IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

HunEmBERT: a fine-tuned BERT-model for classifying sentiment and emotion in political communication

ISTVÁN ÜVEGES[1], ORSOLYA RING[2]

[1]Centre for Social Sciences, Budapest, Tóth Kálmán utca 4., Hungary (e-mail: uveges.istvan@tk.hu)

[2]Centre for Social Sciences, Budapest, Tóth Kálmán utca 4., Hungary (e-mail: ring.orsolya@tk.hu)

„The research was supported by the Ministry of Innovation and Technology NRDI Office and the European Union, in the framework of the RRF-2.3.1-21-2022-00004 Artificial Intelligence National Laboratory project."

ABSTRACT The growing number of digitally accessible text corpora and the accelerating development of NLP tools and methods (particularly the emergence of powerful large-scale language models) have allowed their widespread use in various classification tasks, including the vast field of sentiment analysis. However, these models must often be fine-tuned to perform this task efficiently. Therefore, we aimed to create a transformer-based fine-tuned model for the emotion and sentiment analysis of Hungarian political texts. The training data for the model were the manually annotated parliamentary speech texts from 2014 to 2018, which have the advantage of being rich in various emotions. The compiled corpus can be freely used for research purposes. In our work, we describe in detail the process of fine-tuning the Hungarian BERT model for sentiment and emotion classification, the performance achieved, and the typical classification errors, mainly due to a lack of recognition of pragmatic and other language use features by the fine-tuned models.

INDEX TERMS fine-tuned BERT-model, huBERT, Emotion analysis, Sentiment analysis, political communication

I. INTRODUCTION

The growing number of digitally accessible text corpora and the rapid development of NLP methods, particularly the emergence of powerful large-scale language models, have allowed their widespread use in various social science classification tasks. Among these, sentiment and emotion analysis is one of the most popular, which are related but different research topics. Sentiment analysis, or opinion mining, recognizes positive, negative, and neutral opinions in a text [1]. In contrast, emotion analysis identifies emotions (e.g. joy, anger, sadness) expressed in a text [2].

As is reflected in the relevant literature [3]–[6], the expression of emotions is language-specific, so different analytical models require linguistic adaptation. The available tools for emotion analysis are mainly designed for English texts and require contextual adaptation to give reliable results - especially for morphologically rich languages like Hungarian. In Hungarian, grammatical information that in other languages (e.g. English) is expressed by prepositions is encoded in inflections, i.e. the meaning and grammatical function of words are changed by various additions, especially suffixes, so that a word can have many different forms due to the possible inversion and derivational morphemes.

Recently studies of sentiment or emotion analysis of political texts have increased due to the large quantity of information available online and the development of natural language processing algorithms [7]–[15]. Political discourse cannot be reduced to a statement of facts, the tone of a text is just as important, often vital in decision-making and forming political judgment [15]. Automated sentiment scoring provides a way to measure the tone of political texts [7]. However, to be effective, the scoring must match the specific contexts of political language [13].

Nowadays, more and more studies use texts as data to analyze emotions as part of political discourse, but only a few uses large language models (LLMs). As these are now available in several languages, they promise to improve the

analysis of large amounts of non-English texts significantly. Their fine-tuned versions are also well suited to emotion analysis tasks [16], focusing on determining which emotions can be identified in a text. This approach can be described as a document-, sentence-, or aspect-level classification, where the first two applications focus on the emotion of the whole document or sentence. With target-level classification, the exact relation to each emotion is more accurately identified as we can link emotion to specific objects (e.g., Named Entities).

The paper's main contributions are as follows:

(1.) Using our manually annotated emotion corpus, we fine-tune the BERT-based models' implementation for Hungarian (huBERT) for both sentiment and emotion classification.

(2.) The analysis of the errors in the classification provides insight into the current capabilities and typical failures of the huBERT model and, to some extent, into the limits of what BERT-based models in general can learn.

(3.) By making the models publicly available, we aim to make them accessible and usable for other researchers to classify sentiment and emotions with regard to Hungarian political texts.

Our main goal is to extend the possibility of sentiment and emotion analysis to the political domain of Hungarian texts, which has not yet been investigated. The presented models are intended as basic resources that can contribute to the international discourse and help research in political and social science in general. They do this by adding to the analytical possibilities of political texts the opportunity to analyze the expressed sentiments and emotions.

The study is structured as follows. Section II. briefly describes the most important approaches that can be used for sentiment and emotion analysis in the context of political texts. After that, section III. presents the main steps of the project that forms the basis of this article, from data collection to model training. Section IV. presents the results of the two fine-tuned models, while Section V. discusses the reasons behind their most typical classification errors. The paper is then closed with a short conclusion and a presentation of proposed future work.

II. RELATED WORKS

The text-as-data approach has increasingly been applied in political science over the past decade [17]–[21], and machine learning has been part of the toolkit for more than three decades [22], [23]. Its applications range from methodological work [24] to specific policy research use cases [25], [26]. These studies often employ supervised learning methods for classification tasks using support vector machines, random forest classifiers, logistic regression or Naïve Bayes [27]–[31]. However, a significant problem in social science research is that a large number and a wide range of resources are only available to analyze a few privileged languages (e.g. English and Chinese). In contrast, languages with few resources are more difficult to use in research due to the limited availability of research analysis tools [32].

A. TEXTUAL ANALYSIS OF POLITICAL COMMUNICATION

Political communication encourages political action by eliciting emotional impact and propagating different ideas. As a result of the technical and social changes of the past decades, the number of participants in communication and the number of communication channels available has expanded considerably, which has also impacted the nature and intensity of political communication. Political actors respond to the growing expectations of their role by professionalizing their communication. Political speeches are well-designed actions that aim to inform as well as to persuade an audience. Parliament is a particularly important arena for such communication, where elected representatives discuss submitted bills and other matters of national importance.

During parliamentary debates, various topics arise, arguments and counterarguments collide, and through them, a political agenda is formed that then thematizes public debates [33]. Research on the expression of emotions in political communication has been increasingly emphasized in recent years in international and Hungarian social science research [8], [34]–[38]. These studies primarily analyze the speeches of politicians in the media and on social media [39]–[41]. Analyzing the emotional charge of political and especially parliamentary speeches and their aspects with NLP tools is a novel idea [28], [29], especially in Hungarian.

B. SENTIMENT AND EMOTION CLASSIFICATION

Sentiment and emotion analysis is a significant research topic. However, "sentiment analysis" and "emotion analysis" are often used interchangeably. While sentiments and emotions are related, these two concepts have different meanings. Hence, we should distinguish between them. In this study, we use sentiment analysis only to determine whether the text expresses a positive, a negative, or a neutral opinion. Since sentiment words might not even indicate any real sentiment, or could bear several meanings[1], and the difficulty in detecting the manner of expression – like sarcasm, cynicism, or mockery – the analysis still holds its challenges [42].

By the task of emotion analysis, we mean emotion classification, which means both the task of detecting if a text conveys any emotion and the task of classifying a detected emotion in a text into a set of defined emotions. Emotion analysis is a more complex way of classifying opinions as we move beyond the general distribution by

studying the specific emotion of the texts, for instance, happiness, anger or fear.

1) Dictionary methods

Sentiment lexicons are compilations of so-called sentiment words or phrases, each word usually carrying a positive or negative tone [42]. Sentiment analysis based on dictionaries is much less costly than applying more complex machine learning methods. dictionaries can be sources of features in the machine-learning framework [9].

The dictionary uses the proportion of keywords in a text to rank the full text. However, lacking semantic context, they can misclassify text and are therefore not widely applicable. Although there are methods that enhance both generalizability and vocabulary coverage by using word embeddings to augment dictionaries [43], the performance of dictionary analysis remains limited, especially in complex tasks such as emotion categorization.

Lexicon-based sentiment analysis usually begins with a list of words, but synonym detection may later be employed [44]. Much research is associated with social media data, such as tweets [45]–[49]. O'Connor et al. [50] examined the correlation between public opinion polls and tweet sentiment. No link was found to election results, but there was a connection to the approval of a president. In Russian, machine learning is usually the most successful approach, apart from the case of political news, where a lexicon-based methodology is better due to the variety of topics [51]. Koltsova et al. [52] created a lexicon to examine political sentiment in Russian social media, and it was most effective for negative and less extreme sentiments. In Arabic, the lexicon-based approach achieved 83% accuracy [53]. German sentiment dictionaries were validated for use in political science and found to be better at detecting positive emotions than negative ones [13]. Dilai et al. [54] compared US (2016) and Ukrainian (2014) presidential speeches with emotion detection, finding them to be subjective and mainly positive. More recent research has seen word embedding techniques used to generate or expand sentiment lexicons [55], [56]. Domain-specific embeddings are used with a label propagation framework to create domain-specific sentiment lexicons from seed words [57].

2) Supervised learning approaches

Supervised learning is a common technique for solving sentiment and emotion classification problems [9], [42], [43]. Before the advent of BERT models, the algorithms traditionally used were Naïve Bayes, k-NN, decision trees and Support Vector Machines. Naïve Bayes is a simple set of probabilistic algorithms which is suitable for data sets with small sizes; it has two variants, Multinomial and Bernoulli [58], [59]. Logistic regression is an exponential or log-linear classifier which works by extracting weighted features, taking logs and combining them linearly [60]. Support Vector Machines (SVMs) are effective at traditional text categorization, outperforming Naïve Bayes [61]. SVMs have helpful attributes for text classification, such as the ability to work on a high number of features without overfitting, working with sparse matrices, a built-in kernel trick, and being able to use various domains without adaptation [61]. A significant advantage of supervised methods such as Naïve Bayes, k-NN, decision trees and SVMs over dictionaries is their improved performance and the fact that they provide clear statistics on the performance of the models.

A disadvantage, however, is that such methods require a significant amount of manually labeled data for accurate predictions. In addition, with multi-class classification, there is often a problem of imbalance in the amount of data between classes. Models trained on small or skewed data sets can be optimized by unsupervised pre-training using pre-trained word embeddings that rely on large data sets [62], [63]. The so-called BERT models are unsupervised language models whose context-dependent representations have been generated using a remarkably large amount of text [64]. Thus, BERT helps to create context-specific embeddings by providing a pre-trained universal model. One of the main drawbacks of this method is the non-negligible computational resource requirements of the pre-training process. Before the advent of BERT models, the algorithms traditionally used were Naïve Bayes, k-NN, decision trees and Support Vector Machines [61].

There are many different studies which classify sentiment or emotion by supervised machine learning approaches in the political domain, for example, in social media, online news or the text of speeches [65]–[69]. As Attelveld et al. found, machine learning approaches perform significantly better than off-the-shelf sentiment analysis tools. Although these often do not achieve the level of validity expected of text analysis methods in general, the results of crowd coding can compete with the quality delivered by qualified coders, making them a cheaper and particularly transparent and repeatable alternative [69].

3) Transformer-based models

In Natural Language Processing, Language Models determine the probability of word or word sequences by analyzing textual data and mainly by learning abstractions of syntactic and semantic rules. The model then applies these rules to solve linguistic-based problems (such as part of speech tagging) and to predict or generate new sentences accurately. This general language "knowledge" acquired on a large data set can then be used to solve downstream tasks, such as sequence labeling or named entity recognition [70].

Considering that human languages follow a sequential structure (texts are composed of sentences, their constituent clauses - syntactic constituents, which are built

up by words or word combinations, and words are ultimately built of a sequence of characters in the written representation), the beginning of language modeling was marked by Recurrent Neural Network (RNN) architectures [71]. These were the first neural networks in which the states of individual neurons within a layer could interact. [72].

As RNNs could suffer from "vanishing" or "exploding" gradients when handling longer sequences, an improved architecture for RNNs, the Long Short-Term Memory (LTSM) architecture [73] was developed to address this. Given that the entire history of the processed sequence was to be stored in a single state vector, this was not perfectly efficient in handling longer contexts.

The dramatic surge in available computing power soon led to solutions based on deep-learning neural networks [74]. Focusing on the concept of "attention," the first breakthrough transformer architecture was released in 2017 [75]. The original transformer architecture was based on an encoder-decoder architecture. The former's layers iteratively process the received sequential input (e.g., natural language text) and form encodings that contain information about which parts of the input are relevant to each other. The decoder layers work oppositely, taking all the encodings and storing the contextual information to form an output sequence. Such models (e.g., GPT-1 [76], and BERT [64]) were the first to achieve significant success in 2018 in various NLP tasks, such as language modeling, sentiment analysis and question answering.

This has led to the issue of transfer learning, where the aim is to encode knowledge accumulated while learning a particular task that is also suitable for solving others [70]. Several recent language models, such as XLNET [77], or RoBERTa [78], can be seen as such attempts. These SOTA language models have proven to be of pioneering importance in recent years for sentiment analysis tasks [16].

In the latest period, pre-trained language models have become state-of-the-art solutions for most NLP tasks. Models with hundreds of millions of tunable parameters, such as ELMo [79], GPT [76], BERT [64], or RoBERTa [78] have led to significant improvements in a number of (previously difficult) NLP tasks, such as question answering, machine translation, or, most relevantly for the current paper, sentiment or emotion analysis.

Within sentiment analysis, BERT has mostly been used in aspect-based sentiment analysis [80]–[84], while few authors focused on emotion analysis[16], [85] in connection to a specific event, [86]–[89] or on improving the fine-tuning performance of the BERT model by introducing semi-supervised adversarial learning [90], [91].

III. OUR APPROACH

This section first presents the emotion and sentiment corpus constructed from Hungarian political speeches and used for efficiency measurement and fine-tuning. We then describe the performance of huBERT(-base) in the sentiment and emotion classification tasks performed on it. This is followed by a description of the fine-tuning of huBERT and the evaluation of the resulting HunEmBERT models' performance in both sentiment and emotion classification (from which HunEmBERT3 is applicable to sentiment, and HunEmBERT8 for emotion classification)[1]. Finally, the results are compared with the effectiveness of emotion classification previously achieved on the ISEAR dataset, considering the necessary limitations.

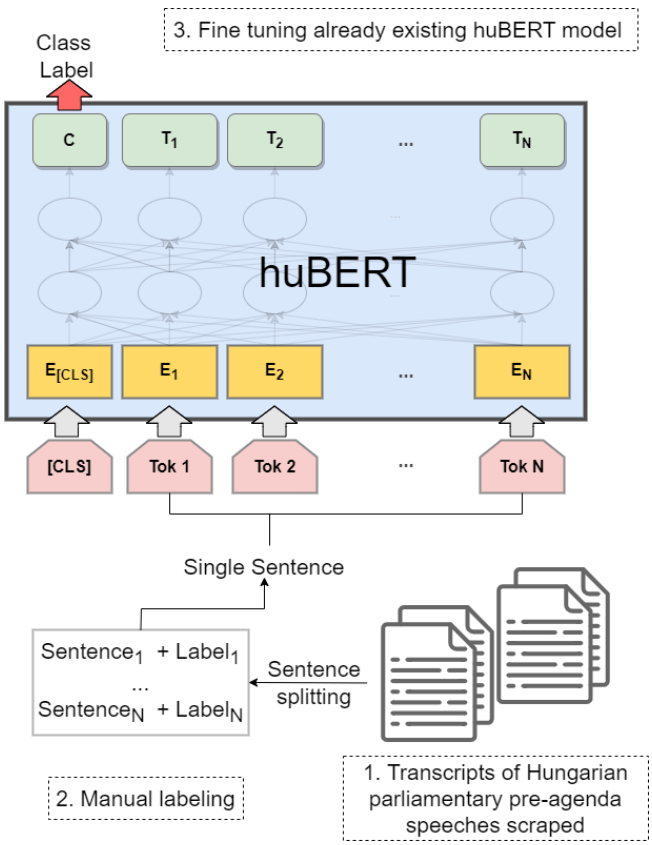

FIGURE 1 . A summary of the main steps in building the fine-tuned sentiment-, emotion classification models for Hungarian political texts ([35], Fig. 4. was partially adapted).

[1] All replication materials are available at: https://osf.io/67zsf/?view_only=a23e5b6ba5ef443892a885a3f1d1d1e7

A. DATA SELECTION AND CORPUS STATISTICS

Previously, there were only two dedicated corpora in the Hungarian language freely available for research purposes. In our project, we first built a language and domain-specific corpus to be used, among other things, to finetune different Large Language Models.

To build an emotion and sentiment corpus (HunEmPoli) [92], we selected Hungarian parliamentary pre-agenda speeches delivered by Members of Parliament from 2014 to 2018. Transcripts of these speeches are publicly available at the official website of the Hungarian National Assembly (parlament.hu). During this period, a total of 1008 speeches were made, all of which are included in the corpus. This amounted to a total of 764008 tokens or 36475 sentences. The annotators were political science students, native Hungarian speakers with no prior experience in automated text analyses, so they were provided with a detailed annotation guideline. The inter-annotator agreement measured during corpus quality assurance is 0.7574 (Kappa), indicating strong agreement.

Category	Count	Ratio	Sentiment	Count	Ratio
Neutral	351	1.85%	Neutral	351	1.85%
Fear	162	0.85%	Negative	11180	58.84%
Sadness	4258	22.41%			
Anger	643	3.38%			
Disgust	6117	32.19%			
Success	6602	34.74%	Positive	7471	39.32%
Joy	441	2.32%			
Trust	428	2.25%			
Sum	19002				

TABLE 1. Statistics of the filtered corpus (Count: number of sentences).

Pre-agenda speeches are presented in the Hungarian legislature at the beginning of each parliamentary session. Frontbencher MPs and members of the government generally give them. The speaker is free to choose the topic of his or her speech, usually followed by a short debate.

Although the texts of our corpus are spoken language data, their style is official, and it differs from the spoken language corpora available in Hungarian, which contain spontaneous speech and/or have an informal style [92]–[94]. They contain many addressing terms and thanks (*Dear House*, *Thank you for giving me the floor*), and they use almost exclusively formal speech. However, the transcripts do not contain the hesitations, short breaks, or false beginnings typical of live speech.

The collected pre-agenda speeches were originally annotated at the clause level. However, our goal was to train a model capable of classifying whole sentences into a sentiment or emotion category. For this reason, texts were first segmented into sentences automatically.

For this purpose, we used the transformer-based pipeline[2] developed for the HuSpaCy [95] natural language processing toolkit for Hungarian[3]. Then, the emotion labels that the sentence has been annotated with at the clause level are determined for each sentence. If two clauses in the same sentence got different emotion labels, the sentence got both.

Several options were available for cases where a sentence had more than one label. The first option is to include "multiple instances" of the sentence in the training set, giving each label to match the original options. Model training could be considered a multiclass + multilabel problem; in this case, for example, using the OVR (One Versus Rest) approach, one model can be trained for each possible label, which can act as binary classifiers one by one. However, our main goal was to create a single model for sentiment and another for emotion classification in the political context, so we clarified the labels for such sentences.

A trivial solution for specifying labels is to choose one of several (correct) options (e.g., randomly or leaving the decision to a master annotator). In practice, however, the models trained in the presence of such train examples (i.e., ones with more than one acceptable label) have performed relatively poorly during the preliminary investigations. The quality of predictions could be significantly improved by removing such ambiguous cases. To achieve this, we removed from the training data all sentences whose clauses were not exclusively annotated with a single label during the manual annotation. With this filtering, about 52.09% of the original data were retained. Table 1 shows the important characteristics of the resulting data set.

B. FINE-TUNING THE huBERT MODEL

For this current experiment, huBERT [96], as the first implementation of the BERT-base for Hungarian, was used[4]. During the pre-training process, the Hungarian Institute of Computer Science and Automation Research (SZTAKI) used 5 TPUs and a 256-core v3-256 TPU pod with 4 TiB memory

[2] https://huggingface.co/huspacy/hu_core_news_trf

[3] Since all starting character positions for annotated clauses were stored during the annotation, after getting the list of sentences, the containing sentence could already be determined automatically.

[4] https://huggingface.co/SZTAKI-HLT/hubert-base-cc

for two weeks under the TensorFlow Research Cloud program. The model was first trained on the smaller 110 million words Hungarian Wikipedia corpus[5], using which the training reaches peak performance on Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) tasks between 300,000 and 400,000 steps. Since diacritics are distinctive in Hungarian, a cased version that waits for unprocessed text as input has been produced based on BERT's original training code.

During the pre-training, and similarly to English BERT, a 30,000-token WordPiece dictionary and sequences with a maximum of 512 tokens were used [97]. There are two methods for using the mentioned pre-trained models for a given (more specific) task: feature-based and fine-tuning. A task-specific architecture is supplemented with a general language representation in the former. In contrast, the number of task-specific components is reduced in the latter, and the desired result is achieved by fine-tuning the pre-trained parameters. BERT models (in general, and also huBERT) implement the fine-tuning approach, similar to the GPT language model [64], [98].

In essence, fine-tuning changes the model's parameters to perform as efficiently as possible in solving the problem. In this context, the usual way of measuring performance is to evaluate the loss of function on both the training and the validation set. For this purpose, the original corpus was split in a standard way into 80% train, 10% validation and 10% test sets, from which the validation set was used for the fine-tuning process and the test set to evaluate the models' final metrics in terms of Precision, Recall and F1.

To achieve the best possible results, all layers of huBERT were set to be trainable instead of "freezing" them. For the fine-tuning process, we chose the Trainer API of the transformers' library, which provides an easy-to-use high-level abstraction to simplify all of the "boilerplate code" one typically has to write when making their training loops in deep learning. The original BERT paper [64] suggests a batch size of 16-32, 2-4 epochs, and a learning rate of 5e-5 - 2e-5 as general values based on the experience of 11 NLP tasks tested there. We also chose hyperparameters within this range. The only exception was the choice of the learning rate, as preliminary tests suggested that the model tended to overfit quickly, so this was set at 5e-6. The batch size was set to 16. A maximum of 10 epochs was specified when the Trainer class was initialized, with an early stop possibility after 2 epochs of failure to improve the fine-tuned model performance. The evaluation metric was accuracy. The inputs to the model were (in a standard way) the "input IDs", "token type IDs" and the attention masks (the latter of which were padded to 512 tokens). These were pre-generated using huBERT's tokenizer.

For both sentiment and emotion classification models, the same experimental setup was used, the difference being the number of labels to be predicted (3 for the former: positive, negative, and neutral, while for the latter 8 labels were possible: fear, anger, etc.). Accordingly, two versions of the training data were available, the first with sentences labeled purely by sentiment value and the second with emotion labels.

C. TRAIN- AND VALIDATION LOSS

In supervised machine learning, the main goal is to produce a model that learns from the training data and generalizes over features of never-before-seen instances.

Solving this essentially results in an optimization problem called Structural Risk Minimization (SRM), which aims to train the most efficient model from a finite set of training data [99].

In this minimisation problem, the loss function is the component that helps determine the distance between the actual output of the model and the expected output, which can be used to modify the model to achieve better results [100].

Epoch	Train loss	Val. Loss
1	0.3042	0.2396
2	0.1676	0.2320
3	0.0979	0.287
4	0.0649	0.3545
5	0.0436	0.365

TABLE 2. Train and validation loss 3 (positive, negative sentiment + neutral) categories

Epoch	Train loss	Val. Loss
1	0.7864	0.6552
2	0.4863	0.6461
3	0.3158	0.7287
4	0.1966	0.8283
5	0.1292	0.945

TABLE 3. Train and validation loss were measured when fine-tuning to 8 (7 emotions + neutral) categories.

By measuring the loss function values achieved on the training and validation sets, we can infer two typical problems of deep learning networks, underfitting and overfitting. In the former case, the loss of function shows a decreasing trend on both sets after the epochs used for

[5] Available as a part of the Webcorpus2.0: https://hlt.bme.hu/en/resources/webcorpus2

training—further training is required. In the latter case, after the initial decrease, the validation loss starts to increase while the value of the training loss continues to decrease; at this point, the model cannot effectively generalize on the new data anymore.

We choose CategoricalCrossentropy as the loss of function and Adam (Adaptive Moment estimation) optimizer [101] with learning rate = 5e-6 (as mentioned before) and decay = 1e6 parameters from TensorFlow.Keras implementation.

D. DETERMINING THE OPTIMAL EPOCH NUMBER

In our case, Table 2 and Table 3 illustrate the measured loss values after the initial 5 epochs fine-tuning phase for the sentiment (3 categories) and emotion (8 categories) classification (the process was then terminated with an early stop).

In both cases, it can be seen that the models perform best after 2 epochs of training. After that, the validation loss started to increase again, while the training loss decreased sharply as a clear sign of overfitting on train instances.

After determining the optimal number of epochs, the final models were trained on this basis, and the results were evaluated on the test set.

IV. RESULTS

The fine-tuned model for sentiment classification achieved 0.866 macro average and 0.9149 weighted average in terms of F-Score. The latter describes the model's performance somewhat better, given that the "neutral" category with the lowest F-value (0.76) was significantly underrepresented in the corpus (the support here was only 35 sentences in the test set).

The results show a more significant variance concerning the fine-tuned model for emotion classification. The measured metrics reflect that the most correctly classified categories (Success, Disgust and Sadness, which account for about 83.43% of the total data set, cf. Table 1) correlate with high item counts.

In terms of Precision, sentences that belong to the Neutral and Anger category both achieved high Precision compared to categories with similar numerosity (0.85 and 0.78, respectively). Regarding recall, Joy achieved 0.63, similar to the result of Sadness (0.62), while the numerosity of the latter is almost 10 times bigger in the corpus. For this 8-emotion category, the macro average F-Score was 0.64. In contrast, the weighted average (which, again, gives us a more realistic picture of the models' expected performance in real scenarios given the category imbalance) was 0.7743. This result is comparable, for instance, with the performance of transformer-based language models (BERT, RoBERTa etc.) obtained in the ISEAR [102] dataset, at least for those categories that were present in the training set in sufficient numbers [103].

Category	P	R	F1
Neutral	0.8333	0.7143	0.7692
Positive	0.8776	0.9198	0.8982
Negative	0.9439	0.9177	0.9306
Macro AVG	0.8849	0.8506	0.866
Weighted AVG	0.9157	0.9148	0.9149

TABLE 4. Best models' metrics on sentiment categories (+neutral) – P: Precision, R: Recall, F1: F-Score

	P	R	F1
Neutral	0.85	0.4857	0.6181
Fear	0.625	0.625	0.625
Sadness	0.8535	0.6291	0.7243
Anger	0.7857	0.3437	0.4782
Disgust	0.7154	0.8790	0.7888
Success	0.8579	0.8683	0.8631
Joy	0.549	0.6363	0.5894
Trust	0.4705	0.5581	0.5106
Macro AVG	0.7134	0.6281	0.6497
Weighted AVG	0.791	0.7791	0.7743

TABLE 5. Best models' metrics on 8 categories (7 emotions + neutral) – P: Precision, R: Recall, F1: F-Score

Here, the authors compared the results of BERT, RoBERTa, DistilBERT, and XLNet pre-trained transformer models in recognizing emotions from the ISEAR dataset. In the case of BERT, the used model was the base uncased version, which has an identical parameter number (110M) as the huBERT model has. The ISEAR dataset itself is a publicly available data collection constructed through cross-culture questionnaire studies from more than 30 countries, and it contains around 7600 sentences classified into seven distinct emotion labels: Joy, Anger, Sadness, Shame, Guilt, Surprise, and Fear in an almost perfectly balanced way. However, the used set of emotion labels is not always the same as the ones applied in the present research.

ISEAR category	F1 (BERT)
Anger	0.57
Disgust	0.67
Fear	0.75
Guilt	0.67
Joy	0.88
Sadness	0.78
Shame	0.6
AVG	0.7029

TABLE 6. BERT models' performance on ISEAR dataset (Data based on [**1**], Table 2)

Given the differences in the category systems, and the imbalance in the HunEmPoli corpus, the comparison makes most sense where HunEmPoli contained sufficient training data, and the category is present in the annotation systems of both corpora. The basis for comparison in this respect is, therefore, Sadness and Disgust. By comparing Table 5 and Table 6, it can be seen that the F-Score of Sadness (0.7243) somewhat underperformed the result obtained at ISEAR with the use of BERT-base. However, the F-Score of Disgust (0.7888) significantly outperformed it.

Regarding the average F-Scores, the BERT-base model achieved a 0.7029 (macro) F1 on ISEAR, while huBERT obtained 0.6497 on HunEmPoli. Since ISEAR is (almost) perfectly balanced, here again, the weighted average is more interesting, as it gives a more accurate picture of the fine-tuned huBERT model's performance. In the case of HunEmPoli, this was the value of 0.7743 mentioned earlier.

This suggests that, despite the complicating factors, the trained model generalizes well when recognizing emotions applied in a political context if sufficient training data was available during the training.

V. DISCUSSION

To better understand the models' typical errors, we used normalized confusion matrices for further evaluation and manually checked a subset of miscategorized sentences to find typical patterns that cause the majority of errors. Confusion matrices and ROC Curves are standard solutions for visualizing the errors of evaluated models in machine learning experiments, and they basically carry the same information [60]. By default, in the case of a confusion matrix, the y-axis of the matrix represents the accurate data labels. In contrast, the x-axis represents the labels predicted by the model, and the corresponding numbers are shown inside the matrix. A variant of this is when the matrix elements are normalized in some way.

A. CONFUSION MATRICES

To get the normalized version from the original matrix, each row element was divided by the sum of the entire row. Since each row here represents the total number of actual values for each class label, the final normalized matrix will show a percentage at every position (i.e., out of all true labels for a particular class, what was the % prediction of each class made by our model for that specific true label). Figures 2 and 3 show these results for the sentiment and emotion classification model.

Concerning emotion classification, it can be seen that cases where the model does not correctly define the emotion typically gives an incorrect label of the same sentiment class. Fear, for instance, is most often confused with Sadness (19%), which also has a negative sentiment value, while Joy is confused with Success (36%), both of which belong to positive sentiment. Perhaps the most ambiguous category in the data seems to be Trust, mixed with a high proportion of positive (Success, 23%) and negative (Disgust, 14%) sentiment, while itself is positive. For neutral sentences, the model tends to falsely assign them to emotions with a positive sentiment value (most often to Success, in 23% of the cases).

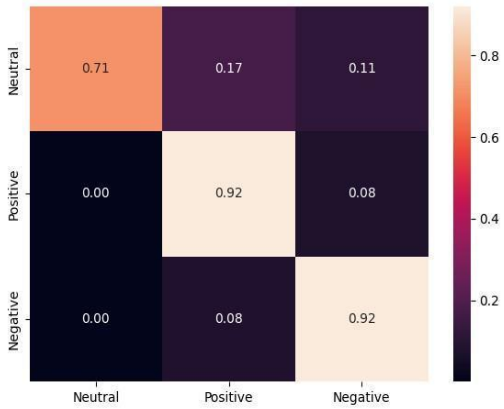

FIGURE 2. Normalized confusion matrix for sentiment classification

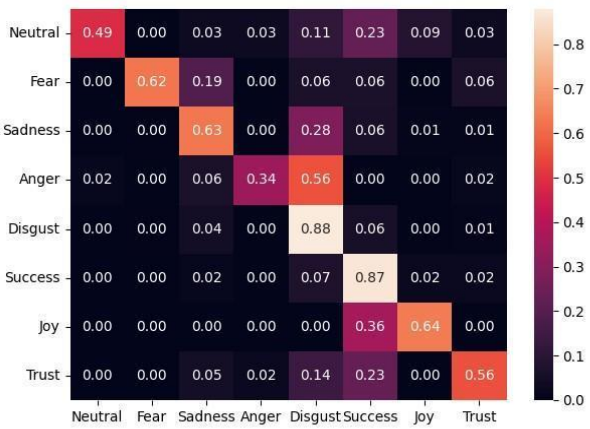

FIGURE 3. Normalized confusion matrix for emotion classification

B. MANUAL EVALUATION

Another way to better understand the trained models' typical errors is to check a random sample from sentences without the correct label. To do this, 200 randomly selected sentences were analyzed using manual validation.

We found several reasons for misclassifying sentences, both in the case of sentiment and emotion classification.

1) SENTIMENT CLASSIFICATION

Again, it is essential to note that the corpus was annotated initially for emotion categories, and the conversion to sentiment labels was done automatically afterwards.

However, it is still possible to analyze errors at the sentiment classification level since errors within the positive or negative emotion classes do not occur at the sentiment level (the conversion masks them). These are discussed in detail in the next section.

Concerning sentiment classification, some illustrative examples can be seen in Table 7, and the main reasons behind them can be summarized as follows:

- Annotation errors: Human annotation is not 100% correct because annotators' knowledge about the political context and background diversity might not be considered in the annotation process (Example 1). Given the current political context and the speaker's identity, this case is a clear example of negative sentiment.
- Use of irony: Irony, by its very nature, results in a speech situation in which the sentence is not meant in its literal sense, but in many cases, in the opposite. This is typically expressed only by prosody. Another common case is when the speaker assumes that the true meaning of the sentence can be inferred through shared contextual knowledge. Such cases can easily confuse the model, as illustrated in Examples 2-3. We observed that this was the most common source of errors.
- Absence of context: in some cases, since the context is missed, the model could not interpret the message in its entirety and predict the label correctly. During labelling, annotators could select labels for each textual unit by considering the full text of the speech. Still, this information was unavailable to the model, which could also result in errors. In the case of Example 4, the word 'auction' can be both positive and negative, depending on which side of the process it is interpreted from. With regard to Example 5, it can be seen that the employment abroad associated with increased mobility is positive in itself but is mentioned in the specific text as a consequence of emigration due to deteriorating living conditions.
- Unbalanced dataset: as the neutral sentences were significantly underrepresented compared to the others, this was also reflected in the predictions about the neutral sentences. The errors here were simply the result of misclassification, with no other specific (linguistic or annotation) cause. Example 6 illustrates a False Positive case from the neutral point of view, and Example 7 is a False Negative one. Note that only 4 examples of the former occurred in the test data.

	Example	**Predicted**	**GS**
1	Magyarországon a nok arra kellenek, hogy szüljenek, szüljék tele a világot. *(Women in Hungary are needed to give birth, to give birth to the world.)*	Negative	Positive
2	A Mészáros családnak például már remekül megy. *(The Mészáros family, for example, is already doing well.)*	Positive	Negative
3	Azt mondták, az önök végső célja az, hogy vasárnap senki ne dolgozzon. *(You said your ultimate goal is that no one should work on Sunday.)*	Positive	Negative
4	A mai napon elindultak a földárverések. *(Land auctions started today.)*	Positive	Negative
5	Szeretném újra felhívni a magyar parlament képviselőinek figyelmét, hogy a tavaly év végi adatok alapján 350 ezer magyar dolgozik az Európai Unió különböző országaiban. *(Once again, I would like to draw the attention of the Members of the Hungarian Parliament to the fact that, according to the figures at the end of last year, 350,000 Hungarians are working in different countries of the European Union.)*	Negative	Positive
6	Valószínűleg képviselőtársaim előtt is ismert, hogy a holnapi napon az Európai Parlament ismételten Magyarországgal fog foglalkozni, és ennek keretében ismételten immár a sokadik határozatot fogja Magyarországról elfogadni. *(As my fellow Members are probably aware, tomorrow the European Parliament will once again be dealing with Hungary, and will once again be adopting its umpteenth resolution on Hungary.)*	Negative	Neutral
7	Biciklik és esernyők lettek a jelképei az oktatás ügyének. *(Bicycles and umbrellas have become symbols of the cause of education.)*	Neutral	Positive

TABLE 7. Examples illustrating typical error types in sentiment classification. Predicted: label predicted by the model. GS: Gold Standard - manually annotated label.

2) EMOTION ANALYSIS

We examined a same-sized random sample of misclassified sentences for emotion analysis and investigated them manually. The cases discussed below illustrate the inherent subjectivity and context-dependence characterizing emotion categories, making their automatic identification often difficult. Many problems are overlayered with those mentioned in the sentiment classification, so only those

unique to the set selected here are reported in detail. Again, some typical examples are illustrated in Table 8.

- Conflicting meanings for a single label: the category "Trust" can often refer to both the presence and the absence of trust. This kind of contradictory nature is illustrated by the fact that, in the former case, the value of the sentiment could be positive, while in the latter case, it could be negative. The category also relies heavily on knowledge of the world and the information present in the text as a whole. In the case of Example 1, the incorrect label 'Success' was presumably calculated by the model based on the literal sentence meaning.
- Sarcasm: This is similar to ignoring contextual meaning. In Table 8 Example 2, words with meanings that conventionally express or relate to recognition are usually found up to the last tag phrase. The sentence's meaning is overwritten only by the last clause's sarcastic tone, which reinterprets the whole sentence's meaning.

	Example	Predicted	GS
1	300 millió forintos éves vagyonosodás fölött progresszív adórendszert vezetünk be. *(We introduce a progressive tax system for annual wealth above HUF 300 million.)*	Success	Trust
2	Önök nemcsak a Demokratikus Ifjúsági Világszövetségbol vagy a Miniszertanács Tájékoztatási Hivatalából ismerhetik jól, de önök 2009-ben még komoly kitüntetést adtak át Fodros István úrnak, a Magyar Köztársasági Érdemrend Lovagkeresztje kitüntetést adták át neki, minden bizonnyal a Buda-Cashben is végzett kiváló munkájának és üzleti tevékenységének az elismeréseként. *(You may not only know him well from the World Youth Democratic Alliance or the Information Office of the Council of Ministers, but you also awarded Mr István Fodros a prestigious medal in 2009, the Knight's Cross of the Order of Merit of the Republic of Hungary, certainly in recognition of his excellent work and business activities in Buda-Cash.)*	Fear	Disgust

TABLE 8. Examples illustrating typical error types in emotion classification. Predicted: label predicted by the model. GS: Gold Standard - manually annotated label.

The above illustrates the problem (often discussed in the literature) that for the automatic recognition of both sentiment and emotion values, it is often not sufficient to consider the concrete text alone. In human communication, decoding such (often subtle) nuances of meaning inherently relies on other meta-information (e.g., pragmatic context, prosody, etc.).

VI. CONCLUSION AND FUTURE WORK

In the present study, we attempted to analyze sentiment and emotion in Hungarian political texts by fine-tuning the Hungarian variant of the BERT-base model (huBERT). To our knowledge, no such model has been previously established.

Based on the results, we can conclude that the models performed at the SOTA level in sentiment analysis on the texts of the political domain and (taking into account the imbalance of the corpus) also produce acceptable results in emotion analysis. At the same time, we consider it important to note that, for example, the hyperparameters used during the fine-tuning of models are generally not considered optimal for sentiment or emotion analysis tasks. This is largely due to the highly task-dependent nature of the optimality of such parameters.

A significant proportion of the errors encountered are due to a lack of detection of linguistic phenomena that are usually identified as a separate research issue (e.g., detection of irony or sarcasm). The present research is the first step towards establishing the sentiment and emotion analysis of political texts in Hungarian literature and contributing to international examples by evaluating Hungarian data.

The models presented here are intended to be fundamental resources that support research in political science and other social sciences by extending the analytical possibilities of political texts with the dimension of sentiment and emotion analysis. In the case of Hungarian, this has not been a previously solved problem (given the different domains of the previous sentiment corpus).

In this section, we also list the possible directions for further development that can be taken to illustrate the potential for further use of the fine-tuned BERT models presented.

A. ASPECT BASED SENTIMENT- AND EMOTION ANALYSIS

Sentiment and emotion analysis can be used to investigate the general polarity of a text or sentence and the emotions it conveys. Still, it is often insufficient to obtain practically useful data. The main reason behind this is that sentences often do not express just a single sentiment or emotion but many of them. For example, there are frequent cases where two clauses referring to two properties of an object have completely different sentiment values. Such cases are, by default, difficult to deal with in classical sentiment analysis

procedures, which cannot detect if negative and positive sentiment values do not refer to the same entity or the same aspect of a given entity.

A potential solution can be using Aspect Based Sentiment Analysis (ABSA). In the most general approach, ABSA systems are designed to identify text aspects of sentiment and determine the sentiment value for each of them. An aspect can be any entity in the real world, such as personal names, companies (traditional targets of Named Entity Recognition – NER) or personal pronouns referring to them, any properties (of a product, for instance), etc. Therefore, using ABSA solutions, the main goal is to identify a sentiment value for a textual unit and find the appropriate entities to which this sentiment is connected [84], [104].

Although there are models already existing in Hungarian that can perform sentiment analysis at the aspect level [105], and aspect-level annotated sentiment corpora are also available [93], [106], we are currently not aware of any similar method that works in the political domain or can perform aspect-based emotion analysis tasks.

B. TESTING ANOTHER HUNGARIAN OR MULTILINGUAL MODELS

Although the huBERT model performs best for the sentiment analysis task in Hungarian [105], there are numerous other models trained exclusively on Hungarian (like HILBERT [107] or HIL-RoBERTa [108] and multilingual models which also included Hungarian texts in their training data (like XLM-RoBERTa base [109], or multilingual cased BERT-Base mBERT [64]) whose performance could be compared on the emotion analysis task as well.

By testing these, we can determine which architecture performs best in the context of the political emotion analysis described in this paper. By evaluating the multilingual models, we can get an idea (with appropriate foreign language test data) of how well the models trained on Hungarian data suit the emotion analysis task in other languages. Similar experiments could also aim to develop NLP tools in languages that do not currently have rich resources in this area (e.g. Czech, Polish, etc.). In the case of multilingual solutions, our main goal would be to support (subject to the availability of suitable training data) resource-poor ('small') languages by improving the results obtained from their teaching data using Hungarian data.

This will also bring us closer to whether transfer learning methods, tested in question-answering systems, for instance [110], are feasible for emotion analysis as well.

REFERENCES

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[2] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 579–586.

[3] C. Bazzanella, "Emotions, language, and context," *Emotion in dialogic interaction*, pp. 55–72, 2004.

[4] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion detection in text: a review," *arXiv preprint arXiv:1806.00674*, 2018.

[5] S. Gururangan *et al.*, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[6] T. Widmann and M. Wich, "Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text," *Polit. Anal.*, pp. 1–16, Jun. 2022, doi: 10.1017/pan.2022.15.

[7] M. Boukes, B. van de Velde, T. Araujo, and R. Vliegenthart, "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools," *Communication Methods and Measures*, vol. 14, no. 2, Art. no. 2, Apr. 2020, doi: 10.1080/19312458.2019.1671966.

[8] M. Haselmayer and M. Jenny, "Sentiment analysis of political communication: combining a dictionary approach with crowdcoding," *Qual Quant*, vol. 51, no. 6, Art. no. 6, Nov. 2017, doi: 10.1007/s11135-016-0412-4.

[9] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in *Emotion measurement*, Elsevier, 2016, pp. 201–237.

[10] T. Mullen and R. Malouf, "A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse.," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 159–162.

[11] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *FNT in Information Retrieval*, vol. 2, no. 1–2, Art. no. 1–2, 2008, doi: 10.1561/1500000011.

[12] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*, 2002.

[13] C. Rauh, "Validating a sentiment dictionary for German political language—a workbench note," *Journal of Information Technology & Politics*, vol. 15, no. 4, Art. no. 4, Oct. 2018, doi: 10.1080/19331681.2018.1485608.

[14] W. Van Atteveldt, J. Kleinnijenhuis, N. Ruigrok, and S. Schlobach, "Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations," *Journal of Information Technology & Politics*, vol. 5, no. 1, pp. 73–94, 2008.

[15] L. Young and S. Soroka, "Affective news: The automated coding of sentiment in political texts," *Political Communication*, vol. 29, no. 2, Art. no. 2, 2012.

[16] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artificial Intelligence Review*, pp. 1–41, 2021.

[17] C. Cardie and J. Wilkerson, *Text annotation for political science research*. Taylor & Francis, 2008.

[18] B. L. Monroe, M. P. Colaresi, and K. M. Quinn, "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict," *Political Analysis*, vol. 16, no. 4, pp. 372–403, 2008.

[19] J. Wilkerson and A. Casas, "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges," *Annu. Rev. Polit. Sci.*, vol. 20, no. 1, pp. 529–544, May 2017, doi: 10.1146/annurev-polisci-052615-025542.

[20] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political analysis*, vol. 21, no. 3, pp. 267–297, 2013.

[21] Á. Máté, M. Sebők, and T. Barczikay, "The effect of central bank communication on sovereign bond yields: The case of Hungary," *PLoS ONE*, vol. 16, no. 2, p. e0245515, Feb. 2021, doi: 10.1371/journal.pone.0245515.

[22] P. A. Schrodt, "Prediction of interstate conflict outcomes using a neural network," *Social Science Computer Review*, vol. 9, no. 3, pp. 359–380, 1991.

[23] P. A. Schrodt, "Artificial intelligence and international relations: An overview," in *Artificial intelligence and international politics*, V. M. Hudson, Ed., Boulder, CO: Westview Press, 1991, pp. 9–34.

[24] J. M. Montgomery, F. M. Hollenbach, and M. D. Ward, "Improving predictions using ensemble bayesian model averaging," *Political Analysis*, vol. 20, no. 3, pp. 271–291, 2012, doi: DOI: 10.1093/pan/mps002.

[25] C. Perry, "Machine learning and conflict prediction: A use case," *Stability: International Journal of Security & Development*, vol. 2, no. 3, p. 56, 2013, doi: 10.5334/sta.cr.

[26] A. Peterson and A. Spirling, "Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems," *Political Analysis*, vol. 26, no. 1, pp. 120–128, 2018, doi: DOI: 10.1017/pan.2017.39.

[27] D. Hillard, S. Purpura, and J. Wilkerson, "Computer-assisted topic classification for mixed-methods social science research," *Journal of Information Technology & Politics*, vol. 4, no. 4, pp. 31–46, 2008.

[28] L. Dun, S. Soroka, and C. Wlezien, "Dictionaries, supervised learning, and media coverage of public policy," *Political Communication*, pp. 1–19, 2020.

[29] P. Barberá *et al.*, "Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data," *American Political Science Review*, vol. 113, no. 4, pp. 883–901, 2019.

[30] M. W. Loftis and P. B. Mortensen, "Collaborating with the Machines: a hybrid method for classifying policy documents," *Policy Studies Journal*, vol. 48, no. 1, pp. 184–206, 2020.

[31] M. Sebők and Z. Kacsuk, "The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach," *Polit. Anal.*, vol. 29, no. 2, pp. 236–249, Apr. 2021, doi: 10.1017/pan.2020.27.

[32] C. Baden, C. Pipal, M. Schoonvelde, and M. A. G. van der Velden, "Three gaps in computational text analysis methods for social sciences: A research agenda," *Communication Methods and Measures*, pp. 1–18, 2021.

[33] M. Bene and F. Nábelek, "A politikai kommunikáció története a külföldi szakirodalomban," *Kiss, B.(szerk.) A szavakon túl. Politikai kommunikáció Magyarországon*, vol. 2015, pp. 11–29, 2015.

[34] G. Szabó, "Emotional communication and participation in politics," *Intersections. East European Journal of Society and Politics*, vol. 6, no. 2, 2020.

[35] A. N. Crigler and M. R. Just, "Measuring affect, emotion and mood in political communication," *The Sage Handbook of political communication*, pp. 211–224, 2012.

[36] M. Wagner and D. Morisi, "Anxiety, fear, and political decision making," in *Oxford research encyclopedia of Politics*, 2019.

[37] J. E. Settle, "Moving beyond sentiment analysis: Social media and emotions in political communication," 2018.

[38] B. Richards, "The emotional deficit in political communication," *Political Communication,* vol. 21, no. 3, pp. 339–352, 2004.

[39] J. T. Aparicio, J. S. de Sequeira, and C. J. Costa, "Emotion analysis of Portuguese political parties communication over the covid-19 pandemic," in *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, 2021, pp. 1–6.

[40] Y. Wang, S. M. Croucher, and E. Pearson, "National leaders' usage of twitter in response to COVID-19: A sentiment analysis," *Frontiers in Communication*, vol. 6, p. 732399, 2021.

[41] S. R. Rufai and C. Bunce, "World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis," *Journal of public health*, vol. 42, no. 3, pp. 510–516, 2020.

[42] B. Liu, "Sentiment analysis and subjectivity.," *Handbook of natural language processing*, vol. 2, no. 2010, pp. 627–666, 2010.

[43] M. Amsler, "Using lexical-semantic concepts for fine-grained classification in the embedding space," University of Zurich, 2020.

[44] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 625–631.

[45] C. Dhaoui, C. M. Webster, and L. P. Tan, "Social media sentiment analysis: lexicon versus machine learning," *Journal of Consumer Marketing*, vol. 34, no. 6, pp. 480–488, 2017.

[46] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," *Procedia Computer Science*, vol. 161, pp. 707–714, 2019.

[47] O. Kolchyna, T. T. Souza, P. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," *arXiv preprint arXiv:1507.00955*, 2015.

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

[49] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[50] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[51] I. Chetviorkin and N. Loukachevitch, "Evaluating sentiment analysis systems in Russian," in *Proceedings of the 4th biennial international workshop on Balto-Slavic natural language processing*, 2013, pp. 12–17.

[52] O. Y. Koltsova, S. Alexeeva, and S. Kolcov, "An opinion word lexicon and a training dataset for Russian sentiment analysis of social media," *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE*, vol. 2016, pp. 277–287, 2016.

[53] M. M. Itani, R. N. Zantout, L. Hamandi, and I. Elkabani, "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)*, IEEE, 2012, pp. 192–197.

[54] M. Dilai, Y. Onukevych, and I. Dilay, "Sentiment analysis of the US and Ukrainian presidential speeches," *Computational linguistics and intelligent systems (2), 2018*, pp. 60–70, 2018.

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

[56] S. Huang, Z. Niu, and C. Shi, "Automatic construction of domain-specific sentiment lexicon based on constrained label propagation," *Knowledge-Based Systems*, vol. 56, pp. 191–200, 2014.

[57] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "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*, NIH Public Access, 2016, p. 595.

[58] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, pp. 41–46.

[59] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and Bernoulli naïve Bayes for text classification," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, IEEE, 2019, pp. 593–596.

[60] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[61] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, C. Nédellec and C. Rouveirol, Eds., in Lecture Notes in Computer Science, vol. 1398. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142. doi: 10.1007/BFb0026683.

[62] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

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

[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers)*, Minneapolis, Minnesota: Association for

Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[65] L. Young and S. Soroka, "Affective news: The automated coding of sentiment in political texts," *Political Communication*, vol. 29, no. 2, pp. 205–231, 2012.

[66] S. Aday, "Chasing the Bad News: An Analysis of 2005 Iraq and Afghanistan War Coverage on NBC and Fox News Channel," *Journal of Communication*, vol. 60, no. 1, pp. 144–164, Mar. 2010, doi: 10.1111/j.1460-2466.2009.01472.x.

[67] J. W. Boumans and D. Trilling, "Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars," *Rethinking Research Methods in an Age of Digital Journalism*, pp. 8–23, 2018.

[68] J. Wilkerson and A. Casas, "Large-scale computerized text analysis in political science: Opportunities and challenges," *Annual Review of Political Science*, vol. 20, pp. 529–544, 2017.

[69] W. Van Atteveldt, M. A. Van der Velden, and M. Boukes, "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms," *Communication Methods and Measures*, vol. 15, no. 2, pp. 121–140, 2021.

[70] S. Singh and A. Mahmood, "The NLP cookbook: modern recipes for transformer based deep learning architectures," *IEEE Access*, vol. 9, pp. 68675–68702, 2021.

[71] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[72] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.

[73] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[74] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[75] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[76] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[77] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[78] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[79] M. E. Peters *et al.*, "Deep contextualized word representations. arXiv 2018," *arXiv preprint arXiv:1802.05365*, vol. 12, 2018.

[80] B. Huang, Y. Ou, and K. M. Carley, "Aspect level sentiment classification with attention-over-attention neural networks," in *Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings 11*, Springer, 2018, pp. 197–206.

[81] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-Dependent Sentiment Classification With BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: 10.1109/ACCESS.2019.2946594.

[82] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-Based Sentiment Analysis using BERT," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland: Linköping University Electronic Press, Sep. 2019, pp. 187–196. [Online]. Available: https://aclanthology.org/W19-6120

[83] A. Karimi, L. Rossi, and A. Prati, "Adversarial Training for Aspect-Based Sentiment Analysis with BERT," *arXiv:2001.11316 [cs, stat]*, Oct. 2020, Accessed: Apr. 22, 2021. [Online]. Available: http://arxiv.org/abs/2001.11316

[84] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," 2022, doi: 10.48550/ARXIV.2203.01054.

[85] Y.-H. Huang, S.-R. Lee, M.-Y. Ma, Y.-H. Chen, Y.-W. Yu, and Y.-S. Chen, "EmotionX-IDEA: Emotion BERT -- an Affectional Model for Conversation." arXiv, Aug. 17, 2019. Accessed: May 11, 2023. [Online]. Available: http://arxiv.org/abs/1908.06264

[86] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 33, 2021.

[87] R. Chandra and R. Saini, "Biden vs trump: modeling us general elections using bert language model," *IEEE Access*, vol. 9, pp. 128494–128505, 2021.

[88] M. S. M. Chowdhury and B. Pal, "BERT-Based Emotion Classification Approach with Analysis of COVID-19 Pandemic Tweets," in *Applied Informatics for Industry 4.0*, Chapman and Hall/CRC, 2023, pp. 109–121.

[89] A. Surolia, S. Mehta, and P. Kumaraguru, "Understanding Emotions: A PoliEMO Datasetand Multi-label Classification in Indian Elections," 2023.

[90] D. Croce, G. Castellucci, and R. Basili, "GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online:

Association for Computational Linguistics, 2020, pp. 2114–2119. doi: 10.18653/v1/2020.acl-main.191.

[91] P. K. Jain, W. Quamer, and R. Pamula, "Consumer sentiment analysis with aspect fusion and GAN-BERT aided adversarial learning," *Expert Systems*, vol. 40, no. 4, p. e13247, May 2023, doi: 10.1111/exsy.13247.

[92] O. Ring, V. Vincze, C. Guba, and I. Üveges, "HunEmPoli: magyar nyelvű, részletesen annotált emóciókorpusz," 2023.

[93] M. K. Szabó, V. Vincze, and G. Morvay, "Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái," *Távlatok a mai magyar alkalmazott nyelvészetben*, p. 282, 2016.

[94] V. Vincze, I. Üveges, M. K. Szabó, and K. Takács, "A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata," 2021.

[95] G. Orosz, Z. Szántó, P. Berkecz, G. Szabó, and R. Farkas, "HuSpaCy: an industrial-strength Hungarian natural language processing toolkit," *arXiv preprint arXiv:2201.01956*, 2022.

[96] D. M. Nemeskey, "Introducing huBERT." [Online]. Available: https://hlt.bme.hu/media/pdf/huBERT.pdf

[97] D. M. Nemeskey, "Natural Language Processing Methods for Language Modeling," 2020. [Online]. Available: https://hlt.bme.hu/media/pdf/nemeskey_thesis.pdf

[98] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[99] V. Vapnik, *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[100] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Annals of Data Science*, pp. 1–26, 2020.

[101] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[102] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning.," *Journal of personality and social psychology*, vol. 66, no. 2, p. 310, 1994.

[103] G. M. Kis, O. Ring, and M. Sebők, "emBERT, a Hungarian emotion classification language model." 2022. [Online]. Available: https://huggingface.co/poltextlab/emBERT

[104] K. Liu, L. Xu, and J. Zhao, "Opinion target extraction using word-based translation model," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012, pp. 1346–1356.

[105] G. Yang Zijian and L. J. Laki, "Neurális entitásorientált szentimentelemző alkalmazás magyar nyelvre," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged: Szegedi Tudományegyetem, 2023.

[106] M. Miháltz, "OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez.," in *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013)*, Szeged, 2013, pp. 343–345.

[107] Á. Feldmann *et al.*, "HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben," 2021.

[108] Z. G. Yang and V. Tamáás, "Training language models with low resources: RoBERTa, BART and ELECTRA experimental models for Hungarian," in *Proceedings of 12th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2021)*, 2021, pp. 279–285.

[109] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[110] M. Bornea, L. Pan, S. Rosenthal, R. Florian, and A. Sil, "Multilingual transfer learning for QA using translation as data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 12583–12591.

ISTVÁN, ÜVEGES was born in Orosháza, Hungary in 1989. He received the B.A. and M.A. degrees in linguistics from the University of Szeged, Hungary and the B.S. degree in computer science also from University of Szeged, Hungary. He is a Ph.D. candidate in computational linguistics at University of Szeged, Hungary.

From 2021 to 2023, he was a Researcher at the Centre for Social Sciences' project of Artificial Intelligence National Laboratory (TK-MILAB) dealing with sentiment analysis of political texts, and from 2023 he is a Researcher at POLTEXTLAB Centre for Social Sciences, Budapest, Hungary). He also works as a Computational Linguist Researcher at MONTANA Knowledge Management Ltd., Budapest, Hungary from 2021. He is the author of 29 articles. His research interests include sentiment- and emotion analysis, Plain Language and automation, computational semantics and legal language.

ORSOLYA, RING was born in Budapest, Hungary, in 1975. She received her PhD in History from the ELTE University of Budapest in 2011. She is a Research Fellow at the Institute for Political Science, Centre of Social Sciences, Budapest, and a senior researcher at the Political and Legal Text Mining & Artificial Intelligence Laboratory. She is the author of 70 articles. Her research interests include social history, critical discourse analysis, and the application of text mining and machine learning methods in these fields.