotion analysis is a more complex way of classifying opinions as we move over the general distribution by studying the specific emotions of the texts, for instance, happiness, anger, or fear. Because these sentiment words might not even indicate any real sentiment, or they could bear several meanings, let alone the problematic detection of the mode of expression—like sarcasm, cynicism, or mockery –, the analysis still holds its challenges (Liu, 2010).

In lexicon-based sentiment analysis, most cases start with a list of words and are complemented with synonym detection later on (Whitelaw et al., 2005). The usage of social media data covers a great amount of related work, such as tweets (Dhaoui et al., 2017; Drus & Khalid, 2019; Kolchyna et al., 2015; Ray & Chakrabarti, 2017; Tumasjan et al., 2010). Furthermore, O'Connor et al. (2010) analyze the correlation between public opinion polls and tweets' sentiment. Their research shows no connection between sentiment and election results, but the sentiment and presidential job approval. In the Russian language, generally, machine learning approaches present the best results except for political news, where the lexicon-based method takes over thanks to the variety of topics in political news texts (Chetviorkin & Loukachevitch, 2013).

Moreover, Koltsova et al. (2016) aim to create a lexicon and study the political issues related to sentiment in social media for the Russian language. Their lexicon's efficiency is higher in the case of negative and less extreme sentiments. In the case of the Arabic language, the lexicon-based method is the most successful way of sentiment analysis with 83% accuracy (Itani et al., 2012). As for the German language, validation of a sentiment dictionary used for political science applications has been carried out. The data present a higher probability of identifying positive than negative emotions (Rauh, 2018). Additionally, the field of study provides evidence in detecting the political orientation of articles through sentiment. Alwan et al. (2021) aim to classify articles with the help of a lexicon and the Rough Set theory into three categories: Reformists, Conservative, and Revolutionary, with an 85% accuracy. A mixed-language automatic and semi-automatic analysis by Dilai et al. focuses on the examination of the US (2016) and Ukrainian (2014) presidential speeches by Donald Trump and Petro Poroshenko in two divisions: emotionally charged or neutral and positive or negative (Dilai et al., 2018). Both methods show that the president's speeches are subjective and charged with positive emotions, offering a further foundation for sentiment research.

Recently, many researchers have shown interest in applying word-embedding methods for sentiment lexicon-generating purposes or enlarging the size of an existing sentiment dictionary (Alshari et al., 2018; Huang et al., 2014). Researchers combine domain-specific word embeddings with a label propagation framework to induce domain-specific sentiment lexicons. They construct lexical graphs by applying a word-embedding method based on the vector representations of the words of the corpus. Then, similarly to using wordnets, they perform some form of label propagation over these graphs to induce domain-specific lexicons from seed words (Hamilton et al., 2016).

2.2 Sentiment analysis based on supervised machine learning approaches

Supervised learning is a common technique for solving sentiment classification problems (Khairnar & Kinikar, 2013; Pang et al., 2002; Ye et al., 2009). Before the emergence of Support Vector Machines, conventionally used algorithms were Naïve Bayes, k-NN, and C4.5 decision trees (Joachims, 1998). Naïve Bayes is a fairly simple group of probabilistic algorithms that is used when the size of the training set is small. There are two Naïve Bayes variants. Multinomial Naïve Bayes method follows a multinomial, Bernoulli Naïve Bayes a multivariate data distribution (Rish, 2001; G. Singh et al., 2019). Logistic regression is an exponential or log-linear classifier. It works by extracting weighted features from the input data, taking logs, and combining them linearly (Genkin et al., 2007). Support vector machines (SVMs) are highly effective at traditional text categorization, generally outperforming Naïve Bayes (Joachims, 1998). SVMs have advantageous attributes when it comes to text classification: the ability to work well with a high number of features without overfitting, the ability to work with sparse matrices, the kernel trick, and the way it can be used on different domains without much specific adaptation (Joachims, 1998). In this work, too, we applied various supervised techniques to analyze our datasets, such as Bernoulli Naïve Bayes, Support Vector Machine, and Logistic Regression.

2.3 Sentiment analysis based on transformer-based models

Language Models in Natural Language Processing (NLP) are designed to determine the probability of word or word sequences by analyzing textual data and learning syntactic and semantic rules. These models are then applied to solve linguistic-based problems like part of speech (POS) tagging, as well as generate new sentences accurately. The acquired 'knowledge' from large datasets can also be used for downstream tasks such as sequence labeling or named entity recognition (Singh & Mahmood, 2021).

Since human languages follow a sequential structure, the inception of language modeling was marked by the use of Recurrent Neural Network (RNN) architectures (Elman, 1990). RNNs were the first neural networks in which the states of individual neurons within a layer could interact (Yin et al., 2017). However, RNNs faced challenges with vanishing or exploding gradients when processing longer sequences. To overcome this, an improved architecture known as Long Short-Term Memory (LSTM) was developed (Hochreiter & Schmidhuber, 1997). Although LSTM stored the entire history of the processed sequence in a single state vector, it wasn't perfectly efficient in handling longer contexts. With the surge in available computing power, deep-learning neural networks have become more prominent (Akleman, 2020). The concept of 'attention' led to the breakthrough transformer architecture, first introduced in 2017 (Vaswani et al., 2017). The original transformer architecture, based on an encoder-decoder design, iteratively processes sequential input (e.g., natural language text) and creates encodings that capture relevant information within the input. On the other hand, the decoder layers store contextual information from all the encodings to generate an output sequence. Models like GPT-1 (Radford

et al., 2018) and BERT (Devlin et al., 2019) achieved significant success in various NLP tasks such as language modeling, sentiment analysis, and question answering in 2018.

These advancements led to the emergence of transfer learning, where knowledge accumulated from learning one task is applied to solve others (Singh & Mahmood, 2021). Recent language models like XLNET (Z. Yang et al., 2019) and RoBERTa (Liu et al., 2019) can be considered as attempts in this direction. These state-of-the-art language models have played a pioneering role in sentiment analysis tasks in recent years (Acheampong et al., 2021).

In the latest period, pre-trained language models have become the state-of-the-art solutions for most NLP tasks. Models like ELMo (Peters et al., 2018), GPT (Radford et al., 2018), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019), with hundreds of millions of tunable parameters, have significantly improved various challenging NLP tasks.

2.4 The specificity of the political news domain for sentiment analysis

Research examining political news sentiment primarily analyses whether it conveys positive or negative attitudes toward the topic under discussion (Haider-Markel et al., 2006; Pang & Lee, 2008). Research examining political news sentiment primarily analyses whether it conveys positive or negative attitudes toward the topic under discussion (Haider-Markel et al., 2006).

Some studies analyze emotions (Kepplinger, 2002; Uribe & Gunter, 2007). Khoo et al. applied Martin and White's (2005) framework for appraisal analysis on a sample of 30 political news articles and analyzed them for various aspects of sentiment. (Khoo et al., 2012).

One of the most significant research on emotion analysis in news texts is associated with Balahur and Steinberger, who attempted to separate good and bad news content and analyzed opinions explicitly expressed in news texts (Balahur & Steinberger, 2009). Tony Mullen and Robert Malouf conducted statistical analyses of political debate postings to analyze informal political communications (Mullen & Malouf, 2006). Sentiment analysis has been used to analyze the media coverage of different politicians and its impact on their electoral performance and public support for their policies (De Vreese & Semetko, 2002; Farnsworth & Lichter, 2005). Van Atteveldt and colleagues used a machine learning model to perform a network analysis of positive or negative relationships between actors and topics in political texts on manually annotated news reports of the 2006 Dutch elections (Van Atteveldt et al., 2008). Haselmayer and Jenny used dictionary and crowdcoding methods to analyze political communication (Haselmayer & Jenny, 2017). Boukes and colleagues, in their articles, compared a range of off-the-shelf sentiment analysis tools to manually coded economic news and examined the agreement between these dictionary approaches themselves (Boukes et al., 2020).

Regarding the analysis of political news texts in under-resourced and morphologically rich languages, Kaya et al. have carried out a sentiment analysis of Turkish news texts (Kaya et al., 2012). Sağlam et al. developed a Turkish sentiment

lexicon for sentiment analysis using online news media (Sağlam et al., 2016). Bakker et al. presented a two-step classification system to detect sentiment in the political news domain for an under-resourced language such as Norwegian (Bakken et al., 2016). Biba and Mane presented the first approach for Sentiment Analysis in Albanian (Biba & Mane, 2014). Bobichev and colleagues developed a corpus of Ukrainian and Russian news and conducted a sentiment analysis of Ukrainian and Russian news (Bobicev & Sokolova, 2017). Suryono and Indra carried out a sentiment analysis of Indonesian online news (Suryono & Indra, 2020).

Research on the expression of sentiment and emotion in political communication has been increasingly emphasized in recent years, also in Hungarian social science research (Bene & Szabó, 2021; Szabó, 2020; Szabó & Szilágyi, 2022). These studies primarily analyze social media (Bene, 2017; Miháltz, 2013) or political discourses (Sarlós, 2015), but there are only several examples of NLP tools being used to analyze the emotional content of political and mainly parliamentary speeches (Citation).

In our case, understanding the political news domain for opinion mining brings us to a challenging obstacle. The question of objectivity and pluralism arises through every bit of journalism, especially with news of political interest. Political articles often project hidden messages, attempting to change and assimilate viewers' attitudes through suggestions, while the audience receives the perception of journalists, which is a specific point of view.

On the other hand, along with the previously mentioned points and—the objective task of journalism—the transmission of information provides the perfect domain for emotion analysis, as it centers the attention on hot topics accordingly, raising and creating emotional responses (Cho et al., 2003). Emotion analysis of news texts can help us understand how the media reacts to political events. Specifically, as articles on current events often project complex, mediated messages, the audience receives journalists' perception, a specific point of view. Even though the apparent influence of journalists can be found in all cases, political sources consistently overpower journalists, as they depend on the specific supply of information provided (Bhowmick et al., 2009). To better understand the defining role of media in providing critical information and interpretation of politics, an automated system suitable for teasing out meaning from subtle textual sources is a valuable tool (Boomgaarden & Schmitt-Beck, 2019). Emotion can be analyzed both from the writer's and the reader's perspective (Bhowmick et al., 2009). In the current study, our task is sentence-level emotion analysis from a reader's perspective. Thus, we have tried to identify the emotion evoked in the readers while reading different news sentences.

As it appears in the psychological literature—the concept of emotion raises several questions while there are no formal criteria for what is and what is not an emotion (Cabanac, 2002; Chapman & Nakamura, 1998; Griffiths, 2008). The empirical analysis of emotion has uncovered the complexity of concepts (Lakoff & Kövecses, 1987), the entanglement of meanings with the specifics of local culture (Wierzbicka, 1999), and the lack of exact equivalents of special emotional expressions in different languages (Russell, 2003). Emotions are considered aspects of complex, interactional systems of the organism, which means there are various relationships between them. As Robert Plutchik (1982) states, emotions are much more complex than most

people realize. These properties of emotions cause them to be challenging to specify and analyze with text mining methods.

At the same time, the question of how the complex system of emotions is structured is still only one of the aggravating circumstances in the task of emotion analysis. The other fundamental question is what carries emotion at the content level. Some previous works (Feng et al., 2013; Loukachevitch & Levchik, 2016) constructed sentiment and emotion lexicons with connotative sentiment value rather than explicit sentiments exclusively. For instance, *awards* and *promotions* have positive connotations, and *unemployment* and *terrorism* have negative ones (Feldman, 2013). As Loukachevitch and Levchik state, "Non-opinionated words with connotations usually convey information about negative or positive phenomena (facts) in social life" (Loukachevitch & Levchik, 2016). In detail, "positive phenomena are usually supported, protected", and "negative phenomena are struggled with, fought against" (Loukachevitch & Levchik, 2016). However, automatic analysis of these connotative semantic contents remains a big challenge.

Connotation is particularly important in analyzing newswire texts since news texts do not explicitly express a positive or a negative opinion of the author but contain factual information implying positive or negative sentiment (Van de Kauter et al., 2015).

Consequently, successful dictionary-based sentiment analysis of the news is impossible without creating a specific sentiment lexicon that allows grasping both explicit sentiment contents and implicit or connotative meanings.

3 Annotation framework to identify emotions in the corpus

Our research experiences have also confirmed the findings of psychological literature. As mentioned earlier, we have experienced that human coders had difficulty determining the positive and the negative polarity of sentences. Still, they have a much better intercoder agreement in emotion annotation. We assume that it can be due to the complexity of semantic content conveyed by the sentences. To solve this problem, we developed a new sentiment and emotion annotation framework that uses inductive approaches to identify 12 emotions in the corpus (See Table 1). Our category system is based on Plutchik's categories, but only two categories (*fear and joy*) were clearly present in the corpus. We had to divide *anger* into two subcategories (*guilt and anger*), *disgust* into two (*conflict, contempt*), and *sadness* into three (*misfortune, suffering, and sorrow*). Instead of *expectation*, an *improvement/success* category was identified. In addition, we inductively identified two more categories close to, but not equivalent to, Plutchik's category of *trust*. Namely, *trust* has a positive emotional content, whereas the *justice/investigation* and *rescue* we identified have partially negative emotional content. Among Plutchik's categories, we could not identify *surprise* in the corpus.

After annotation, the 12 emotions identified were aggregated into positive, negative, and mixed emotion categories.

Figure 1 shows eight negatives, two positives, and two mixed emotions identified in the corpus. The emotion categories of success and joy as positive, while fear, crime,
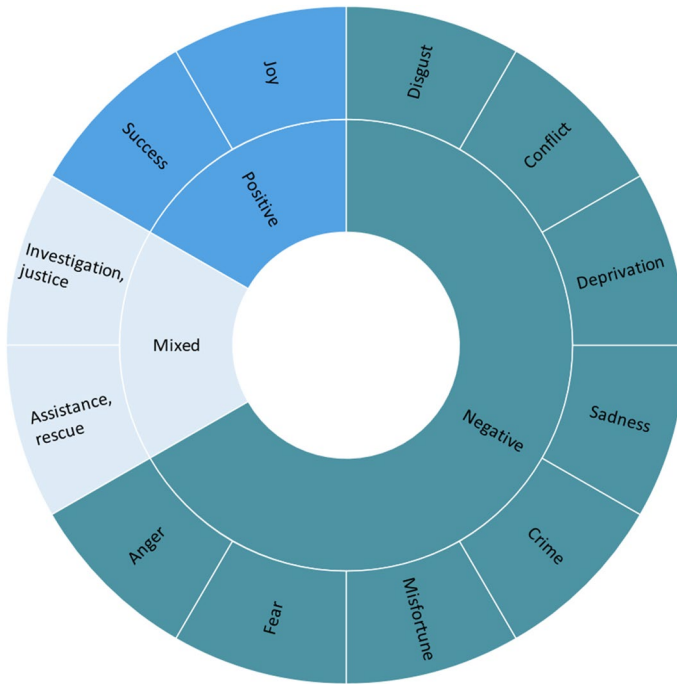
Fig. 1 Typology of the identified emotions

disgust, misfortune, sadness, deprivation, conflict, and anger as negative sentiments are consistent with the findings of the relevant literature (Demszky et al., 2020; Koljonen et al., 2022). As far as the mixed category is concerned, in the aggregation, we could not classify two emotions (assistance/rescue and justice/investigation) as positive or negative. Because in these sentences, the positive emotion always occurred along with a negative one, we handled these sentences as mixed sentiment categories. For example, the sentence "*A man collapsed on the street, passers-by saved his life.*" contains both positive and negative sentiments, and this causes the coders to disagree when it is only possible to choose between positive and negative categories. Nevertheless, if they could sign the sentence as "assistance", the intercoder agreement would be better. Same as "*The court imposed a prison sentence on the killer.*" This sentence couldn't be annotated as a pure positive sentence because of the reference to killing; however, it isn't a pure negative sentence as it contains a reference to the judgment of justice (See Tables 1 and 2).

The following figure demonstrates the system of emotions we applied during our work:

Table 1 The emotion classification system

Sentiment	Category	Example
Neutral	Neutral	Cambridge econometrics has examined several possible scenarios.
Negative	Fear, threat, intimidation, dread, anxiety	However, he has threatened the United States: if he does not stop his unjust war against Islam, then further assassinations will be committed against the country on its territory, similar to the one last September.
	Suffering, deprivation, misery, poverty, torment, failure, negative change	Moreover, in Ukraine, the difficult economic situation and the inefficient operation of leadership have created the precondition for Kyiv to give up its final defence line.
	Sorrow, despair, hopelessness, melancholy	On behalf of the municipality, the assembly and the residents of Székesfehérvár, mayor Tihamér Warvasovszky said goodbye to the deceased, and at the grave, on behalf of friends and relatives Mihály Pénzes, the president of MSZP organization of Székesfehérvár said goodbye to Zoltán Vancsik, a representative of the 1st constituency of Fejér county.
	Misfortune, catastrophe	On January 15th, 2011, three people died in the West Balkan, when panic escalated.
	Crime, terror, assassination, persecution, cruelty, organized crime, vandalism, intentional harm, violence	So far the autonomous region of Abkhazia north-western Georgia has been operating with kidnappings; its independence was declared in 1991.
	Anger, fury, hatred	According to the Swedish press, he hated the Foreign Minister for not forgiving him for supporting NATO bombings against Serbia during the 1999 Kosovo war.
	Conflict, confusion, conflict of interest, revenge, punishment	According to Trump, the north-Korean leader may not even want to meet him; anyway, if Kim Jong Un gives up on his nuclear weapons, then he will get extremely strong protection". Iran has been recently appearing in the world press with its opposition to Saudi Arabia, on which we recently wrote a summary article.
	Contempt, mockery	Hope kept his voters as well, who have reasoned that Fidesz is truly unacceptable here and there, truly runs idiots, truly campaigns with unrealistic communist promises – but if it is needed for the win, if it is needed for them to govern well afterwards, then let them do what they have to.
Positive	Improvement, relief, development, success, positive change	As a guest of the foreign representation meeting at the Ministry of Foreign Affairs, Cook said that Great Britain remains a committed supporter of our country's accession to the EU as soon as possible and that the Union has to adhere to the deadline of 2003 addressed at the Helsinki Summit.

Table 1 (continued)

Sentiment	Category	Example
Neutral	Neutral	Cambridge econometrics has examined several possible scenarios.
	Joy, enjoyment, merriment, serenity, love, acceptance,tolerance	Forty-five years ago, a huge crowd was waiting for John F., a young American president bearing the promise of change.
Mixed	Assistance, rescue,relief, healing, care,deliverance	The decision is related to this too, in which the EP would guarantee employment, education, and health rights for asylum seekers, who are not considered refugees, but their lives would be in danger if they were sent back to their countries.
	Justice, investigation	According to the organization, it is now the National Assembly's turn to provide law for the compensation of people, who have been sentenced to imprisonment or remained in custody based on regulations now classified as unconstitutional.

Table 2 Basic POS distribution of the validated dataset

POS-tag	Quantity	Percentage (%)
NOUN (common noun)	38719	24.15
PUNCT (punctuation)	21381	13.34
ADJ (adjective)	20815	12.98
DET (determiner)	18636	11.62
VERB (verb)	15165	9.46
PROPN (proper noun)	11353	7.08
ADV (adverb)	10001	6.24
CONJ (conjunction)	6251	3.90
PRON (pronoun)	5922	3.69
NUM (numeral)	4849	3.02
ADP (adposition)	3311	2.07
SCONJ (subordinator)	3221	2.01
PART (particle)	475	0.30
X (unknown)	104	0.06
INT (interjection)	86	0.05
SYM (symbol)	32	0.02
Σ	160321	

4 Dataset

After developing this annotation system, human coders labeled the sentences into such categories. The annotators were political science students and Hungarian native speakers with no prior experience in automated text analyses, so they were provided with a detailed annotation guideline. Our manually annotated corpus contains 5,700 double-blind coded sentences.

The validated label distribution can be seen in Fig. 2. Based on this, our dataset contained 1710 (*30%*) instances of positive sentences, 2,752 (*48.28%*) negative (the details of the aggregation to mere positive/negative classes can be seen below) while 647 (*11.35%*) sentences were classified as neutral and 591 (*10.36%*) mixed sentences.

The preliminary linguistic analysis of the manually annotated corpus was carried out by using the magyarlanc 3.0 toolkit (Zsibrita et al., 2013), which uses the UD morph tag set for POS-tagging (Nivre, 2015). Based on the results, the whole dataset contains 5700 sentences and 160,319 tokens. The basic part-of-speech (POS) statistics can be seen in Table 2.

Finally, the Inter-Annotator Agreement (IAA) was measured regarding these 4 aggregated sets of categories (Mixed, Neutral, Positive, and Negative) by the use of Cohen's Kappa score (Cohen, 1960). Cohen's Kappa coefficient (because of its simplicity and robustness) is a widely used method to calculate IAA between the annotators (Bhowmick et al., 2009; Bobicev & Sokolova, 2017; Krippendorff, 1980; Pyry et al., 2014), in case of variables that are describing a nominal scale (See Fig. 3).
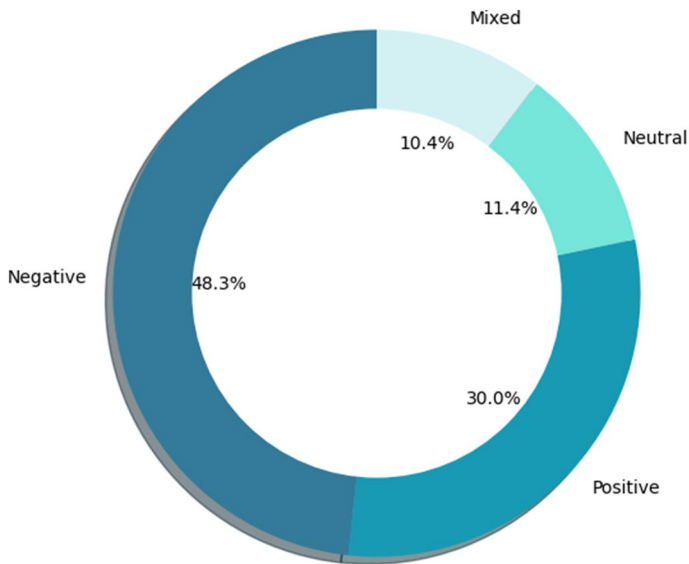
Fig. 2 Occurrences of the category labels in the validated dataset

The results showed moderate agreement in 3 out of 4 category sets and almost perfect in the case of positive sentences. These numbers indicate that such grouping of categories can be considered a feasible approach.

To have a benchmark corpus for all sentences where there was no agreement before, a qualified senior annotator decided on the emotion categories. We could use these labels as a gold standard in the later investigation phases.

5 Methods

As mentioned earlier, we offer different tools for sentiment analysis of political news in the Hungarian language. We established at the very beginning of our work, that the off-the-shelf dictionaries do not automatically lead to valid conclusions. More precisely, these general-purpose lexicons are not very effective in the domain analyzed. First of all, there are many words whose value in this domain is not the same as in another domain, which may cause biases in the analyses and limit the performance of sentiment analysis.

In this chapter, we first describe how the domain-specific dictionary was developed. We then discuss various machine learning techniques that we have used.

Figure 4 illustrates the overall dictionary creation process along with the comparisons made.

5.1 Dictionary creation

During the construction of the dictionary, both needed to provide as many relevant terms as possible with their corresponding sentiment value and to ensure that the resulting word lists contained as little noise as possible.

In the first phase, the wordlists were built using the SentProp algorithm, a label propagation framework (Hamilton et al., 2016). As in the original study, we started with the manual definition of a seed set, which in our case consisted of 120 words with positive and 120 words with negative polarity selected from the Hungarian Gigaword Corpus (Oravecz et al., 2014). The initial semantic vectors were created using a corpus of 31,376 texts containing Hungarian political texts. We have used the skip-gram model of Word2vec with a window size of 3, also setting the threshold to filter out tokens occurring less than five times to enrich our seed set (Mikolov et al., 2018). Negative sampling was also applied with a default value, while the created embeddings contained a standard 300-dimension (iterations were set to 10). We only included n-grams with a character length of more than 1, following the implications of Döbrössy et al. (2019), and excluded numbers from the n-grams (since they do not carry valuable sentiment-value).

In the next phase, the algorithm needs a lexical graph with edges representing the cosine distances between each node (word) and its K-Nearest Neighbors. In this step, we automatically selected the 100 semantically closest words of all seed words and assigned each word the same polarity as the given seed word had. However, this was a potential source of error in practice, as the words with similar contexts but opposite sentiment polarities are mapped into close word vectors in the embedding space (Fu et al., 2018). All words not assigned to the correct polarity have been removed from the lists during a manual check.

Based on this version of dictionaries, we performed a further expansion process in two steps. First, we used the latest version of the Hungarian Wordnet (Prószéky & Miháltz, 2008) to extend our positive and negative word lists. The Hungarian Wordnet database consists of approximately 42,000 synsets, organized in terms of classical semantic relations; synonymy, antonymy, hypernymy, and hyponymy. It is available for any non-commercial purpose as a GitHub repository. In this step, the synonyms of all the words already on the list were searched for and added to the list.

After this, we again manually reviewed all new lexical elements to ensure that they were valid extensions from the viewpoint of our research goal. This has crucial importance since inappropriate words were quite common, which would also have been a source of error in subsequent analyses and would have negatively affected the overall effectiveness of the created dictionary. The final version of our dictionary (hereafter: POLTEXTLAB Dictionary) contains 2,585 positive and 2,566 negative words.

5.2 Lemmatization

The mentioned dictionary size was formed so that only the lemmas of each word were added to each (positive and negative) list.

Lemmatization is often used in natural language processing (NLP) when dealing with morphologically rich languages like Hungarian. Its main difference from stemming (a technique often used in the case of the English language, for example) can be described by the fact that stemming primarily only removes the various suffixes from the end of the word (therefore, it does not necessarily result in a meaningful word form). At the same time, lemmatization always returns the dictionary form of the word as a result (Jurafsky & Martin, 2000).

Both are stemming and lemmatization used in constructing the vector space model to bring the different word forms of each word into a common canonical form, allowing for their unified representation. In practice, this means that the information related to the word form is lost, but the dimension of the vector space can be significantly reduced. In English, this decrease can reach 40–70%, while in Hungarian, according to some observations, it can be as high as 90% (Tikk, 2007). Whether the separate representation of individual word forms is unnecessary noise or valuable information can be highly domain- and application-dependent. In our case, the conjugated forms of individual words would not have helped the construction of the dictionary. On the other hand, lemmatization resulted in much more expressive lists, which is why we decided to use it.

5.3 Comparison with other dictionaries and the gold standard

The next step was to validate the dictionary's effectiveness and put its performance into context. For this purpose, we first compared the performance of the dictionary with the manually annotated Gold Standard.

The sentiment score of each sentence here was calculated via a simple algorithm: if a word was found from the negative domain-dependent list, then -1 was added to the score of the sentence. If a positive one was found, then $+1$ was added to it. Henceforward, the summed result of the above two was divided by the overall token count of the given sentence (as a normalization step). If the result is a positive number, the sentence is considered as a positive one, and similarly for negatives:

$$\sum_{n=1}^{N_i} w_{ip} - w_{in}$$

here, N is the number of tokens with sentiment value in a sentence I, w_{ip} is the number of positive and w_{in} is the number of negative tokens within the sentence. Finally, the calculated sentiment score (given by the dictionary-based approach) and the manual annotations were compared in terms of the F-score regarding each category.

We have already explained that general-purpose lexicons cannot be as accurate as a special sentiment dictionary created for the political domain. To support this

statement, we have compared their performance with the Hungarian general-purpose sentiment lexicon (with 2699 positive and 6811 negative terms), with NRC Word-Emotion Association Lexicon (with 1884 positive and 2584 negative terms) and with the Bing-Liu sentiment lexicon (with 2005 positive and 4783 negative terms), from which NRC provide versions of the lexicon in over 100 languages by translating the English terms using Google Translate, the Bing-Liu was translated automatically with Google Translate from English to Hungarian.

All three dictionaries were evaluated in two ways. First, we worked with a version of the corpus where sentences belonging to mixed categories were removed since these clearly cannot be classified correctly using just a positive–negative opposition.

Besides this, it was necessary to prepare a second version of the corpus due to the problematic nature of the neutral elements. These are difficult to manage since sentiment dictionaries contain only positive and negative lists, so the sentiment score calculated using them can only be positive or negative. In this relation, neutral sentences could best be associated with a score of 0 (if, for instance, they contain neither positive nor negative words). However, a score of 0 can also occur when a sentence contains an equal number of positive and negative sentiment scores, and such cases cannot automatically be considered as merely neutral sentences. To eliminate this confounding effect and to allow for the most rigorous possible evaluation of the dictionaries' performance, we have removed sentences in this second version of the corpus that were given a neutral label during manual annotation.

5.4 Sentiment analysis with different machine learning algorithms

By developing the sentiment dictionaries, we wanted to provide a solution that allowed researchers with little or no IT expertise to conduct basic sentiment analysis. However, to put the performance of the dictionaries into context, it was also worth testing some simpler machine learning algorithms so that the effectiveness of the dictionaries could be evaluated in their light.

To achieve this goal, we tried Naïve Bayes (NB), Support Vector Machines (SVM), and logistic regression (LR) classifiers with TF-IDF vectorization. Here we have also attempted to evaluate the performance of the dictionaries in an optimal setting, so we have removed the neutral sentences from the corpus. Accordingly, the machine classifiers were trained for binary classification (to separate positive and negative sentences). Separation of train and test data was performed in a standard way, using 70% of the corpus as train and 30% as test data (because of the relatively low number of annotated sentences) in all cases.

Before running the machine learning algorithms, standard pre-processing steps were applied:

- lowercasing,
- defining a stopword list of Hungarian, and removing these stopwords (like determiners, personal pronouns, etc.),
- removing punctuation and numeric numbers (based on simple regex patterns),
- and tokenization.

In TF-IDF vectorization, the size of the resulting dictionary and the informativeness of the words it contains can significantly affect the efficiency of machine learning algorithms. For morphologically rich languages, the size of the vectorizer's dictionary can be reduced significantly and without loss of information by lemmatizing the included words. Compared to the usual stemming in English, it is different in that it not only removes the conjunctions at the end of words when stemming conjugated words but also takes morphological information into account in the process. This ensures the recovered word is always a meaningful dictionary form (Zsibrita et al., 2013).

To demonstrate this difference in practice between simple tokenization and the more advanced lemmatization, we tested NB, SVM, and LR approaches on the data preprocessed in these different ways. In the first version, the sentences were merely tokenized and then vectorized with TF-IDF vectorizer. In the second version, instead of simply tokenizing the sentences, a lemmatization process was carried out on the dataset sentences before vectorization. Spacy carried out both lemmatization and tokenization using a language model developed by György Orosz.[1]

Although the primary aim of preparing the dictionaries was to create a simple and easy-to-use tool for sentiment analysis tasks, it seemed appropriate to compare their effectiveness with state-of-the-art methods. In this way, the results obtained can again be put into context more easily. To this end, we fine-tuned the huBERT (Nemeskey, 2020) model, which was the first Hungarian implementation of the BERT-based case model. The fine-tuning was done in a standard way like the way we have previously done for sentiment and emotion analysis of political texts (Üveges & Ring, 2023).

6 Results

The results obtained with POLTEXTLAB sentiment dictionaries are first compared with other (general purpose) sentiment dictionaries and then with the machine learning algorithms mentioned above.

6.1 Comparison of dictionaries

To get an accurate picture of the effectiveness of the dictionaries, we tested their predictive accuracy under two possible scenarios. As described in the previous chapter, each of the dictionaries (the general Hungarian-language sentiment dictionary, the machine-translated version of the Bing-Liu dictionary, the NRC Word-Emotion Association Lexicon, and the POLTEXTLAB dictionary) was evaluated on two versions of the corpus.

Table 3 presents the results obtained. The first column shows the results where mixed sentences have been removed from the corpus before evaluation ('–M'), and

---

[1] https://github.com/spacy-hu/spacy-hungarian-models.

Table 3 Comparison of results achieved with different dictionaries on different corpus-variants

		GPD		Bing-Liu		POLTEXTLAB		NRC	
		–M	–M, N	–M	–M, N	–M	–M, N	–M	–M, N
Positive	P	0.43	0.47	0.43	0.48	0.46	0.53	0.4	0.46
	R	0.69	0.69	0.52	0.52	0.7	0.7	0.72	0.72
	F1	0.53	0.56	0.47	0.5	0.55	0.6	0.51	0.56
Negative	P	0.86	0.91	0.79	0.84	0.81	0.88	0.69	0.77
	R	0.12	0.12	0.23	0.23	0.38	0.38	0.23	0.23
	F1	0.21	0.21	0.36	0.36	0.52	0.53	0.34	0.35
AVG	P	0.65	0.69	0.61	0.66	0.64	0.71	0.55	0.62
	R	0.41	0.41	0.38	0.38	0.54	0.54	0.48	0.48
	F1	0.37	0.39	0.42	0.43	0.54	0.57	0.43	0.46

GPD: general purpose Hungarian sentiment dictionary, Bing-Liu: translated dictionary, POLTEXTLAB: POLTEXTLAB sentiment dictionary, NRC: Word-Emotion Association Lexicon, AVG: average of the given metric measured on positive and negative sentences
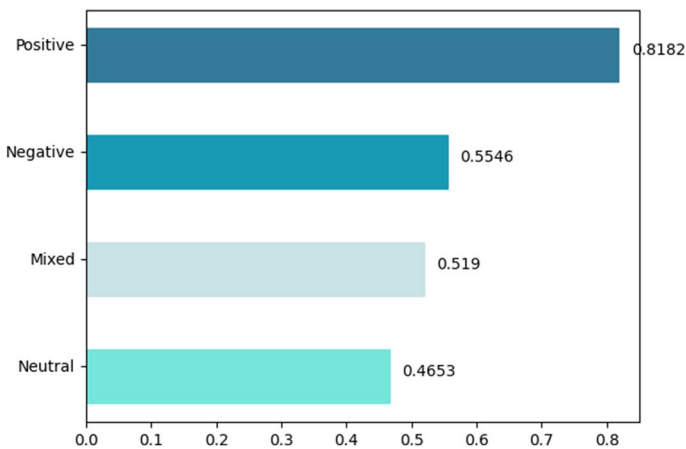
Fig. 3 Cohen's Kappa scores regarding the four major groups of categories

the second column shows the evaluation results after removing mixed and neutral sentences ('–M, N'). It should be noted that both scenarios can be considered somewhat idealized compared to real usage conditions. However, the limitations of sentiment analysis methods using dictionaries (i.e. that dictionaries containing words with positive and negative connotations can only predict positive and negative sentiment values by themselves) should be considered. We believe that an environment optimized under these conditions best demonstrates the strength of the developed domain-dependent solution compared to the translated (Bing-Liu), the NRC Word-Emotion Association Lexicon, and the general-purpose dictionaries.

The results show that the POLTEXTLAB dictionary (as expected) yielded the best average F-value of the four dictionaries for all three corpus variants tested. The
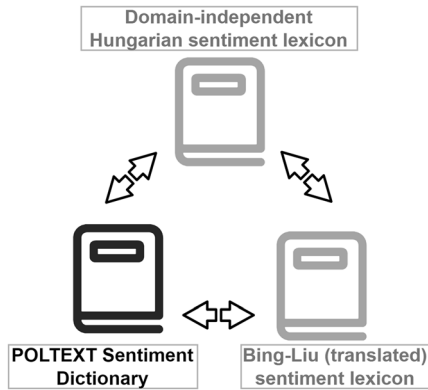
Phase III.

Comparison with simpler Machine Learning algorithms

POLTEXT Sentiment Dictionary

1. Naive Bayes
2. Support Vector Machine
3. Logistic Regression
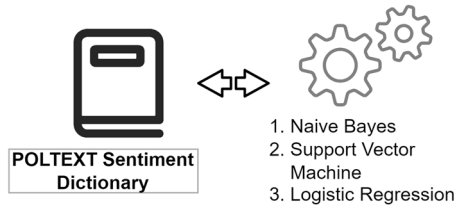
Phase II.

Comparison with other dictionaries to determine the effectiveness of POLTEXT Sentiment Dictionary

Domain-independent Hungarian sentiment lexicon

POLTEXT Sentiment Dictionary

Bing-Liu (translated) sentiment lexicon

Phase I.

Creating domain-specific sentiment dictionary for the political domain

Dictionary creation

POLTEXT Sentiment Dictionary

Manual validation

Adding synonyms from Hungarian Wordnet

Manual validation

SentProp algorithm with 120 positive, 120 negative word (seed set)
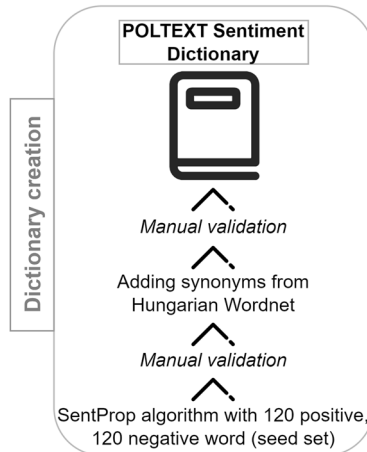
Fig. 4 Summary of the applied methods

advantage of the dictionary is most obvious in the purest context (–N, M case), but it is not much less so in the other corpus versions.

In general, POLTEXTLAB performed better and more balanced than all three other dictionaries for both positive and negative sentiments. Bing-Liu and the NRC lexicon performed similarly, which is not surprising, as both include automatic translation of English words with Google Translate. When averaging the positive

Table 4 Performance of different machine learning algorithms

		Positive			Negative			Average		
		P	R	F1	P	R	F1	P	R	F1
–M	NB	0.83	0.02	0.04	0.53	1	0.69	0.68	0.51	0.37
	SVM	0.59	0.48	0.53	0.63	0.84	0.72	0.61	0.66	0.63
	LR	0.62	0.38	0.47	0.6	0.89	0.72	0.61	0.64	0.60
	HunMediBERT3	0.74	0.84	0.79	0.89	0.89	0.89	0.82	0.87	0.84
–M, N	NB	0.9	0.04	0.07	0.63	1	0.77	0.77	0.52	0.42
	SVM	0.68	0.49	0.57	0.74	0.86	0.79	0.71	0.68	0.68
	LR	0.71	0.33	0.45	0.7	0.92	0.79	0.71	0.63	0.62
	HunMediBERT2	0.86	0.89	0.88	0.93	0.91	0.92	0.9	0.9	0.9

–N: corpus with neutral sentences removed, –N, M: corpus with neutral and mixed sentences removed, *NB* Naïve bayes, *SVM* Support vector machine, *LR* Logistic regression

and negative class prediction metrics, it can be seen that in terms of F1 score, the POLTEXTLAB dictionary performs about 0.17 and the machine-translated dictionary performs 0.12 better than GPD when mixed sentences are removed (–M). After removing mixed and neutral sentences (–M, N), this advantage increases to 0.18 and 0.14, respectively.

In the next step, we checked how the POLTEXTLAB sentiment dictionary performs against some classical machine learning algorithms. Note that none of these can be considered state-of-the-art solutions. In almost all areas of NLP, transformer-based solutions can achieve results orders of magnitude better than previous machine learning algorithms (even those presented here). However, using them requires considerable expertise and, in many cases, large amounts of hardware resources (mainly GPU capacity). For all these reasons, comparisons with the dictionary-based solutions presented here could only lead to trivial conclusions, such as that neural network-based AI algorithms are significantly better in performance than simple dictionary-based solutions. The aim of this paper, however, is to demonstrate the effectiveness of dictionaries that can be applied with minimal IT expertise and low resource requirements. A comparison with the machine learning algorithms presented here fits much better into this line of thought. As mentioned above, the comparison with the fine-tuned huBERT model (named HunMediBERT) is only included in the analysis for comparability. Note that the huBERT model was separately finetuned for 3-class classification (-M case) and then for binary classification for sentiments (–M, N case). Models were named HunMediBERT2 and HunMediBERT3 respectively.[2]

Table 4 illustrates the classification results achieved with the mentioned ML algorithms. To evaluate the algorithms in this case, binary classifiers were trained and tested on two corpus versions. One was obtained by removing neutral sentences,

[2] Available at: https://huggingface.co/poltextlab/HunMediBERT2 and https://huggingface.co/poltextlab/HunMediBERT3.
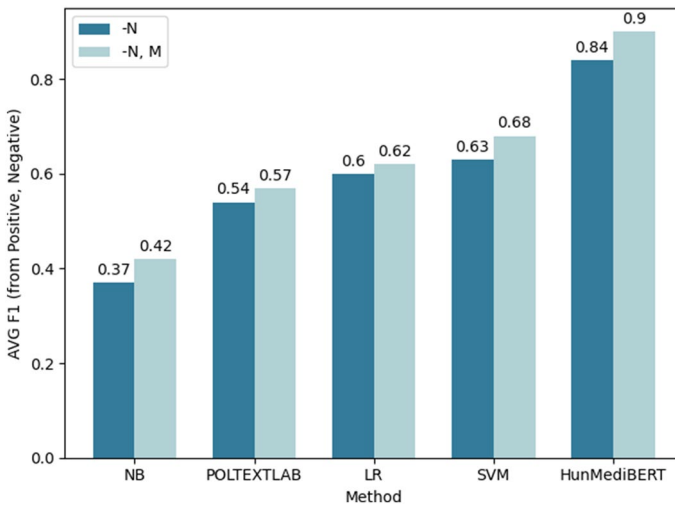
Fig. 5 Average F-scores with different ML-based approaches and with the POLTEXTLAB sentiment dictionary

as mentioned earlier, and the other by removing neutral and mixed sentences. As expected, among the tested algorithms, NB performed the worst in all cases. This was presumably because the classifier simply labeled almost all sentences as negative. SVM and LR performed relatively better, with a difference of only 0.03 for the F1-score of the corpus version without mixed sentences and 0.06 for the version without neutral and mixed sentences (both in favor of SVM). As expected, the BERT-based solution proved to be the most effective in all cases.

Overall, the recognition of negative sentences was more efficient for all algorithms, presumably due to the highly unbalanced nature of the corpus (the proportion of negative sentences exceeded that of positive ones by about 60%, with 2752 sentences against 1710). This is somewhat in contrast to the IAA results achieved during manual annotation, which show that the annotators achieved much higher agreement when labeling positive sentences.

This kind of asymmetry calls for caution during the practical application of the models, as it indicates that the models are somewhat overfitted for the negative sentiment class. However, the difference is not significant enough to predict a major reduction in the generalization ability of the model (at least for SVM and LR).

6.2 Dictionaries versus machine learning algorithms

The question of where the performance of the POLTEXTLAB dictionary is compared to the ML solutions can best be answered by averaging and ranking the obtained F1 scores of the different methods.

Figure 5 illustrates the performance of the POLTEXTLAB dictionary and the selected machine-learning algorithms. Each bar in the figure illustrates the F1-Score achievable with the given approach, again for the –M (without mixed) and –M, N

(without neutral and mixed sentences) corpus variants. Note that the F1-Scores reported here are the averages of the F1s calculated separately for the positive and negative sentiment classes for each method. The values on which the averaging is based can be traced in Tables 3 and 4 (see 'AVG' F1 row in Table 3, and 'Average' columns in case of Table 4).

From the results, it is clear that the NB lags behind the performance of the dictionary due to its overfit on negative samples, but somewhat more sophisticated machine learning solutions (LR, SVM) already perform better. In terms of the order, there is no difference between the -M and -M, N corpus versions. The results show that removing neutral sentences in addition to mixed sentences improved the results in all cases (in the case of the POLTEXTLAB dictionary, improving the F1 value by 0.03). This shows that the vocabulary of sentences belonging to the mixed category showed a significant overlap between the positive and negative words included in the POLTEXTLAB dictionary. Here, again, the clear superiority of the BERT-based solution can be observed.

7 Future work

We hope that the basic resources created and presented will provide baseline results to which the performance of increasingly advanced machine learning algorithms can be compared in the future.

In this paper, we have started from the simplest dictionary approaches, through different machine learning algorithms, to presenting the possibilities provided by state-of-the-art context-dependent embeddings, as an important goal of this paper was to help researchers with less comprehensive NLP expertise to implement sentiment analysis procedures with some easy-to-implement solutions.

The role of context-dependent embeddings that can be extracted from neural network-based models (e.g. BERT) is twofold; on the one hand, it can serve as the cornerstone of the entire dictionary construction pipeline described in the study when applying the SentProp algorithm, and machine learning models can also be built with their help, for example during the fine-tuning of transformer-based models. In the future, we plan to use the latter to build a model that will be able to perform sentiment analysis tasks on political texts using cutting-edge methods more efficiently, also identifying the aspect of the sentiment (Aspect Based Sentiment Analysis), than the cutting-edge methods presented here, although requiring more NLP expertise.

8 Conclusion

In this paper, we presented different approaches for the automated sentiment analysis of Hungarian political news sentences. We presented a novel sentiment annotation framework, a political sentiment dictionary, a benchmark corpus, and a fine-tuned BERT model, and compared the performances of different dictionaries and machine learning algorithms.

We presented the elaboration of a new political sentiment dictionary with an average of 0.57 F1 score in positive and negative dimensions when only positive and negative sentences were present in the test corpus. In comparison with the Hungarian general-purpose dictionary, the NRC Word-Emotion Association Lexicon, and the Bing-Liu sentiment dictionary, the new POLTEXTLAB dictionary had better performance in all categories. It was proved that the developed domain-specific dictionary performed better on political texts than the domain-independent ones used for the comparison.

We have presented an example of why morphologically rich languages, such as Hungarian, require special treatment during every step of Natural Language Processing tasks and how significant improvement can be achieved by using appropriate pre-processing methods.

We found that negative emotions are expressed in a much more complex way, as shown by the system of inductively identified emotions. Later we aggregated these emotions into positive, negative, and mixed sentiment categories. Our important conclusion was that the results achieved with dictionaries could surpass the simplest tested machine learning algorithm (Naive Bayes). This may be related to the proportion of negative sentences with a significant preponderance in the corpus and the consequent overfitting of the NB.

We also tested the efficiency of various machine learning algorithms, including a fine-tuned BERT model, which has shown excellent performance.

Overall, we believe that our paper demonstrates well the advantages and disadvantages of the different approaches. For researchers with limited NLP skills, we offer a domain-specific dictionary option, which is a fast and cost-effective solution, although its performance is significantly below that of machine learning approaches. For those who have the required expertise and can afford the cost of manual annotation and fine-tuning, the use of a transformer-based model is the best choice.

Data availability Our data and scripts are available: https://osf.io/ktn9p/?view_only=3943d011e2ac4cf5a4600fa877c24d72.

References

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review, 54*(8), 5789–5829.

Akleman, E. (2020). Deep learning. *Computer, 53*(09), 17–17.

Alshari, E. M., Azman, A., Doraisamy, S., Mustapha, N., & Alkeshr, M. (2018). Effective method for sentiment lexical dictionary enrichment based on Word2Vec for sentiment analysis. *Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), 2018*, 1–5.

Alwan, J. K., Hussain, A. J., Abd, D. H., Sadiq, A. T., Khalaf, M., & Liatsis, P. (2021). Political arabic articles orientation using rough set theory with sentiment Lexicon. *IEEE Access, 9*, 24475–24484.

Bakken, P. F., Bratlie, T. A., Sánchez-Marco, C., & Gulla, J. A. (2016). Political news sentiment analysis for under-resourced languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2989–2996).

Balahur, A., & Steinberger, R. (2009). Rethinking sentiment analysis in the news: From theory to practice and back. *Proceeding of WOMSA, 9*, 1–12.

Bene, M. (2017). Go viral on the facebook! interactions between candidates and followers on facebook during the Hungarian general election campaign of 2014. *Information, Communication & Society, 20*(4), 513–529.

Bene, M., & Szabó, G. (2021). Discovered and Undiscovered fields of digital politics: Mapping online political communication and online news media literature in Hungary. *InterSections, 7*(1), 1–21.

Bhowmick, P. K., Basu, A., & Mitra, P. (2009). Reader perspective emotion analysis in text through ensemble based multi-label classification framework. *Comput. Inf. Sci., 2*(4), 64–74.

Biba, M., & Mane, M. (2014). Sentiment analysis through machine learning: An experimental evaluation for Albanian. In *Recent Advances in Intelligent Informatics: Proceedings of the Second International Symposium on Intelligent Informatics (ISI'13), August 23–24 2013, Mysore, India*, (pp. 195–203).

Bobicev, V., & Sokolova, M. (2017) Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, (pp. 97–102). https://doi.org/10.26615/978-954-452-049-6_015

Boomgaarden, H. G., & Schmitt-Beck, R. (2019). The Media and Political Behavior. In H. G. Boomgaarden & R. Schmitt-Beck (Eds.), *Oxford Research Encyclopedia of Politics.* Oxford University Press.

Boukes, M., Van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures, 14*(2), 83–104.

Boumans, J. W., & Trilling, D. (2018). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. In M. Karlsson & H. Sjøvaag (Eds.), *Rethinking Research Methods in an Age of Digital Journalism* (pp. 8–23). Routledge.

Cabanac, M. (2002). What is emotion? *Behavioural Processes, 60*(2), 69–83.

Chapman, C. R., & Nakamura, Y. (1998). A bottom up view of emotion. *ASSC* Seminar, http://Server.Phil.vt.Edu/Assc/Watt/Chapman1.Html.

Chetviorkin, I., & Loukachevitch, N. (2013). Evaluating sentiment analysis systems in Russian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, (pp. 12–17).

Cho, J., Boyle, M. P., Keum, H., Shevy, M. D., McLeod, D. M., Shah, D. V., & Pan, Z. (2003). Media, terrorism, and emotionality: Emotional differences in media content and public reactions to the September 11th terrorist attacks. *Journal of Broadcasting & Electronic Media, 47*(3), 309–327.

Cohen, J. (1960). A coefficient of agreement for nomimal scales. *Educational and Psychological Measurement, 20*, 37–46.

De Vreese, C. H., & Semetko, H. A. (2002). Cynical and engaged: Strategic campaign coverage, public opinion, and mobilization in a referendum. *Communication Research, 29*(6), 615–641.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. ArXiv Preprint ArXiv:2005.00547.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint ArXiv:1810.04805.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs]. http://arxiv.org/abs/1810.04805

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing, 34*(6), 480–488.

Dilai, M., Onukevych, Y., & Dilay, I. (2018). Sentiment analysis of the US and Ukrainian presidential speeches. *Computational Linguistics and Intelligent Systems, 2*(2018), 60–70.

Döbrössy, B., Makrai, M., Tarján, B., & Szaszák, G. (2019). Investigating sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, (pp. 187–193).

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science, 161*, 707–714.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.

Farnsworth, S. J., & Lichter, S. R. (2005). Reporting on two presidencies: News coverage of George W. Bush;s first year in office. *Congress & the Presidency: A Journal of Capital Studies, 32*(2), 91–108.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89.

Feng, S., Kang, J. S., Kuznetsova, P., & Choi, Y. (2013). Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 1774–1784).

Fu, P., Lin, Z., Yuan, F., Wang, W., & Meng, D. (2018). Learning sentiment-specific word embedding via global sentiment representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-dependent sentiment classification with BERT. *IEEE Access, 7*, 154290–154299. https://doi.org/10.1109/ACCESS.2019.2946594

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics, 49*(3), 291–304.

Griffiths, P. E. (2008). *What emotions really are*. University of Chicago Press.

Haider-Markel, D. P., Allen, M. D., & Johansen, M. (2006). Understanding variations in media coverage of US supreme court decisions: Comparing media outlets in their coverage of Lawrence v Texas. *Harvard International Journal of Press/politics, 11*(2), 64–85.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, *2016*, (p. 595).

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity, 51*(6), 2623–2646.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 168–177).

Huang, S., Niu, Z., & Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems, 56*, 191–200.

Itani, M. M., Zantout, R. N., Hamandi, L., & Elkabani, I. (2012). Classifying sentiment in arabic social networks: Naive search versus naive bayes. In *2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, (pp. 192–197).

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98* (pp. 137–142). Springer.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Kaya, M., Fidan, G., & Toroslu, I. H. (2012). Sentiment analysis of Turkish political news. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, *1*: 174–180.

Kepplinger, H. M. (2002). Mediatization of politics: Theory and data. *Journal of Communication, 52*(4), 972–986.

Khairnar, J., & Kinikar, M. (2013). Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications, 3*(6), 1–6.

Khoo, C. S.-G., Nourbakhsh, A., & Na, J.-C. (2012). Sentiment analysis of online news text: A case study of appraisal theory. *Online Information Review*.

Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *ArXiv* Preprint ArXiv: 1507.00955.

Koljonen, J., Öhman, E., Ahonen, P., & Mattila, M. (2022). Strategic sentiments and emotions in post-Second World War party manifestos in Finland. *Journal of Computational Social Science, 5*(2), 1529–1554. https://doi.org/10.1007/s42001-022-00181-9

Koltsova, O. Y., Alexeeva, S., & Kolcov, S. (2016). An opinion word lexicon and a training dataset for Russian sentiment analysis of social media. *Computational Linguistics and Intellectual Technologies: Materials of Dialogue, 2016*, 277–287.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Sage Publications.

Lakoff, G., & Kövecses, Z. (1987). The cognitive model of anger inherent in American English. In D. Holland & N. Quinn (Eds.), *Cultural Models in Language and Thought* (pp. 195–221). Cambridge University Press.

Lengauer, G., Esser, F., & Berganza, R. (2012). Negativity in political news: A review of concepts, operationalizations and key findings. *Journalism, 13*(2), 179–202.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv: 1907.11692. http://arxiv.org/abs/1907.11692

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing, 2*(2010), 627–666.

Loukachevitch, N., & Levchik, A. (2016). Creating a general Russian sentiment lexicon. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1171–1176.

Lutz, B., Pröllochs, N., & Neumann, D. (2018). Sentence-Level sentiment analysis of financial news using distributed text representations and multi-instance learning. ArXiv Preprint ArXiv: 1901.00400.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Miháltz, M. (2013). OpinHuBank: Szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013)* (pp. 343–345).

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In P. Gowtham (Ed.), *Emotion measurement* (pp. 201–237). Elsevier.

Mullen, T., & Malouf, R. (2006). A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, (pp. 159–162).

Nemeskey, D. M. (2020). *Natural Language Processing Methods for Language Modeling*. https://hlt.bme.hu/media/pdf/nemeskey_thesis.pdf

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 3–16).

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1).

Oravecz, C., Váradi, T., & Sass, B. (2014). The Hungarian Gigaword Corpus. In *Proceedings of LREC*, 1719–1723.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *ArXiv Preprint Cs/0205070*.

Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis (foundations and trends (R) in Information Retrieval)*. Now Publishers Inc.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. ArXiv 2018. *ArXiv* Preprint ArXiv:1802.05365, 12.

Plutchik, R. (1982). *A psychoevolutionary theory of emotions*. Sage Publications.

Prószéky, G., & Miháltz, M. (2008). Magyar WordNet: Az első magyar lexikális szemantikai adatbázis. *Magyar Terminológia, 1*(1), 43–57.

Pyry, T., Pekka, M., Ankur, S., & Oskar, A. (2014). *Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2152–2157.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Rauh, C. (2018). Validating a sentiment dictionary for German political language—A workbench note. *Journal of Information Technology & Politics, 15*(4), 319–343.

Ray, P., & Chakrabarti, A. (2017). Twitter sentiment analysis for product review using lexicon method. *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, (pp. 211–216).

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22): 41–46.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145.

Sağlam, F., Sever, H., & Genç, B. (2016). Developing Turkish sentiment lexicon for sentiment analysis using online news media. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, 1–5.

Sarlós, G. (2015). Risk perception and political alienism: Political discourse on the future of nuclear energy in Hungary. *Central European Journal of Communication, 8*(14), 93–111.

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, (pp. 3–596).

Singh, S., & Mahmood, A. (2021). The NLP cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access, 9*, 68675–68702.

Song, Y., Wang, J., Liang, Z., Liu, Z., & Jiang, T. (2020). Utilizing BERT Intermediate Layers for Aspect Based Sentiment Analysis and Natural Language Inference. ArXiv: 2002.04815 [Cs]. http://arxiv.org/abs/2002.04815

SURYONO, R. R., & Indra, B. (2020). P2P Lending sentiment analysis in Indonesian online news. In *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, (pp. 39–44).

Szabó, G. (2020). Emotional communication and participation in politics. *Intersections. East European Journal of Society and Politics, 6*(2).

Szabó, G., & Szilágyi, S. (2022). *Morál a médiában: Az ukrajnai háború az online hírportálokon a 2022-es országgyŰlési kampány idején*.

Szabó, M. K. (2015). Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. *Nyelv, Kultúra. Társadalom. Segédkönyvek a Nyelvészet Tanulmányozásához, 177*, 278–285.

Tikk, D. (2007). *Szövegbányászat*. Typotext.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1).

Uribe, R., & Gunter, B. (2007). AreSensational'News Stories more likely to trigger viewers' emotions than non-sensational news stories? A content analysis of british TV News. *European Journal of Communication, 22*(2), 207–228.

Üveges, I., & Ring, O. (2023). HunEmBERT: A fine-tuned BERT-model for classifying sentiment and emotion in political communication. *IEEE Access, 11*, 60267–60278. https://doi.org/10.1109/ACCESS.2023.3285536

Van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schlobach, S. (2008). Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations. *Journal of Information Technology & Politics, 5*(1), 73–94.

Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications, 42*(11), 4999–5010.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, *30*.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures, 11*(4), 245–265.

Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 6 (pp. 25–631).

Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge University Press.

Yang, C., Lin, K. H.-Y., & Chen, H.-H. (2007). Emotion classification using web blog corpora. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, (pp. 275–278).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*., 5753–5763.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*(3), 6527–6535.

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *ArXiv Preprint* ArXiv: 1702.01923.

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political CommunicAtion, 29*(2), 205–231.

Zsibrita, J., Vincze, V., & Farkas, R. (2013). magyarlanc: A tool for morphological and dependency parsing of hungarian. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP, 2013*, 763–771.