SN Computer Science (2025) 6:13
https://doi.org/10.1007/s42979-024-03553-2

ORIGINAL RESEARCH

Boosting Classification Reliability of NLP Transformer Models in the Long Run—Challenges of Time in Opinion Prediction Regarding COVID-19 Vaccine

Zoltán Kmetty[1,2] · Bence Kollányi[1] · Krisztián Boros[1]

Received: 20 October 2023 / Accepted: 19 November 2024

Abstract

Transformer-based machine learning models have become an essential tool for many natural language processing (NLP) tasks since the introduction of the method. A common objective of these projects is to classify text data. Classification models are often extended to a different topic and/or time period. In these situations, deciding how long a classification is suitable for and when it is worth re-training our model is difficult. This paper compares different approaches to fine-tune a BERT model for a long-running classification task. We use data from different periods to fine-tune our original BERT model, and we also measure how a second round of annotation could boost the classification quality. Our corpus contains over 8 million comments on COVID-19 vaccination in Hungary posted between September 2020 and December 2021. Our results show that the best solution is using all available unlabeled comments to fine-tune a model. It is not advisable to focus only on comments containing words that our model has not encountered before; a more efficient solution is randomly sample comments from the new period. Fine-tuning does not prevent the model from losing performance but merely slows it down. In a rapidly changing linguistic environment, it is not possible to maintain model performance without regularly annotating new texts.

Keywords Comment classification · BERT · Fine-tunning · Temporal analysis · Concept drift · Vaccination

Abbreviations

NLP	Natural Language Processing
OOD	Out-of-distribution
NER	Named Entity Recognition
SVM	Support Vector Machine
WHO	World Health Organisation

Introduction

Temporality is important in machine learning, especially in areas such as time series analysis, dynamic systems and sequential decision-making processes. Many real-world phenomena change over time, and the order in which the data points occur affects their relationships. In stock market, sales or weather forecasts, for example, patterns over time are crucial for accurate predictions. Another well-known example is Google Flu Trends, which was able to accurately predict flu outbreaks—compared to the US Centers for Disease Control and Prevention, the results were 97% accurate [1]. The performance of GFT deteriorated significantly over time, and as David Lazer and colleagues [2] noted, the metrics were not stable over time due to both internal and external factors.

In the real-world application of machine learning models, the environment or underlying data on which the original model was trained changes, a phenomenon often referred to as model drift or concept drift [3]. These concept drifts can occur gradually or rapidly, e.g. the introduction of a new policy measure can lead to a sudden drift, and data distribution can also fluctuate over time, particularly in cases like seasonal trends [3]. To mitigate the effects of such concept changes, researchers in machine learning have developed models capable adapting to evolving data over time [4, 5]. According to a systematic review of concept drift, this phenomenon may affect various machine learning tasks, including prediction, classification, recognition, clustering and item set mining [6]—leading to challenges in many

✉ Zoltán Kmetty
kmetty.zoltan@tk.hu

[1] Hungarian Research Centre, Centre for Social Sciences, CSS-RECENS Research Group, Tóth Kálmán U. 4., 1097 Budapest, Hungary

[2] Faculty of Social Sciences, Sociology Department, Eötvös Loránd University, Budapest, Hungary

Published online: 19 December 2024

application areas, from fraud detection to product or service recommendations to fake news detection [7–9].

In a "simple" classification task, both the training and the data to be classified come from the same period and data set. In practice, however, it may often be the case that a particular classification is extended to a different set of data and/or to a different period. If we expect a concept drift between the training period and the application period, we need to rethink the initial classifiers. The need and the possibility to extend classification over time is strongly supported by increasing digitization as updated datasets are more frequently available for industrial and scientific research purposes.

Concept drift may be particularly relevant in cases where the domain under study changes rapidly over time. In the various natural language processing (NLP) projects, this problem is common, as language and language use can change within a short period of time on a given topic [10]. Furthermore, neural network-based black-box models complicate the problem, as they offer little insight into how a particular classification model works, making it harder to identify what changes in content or context might cause the model to break down. This black-box feature is a problem with the transformer-based NLP models currently used for classification tasks in data-mining projects.

Transformer-based machine learning models have become an important tool for many NLP tasks since the introduction of the method [11]. Connecting the encoder and decoder through an attention mechanism (which can capture the meaning of an input sequence, e.g., a sentence) and effective parallelization of the training process led to robust generalized language models. Transformer models are usually pre-trained on a generic language corpus and can later be tuned to domain-specific tasks. These models can be used for a wide array of general NLP tasks, from classification problems to document summarization, sentiment analysis, and named entity recognition.

When a model is trained and fine-tuned on a range of training materials, it is often difficult to apply it to another field [12]. A growing body of literature suggests that domain shifts affect the performance of general models. Performance can significantly drop when a model is applied to a dataset different from the data used for pre-training, often referred to as an out-of-distribution (OOD) dataset [13–16].

Language models trained on general texts are often insufficient for tasks related to special domains, such as medical texts or legal documents. The performance of Bidirectional Encoder Representations from Transformers (BERT) can also degrade when the general model is applied to a specific domain, so further fine-tuning is often required for specific NLP tasks. BERT, one of the most popular transformer-based models, was originally pre-trained on the complete Wikipedia corpus and books digitized by the Google Books project [17]. After BERT was released as open source in 2019, researchers focusing on specific domains created their own specialized versions of BERT by extending the base model and pre-training their models on domain-specific text. For example, BioBERT was trained on biomedical corpora (PubMed abstracts and full-text articles) to learn the biomedical language [18], SciBERT was pre-trained on a random sample of 1.14 million scientific (mainly computer science and biomedical) articles [19], while FinBERT was pre-trained on a large financial corpus consisting of more than 46,000 news articles with 29 million words [20]. These domain-specific models can be fine-tuned to downstream NLP tasks with great success and significantly improve the performance of BERT on these domain-specific tasks. Using a large domain-specific corpus and pre-training the model with it is often difficult and expensive. Another approach is to pre-train the model using the training classification dataset, which is often a much smaller corpus [20].

In addition to the difficulties of applying a language model to an OOD dataset (a domain distinct from the materials used to train the model), the performance of trained language models may also degrade over time. Measuring the robustness of such a model over time has recently attracted considerable research attention. Measuring the performance of transfer models is challenging because transformer models tend to overestimate their performance when the training and evaluation periods overlap [21]. The literature suggests temporal misalignment degrades performance when the testing period is temporally distant from the training period [22, 23].

Previous work also examined the changes in the language itself over time. The pace of language change has become increasingly rapid in the context of social media [24]. Linguistic shifts can also affect the performance of classification tasks based on machine learning. Both changes in word usage and altering the context of the same words can lead to performance degradation [25]. Temporal shifts have been studied in the context of several NLP tasks, including domain classification [26, 27], named entity recognition (NER) [28, 29], and sentiment analysis [30].

Florio et al. [31] evaluated the temporal robustness of an Italian version of the model BERT and found that the model's efficiency is highly time-dependent. Although the Italian BERT model was trained on social media data and it was fine-tuned on a labeled dataset on hate speech, the content of the texts used for testing changed rapidly. When the temporal distance between the training set and the test set increased, the model's precision and F1 values decreased significantly. The authors concluded that the language of hateful content is closely linked to current events in the news that trigger hate speech and is therefore highly sensitive to changes over time.

Florio et al. [31] also found that training the model repeatedly with new data significantly helps the model to maintain

its performance over time. The authors compared two different solutions to this problem. First, they used a sliding window approach in which the model was fine-tuned on a data set that was collected one month prior to the test data set. The other solution used an incremental approach, meaning that the model was trained with data from all available data before the test period so that after a longer period, data from several months were used. BERT performed significantly better in their experiment than a Support Vector Machine (SVM) model. It proved to be more robust and less sensitive to changes over time, even when fine-tuned on a static dataset and not adjusted by using more recent labeled data.

In this paper, we answer two research questions and five hypotheses. On the one hand, we analyze the best approaches to fine-tune a BERT model on a long-running classification task regarding the temporal selection of fine-tuning data (RQ1). We formulate four hypothesis here:

H1.1 Fine-tuned classification can achieve better results than base models.
H1.2 The best fine-tuned approach is to use unlabeled data from the time period to be classified.
H1.3 If we need to predict an outcome after a concept drift, with training data before the concept drift, than adding unlabeled data from the period after concept drift to the fine-tuning could boost the classification performance.
H1.4 In the case of fine-tuning, it may be worthwhile to combine the unlabeled data of the whole period (shallow learning, with minimal epoch number) with the unlabeled data of the specific period (deep learning, with many epoch).

We also test a particular re-classification approach where we use comments selected by their temporal uniqueness (RQ2). We test the following hypothesis related to RQ2:

H2.1 In the case of new annotations, it is worth over-representing those comments that are unique comments compared to the previous annotation phase. This annotation approach leads to more accurate classification models.

Data

This paper uses comments about COVID-19-related vaccination in Hungary collected between September 2020 and December 2021. We used a social listening platform called SentiOne to collect the data. SentiOne (www.sentione.com) is an international social listening software, a content-based web analytics platform that covers and recognizes 30 languages across Europe. It collects, indexes, and analyses online public content from social media platforms, blogs, forums, and websites. We set the start date to September 2020, as this is when the debate around Covid-19 vaccines started in Hungary. For creating our corpus, we used keywords related to vaccination. Then, we filtered out stop-words and used topic-modeling approaches to find the relevant content. There were several articles that were not related to the topic in the initial corpus, for example, articles related to animal vaccination. We retained only those articles and posts that relate to Covid-19 vaccination. The cleaned corpus consisted of about 295,152 articles and posts on social media. In this paper, we do not use the articles themselves, just the comments under the articles and relevant social media posts. At the comment level, we retained those comments written under the selected articles or containing vaccination or vaccine-related keywords.

Given the topic's sensitive nature, we used a strict data management protocol. All these comments are posted on public pages, so we did not include any private content. In order to preserve the anonymity of the commenting people, we did not store their names. We also stored the database on a secure server of our Research Centre with limited access by people outside of the core research team.

One of our initial project goals was to find the anti-vaccination content in the comments. To this end, we developed a transformer-based classifier in early 2021. As the first step, the researchers in the project coded a random sample of 1,000 comments to anti-vax and non-anti-vax or not related to vaccination category. Based on this initial classification, we defined a set of keywords that covered well (over 98 percent recall) the anti-vax comments (see the supplementary for the list of keywords). In this step, we wanted to remove those comments from the classification where we did not expect to find any anti-vax content. This method filtered out 27.5 percent of the comments.

After filtering the comments with the keywords, we randomly selected 10,000 comments from the pool of 1.7 million. The selected comments covered the period from September 2020 to February 7, 2021. We double-annotated these comments to anti-vaxxer comments and pro-vaxxer or neutral comments. In case of disagreement between annotators, a supervisor selected the final category. The supplementary contains the details of the annotation process. The kappa value for the inter-coder agreement was 0.73.

These manually classified 10,000 comments formed the basis for categorizing the rest of the comment pool. For the comment categorization, we used a transformer-based NLP model based on the Hungarian version of BERT (huBERT) [32]. The average macro-precision of the classification was 0.79, and the recall and F1 scores were 0.78. This model was not fine-tuned to the topic. We call this model the base model in this paper.

We continued to collect COVID-related content in the project daily and classified the new comments using the above-described transformer model. The first wave of the

COVID-19 pandemic in Hungary subsided in the summer of 2021, but infection rates started to rise again around August. This was the start of the second Covid-19 wave in Hungary. In this wave, more people, including many young people, were infected. Regular monitoring of vaccine-related comments showed that the discourse on vaccination also changed with the rise of the second wave, so we decided to run a second round of classifications to check whether our BERT classifier could still classify the comments with high validity. We started the second round of annotation in December 2021. We wanted to annotate content from the closest pool as possible. So, we decided to select comments from November 2021. To overcome the unbalanced ratio of anti-vax and non-anti-vax comments, we increased the weighting of the anti-vax comments in this annotation round. Before annotation, we classified the comments using our BERT classifier and selected 3,000 anti-vax and 1,500 pro-vaxxer or neutral comments. The 4,500 comments were annotated in the same way as in the first round of annotation. Two annotators annotated all the comments. The annotators did not know how the BERT classifier classified the comments. If there was a difference between the annotators, a supervisor selected the final category. The kappa value for the inter-coder agreement was 0.63. In this paper, we use some of these annotated comments from the second round of annotation to test the quality of the different classification approaches (see the details in the method section).

The final corpus contains around 5.5 million comments from the selected period.

Methods

Our primary research question and the related hypothesis (RQ1, H1.1-H1.4) were how we could optimize the classification of vaccination-related comments in the long run. We use data from different periods in the first set of tests to fine-tune the original BERT model. The classification logic was the following. We started with the original huBERT model; then, we fine-tuned the model with non-labeled data. After the fine-tuning, we add the labeled data from the first annotation round to the model to learn the classification. The quality of this classification was measured on two bases. We leave out 30 percent of the comments from the labeled training set for test purposes. These comments cover the period from September 2020 to January 2021. The second quality check was on the test comments selected from the second annotation round. These comments are coming from November 2021. For the fine-tuning, we used all available non-labeled data from three different periods:

- 09/2020–01/2021—1.3 m comments
- 10/2021–11/2021—777 k comments
- 09/2020–12/2021—5.5 m comments

For the first fine-tuning period, we selected the exact dates that the first annotation round covers. In the second annotation round, we selected comments from November 2021. For the fine-tuning data, we decided to include not only this November month but also October. We extended the November time frame to get a bigger pool of comments for fine-tuning.

To test H1.4, we used a special approach. We used unlabeled data from the entire period (09/2020–12/2021) for fine-tuning, but we only allowed the model to specify the weights for two epochs; then, we used these model weights to do further fine-tuning with data from October and November 2021. Here, we wanted to test whether it is worth combining shallow and deep learning techniques to maximize classification quality.

In addition to testing the different fine-tuning approaches, we also measured how particular annotation approaches could boost the classification quality (RQ2, H2.1). For this, we tested an optimization approach for selecting second-round annotated comments. To do this, we computed a weirdness statistic for all words in the comments. Weirdness is a simple statistic on a word level, which shows the ratio of a given word between two corpora or two time-periods [33]. In our case, we calculated the baseline ratio of a word for the period between September 2020 and January 2021. This is a simple division of the frequency of the word and the total counts of all words in the selected period. We then calculated a weirdness statistic for words on a monthly basis. We calculated the ratio of one word compared to all words in the selected month and divided this number by the baseline ratio. This weirdness number is one if the occurrence of a word in a month is the same as in the baseline period, and it rises above one if the words appear more often in that month. To compute the weirdness statistics, we used a sample of comments. We selected 40,000 comments each month and lemmatized the texts with a Hungarian lemmatizer e-magyar [34]. We used the lemmatized comments for the calculation.

Figure 1A presents the weirdness value of the word "Delta" and Fig. 1B the Google searches for this word. WHO began using the delta-variant name in June 2021. The weirdness statistic for this word goes parallel with the number of Google searches, which started to increase in the same month and peaked in August 2021.

Weirdness is a perfect statistic to measure the change in our corpus compared to a previous period. Figure 1C shows the change in the standard deviation of the weirdness statistic. A standard deviation of zero would mean that the word ratio in the observed month is the same as in the baseline period. As expected, the standard deviation of the weirdness statistic was the lowest in February 2021, the month

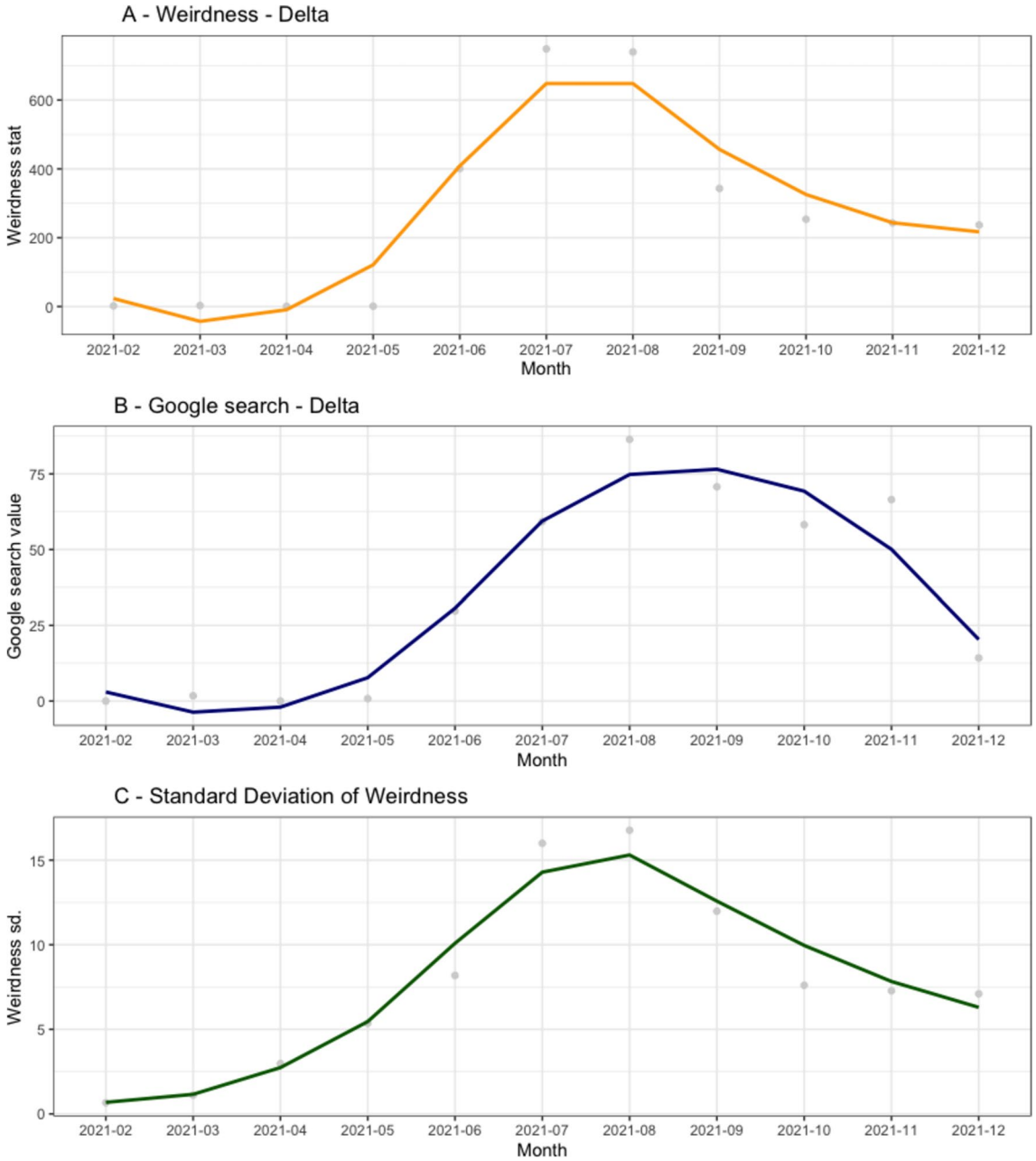

Fig. 1 **A** Weirdness value for the word Delta. **B** Google searches in Hungary for the word Delta. **C** The monthly standard deviation of the weirdness statistics

closest to the baseline period. The standard deviation value increased until August 2021, and then decreased. Seasonality might explain this decline in word usage.

We calculated a comment level weirdness based on the word level weirdness statistic. It is a simple average of the word's monthly weirdness value per comment. We use this comment level weirdness value to select a particular sample from the second round of annotated contents by using weirdness statistics as a weight in the sampling process. We selected 1500 comments based on this approach and 1500 comments randomly from the pool of secondly annotated comments. There was some overlap (693 comments) between the weirdness based 1500 comments and the randomly selected 1500 comments. Both samples were selected from the second annotation round of annotation, covering November 2021. The 2193 comments from the second annotation round used for test purposes did not overlap with these samples. We used the two samples of annotated comments to boost the classification. We added these comments to the original training dataset and trained the models with these

extended training sets. In these tests, we used all the available unlabeled data for fine-tuning.

Results

First Annotation Period (September 2020 and January 2021)

We first present the model performance for the first period of annotation. Of the 10,000 comments annotated in the first round, we withheld 30 percent for testing and trained the models on the remaining comments. This round of annotation covered the period between September 2020 and January 2021. Figure 2. Summarizes the results. The base huBERT model, without any fine-tuning, achieved an F1-score of 0.65 in the anti-vaxxer category and 0.92 for the non-anti-vaxxer and neutral comments. As expected, fine-tuning boosted the performance of our transformer model. This result confirmed our H1.1 hypothesis. We obtained an F1-score of 0.73 for the anti-vaxxer comments when we used the complete unlabeled data for fine-tuning. We obtained the same value when we fine-tuned with data selected from the same period as the test comments, namely September 2020 to January 2021. As expected, the other models optimized for the second annotation period achieved weaker performance in this case. The model with an extra 1,500 randomly selected comments from the second annotation

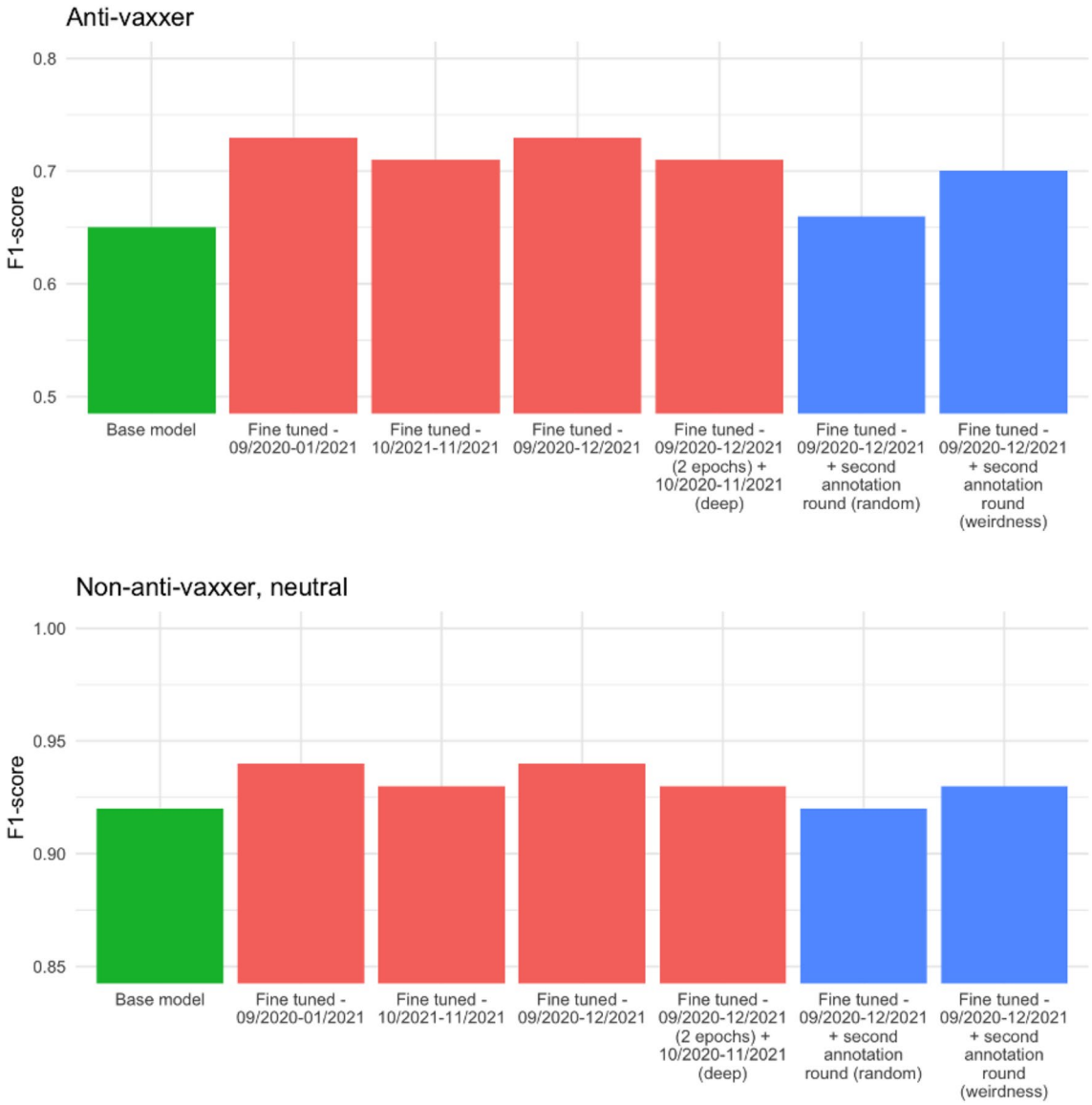

Fig. 2 F1-score for models tested in the first period of annotation

achieved almost the same low F1-score as the base model without fine-tuning.

Second Annotation Period (November 2021)

Prior to the detailed analysis of classification performance, we examined how confident the models in the prediction were in the case of the test dataset coming from the second annotation period. To do this, we calculated a simple confidence value which was the absolute value of 0.5 minus the probability of anti-vaxxer classification given by the BERT model. The minimum value of this statistic is zero, and the maximum value is 0.5. A higher value means higher model confidence. We calculated this value for the test data and a subset of the test data with lower and higher weirdness values (see Table 1). As expected for all the models, the confidence value was low for comments with high weirdness and high for comments with low weirdness scores. Our base model had the lowest confidence value. The fine-tuned model had a higher confidence value, and the confidence value of the fine-tuned models and the models with the boosted classification samples did not differ significantly. Without knowing the true classification results based on model confidence, we would expect the same performance for the fine-tuned and the extended models. As we will see in the next section, the results did not confirm this expectation.

Figure 3 presents the results of the comments used to test the classification quality for November 2021. There were 2193 comments in the test database, none of which were used for fine-tuning or model classification. As expected, the overall quality of classification for these comments was much lower. The base model using annotated comments from the first period without fine-tuning only achieved an F1-score of 0.5 for anti-vaxxers and 0.66 for the rest of the comments. Fine-tuning without new training data did not improve the model quality for non-anti-vaxxers and neutral comments; in some cases, it even decreased it. For the anti-vaxxer category, fine-tuning boosted the F1 score. Time specific unlabeled data did not improve the model performance compared to unlabeled data from the full period. We got the same results regarding the first annotation round. We got the same results regarding the first annotation round. So, we can reject the H1.2 hypotheses. It seems that it is worth to use all available data for fine-tuning. We can also reject the H1.3 hypothesis—fine-tuning with unlabeled data selected after the concept drift period could not help us avoid the decrease in model performance measures.

The model with the full unlabeled data and the model with the combination of swallow and deep learning had the same level of performance (Fig. 3). Based on this result we can also reject H1.4—combination of swallow and deep learning did not increase the model's performance. The last two sets of models used classified comments from the time period of the test data. The performance of these two models was much higher than the earlier ones, so it seems that reclassification is inevitable for a changing corpus.

The random sample from the annotated comments performed better based on the test data, achieving higher F1-scores both for the anti-vaxxer and the other class, so the special sampling design for the training dataset did not help the overall classification. Based on this result we can reject H2.1 hypothesis.

In the next step, we checked how the classification model worked for the comments where the weirdness value was above 1.2 (see Fig. 3.). About 10 percent of the annotated comments were above this threshold. The test data contained 223 comments that were above the 1.2 weirdness score. As expected for most models, the classification of these comments was more challenging than that of the "normal" comments; the model performance decreased sharply. There was only one exception, namely the extended model with the random sample from the second annotation. In contrast, the extended model with the sample based on weirdness probability had a weaker F1-score for the anti-vaxxers.

In order to evaluate the global performance of the models, we contrasted the results of three studies that attempted a similarly difficult classification task. In the study by Florio et al. [31], cited in the literature, tweets containing hate speech were classified using an Italian BERT model, testing different fine-tuning solutions. Gozuacik et al. [35] performed sentiment and opinion prediction with different models. The Word2Vec models with the best results in their

Table 1 Confidence level of models

	Weird < 0.9	All test data	Weird > 1.2
Base model	0.39	0.18	0.16
Fine tuned—09/2020–01/2021	0.44	0.28	0.25
Fine tuned—10/2021–11/2021	0.42	0.27	0.26
Fine tuned—09/2020–12/2021	0.43	0.31	0.27
Fine tuned—09/2020–12/2021 (2 epochs) + 10/2020–11/2021 (deep)	0.44	0.28	0.26
Fine tuned—09/2020–12/2021 + second annotation round (random)	0.41	0.27	0.26
Fine tuned—09/2020–12/2021 + second annotation round (weirdness)	0.44	0.26	0.24

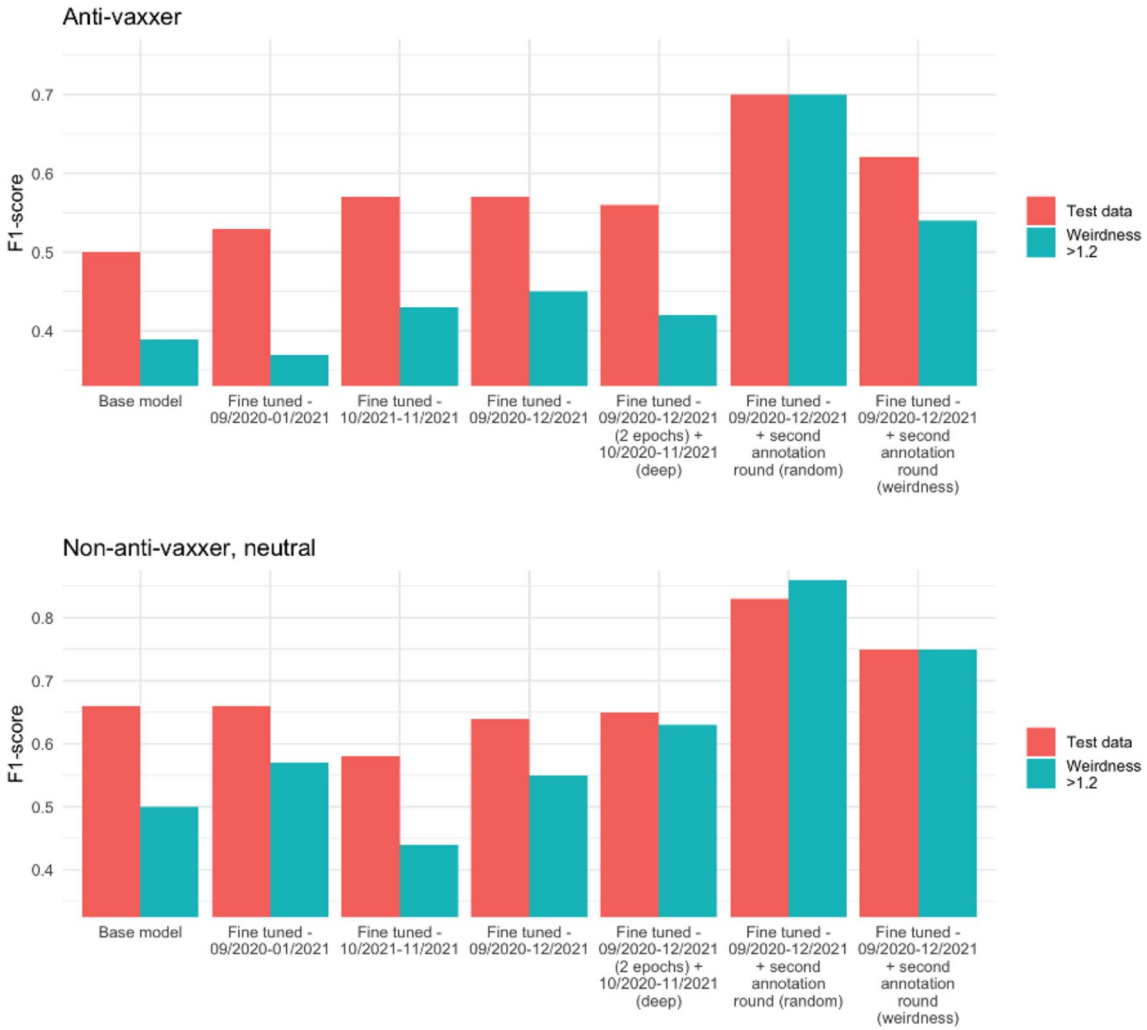

Fig. 3 F1-score for models tested in the second period of annotation

comparison were included in the comparison. In the study by Fahfouh et al. [36], the focus was on Deceptive Opinion detection. They used, among others, contextual word embedding models for prediction and complex neural network-based models developed explicitly for Deceptive Opinion detection. The Contextual Relationship Model (CRM) was highlighted as the best-performing of the latter.

The macro F1 values were used to compare the performances (see Table 2). The base model run in the first coding period achieved a very high F1 value. Of the models compared, only CRM performed better, but the latter was a solution developed for a specific problem, not a general model like BERT. The fine-tuned models of the first period scored above 0.83, which is even more outstanding. As seen from the results, the second period proved to be much more challenging to classify, which was also reflected in the performance of the models, with an F1 value of around 0.6 being a distinctly low value for a fine-tuned model. Including the extra coding for that period in the model brought the average F1 value up to 0.76, which is a good value by comparison.

Discussion

The rate of vaccinated people was a critical factor in preventing the further spreading of the Covid-19 virus. The acceptance or rejection of vaccination has been the subject of considerable social debate in Hungary, and conspiracy theories about vaccination have flourished [37]. Given the sensitivity of the issue, it was of paramount importance to be able to measure as accurately as possible in real time the extent of vaccine resistance. Exaggerating or covering up the issue would have had negative consequences, both from a policy and public discourse perspective. Our methodological investigation was primarily motivated by these considerations.

Table 2 Model comparison

Author	Domain	Specification	Model	F1 Score
K. Florio et al. 2020 [23]	Hate speech	Fixed	AIBERTo	0,628
K. Florio et al. 2020 [23]	Hate speech	Sliding Window	AIBERTo	0,679
K. Florio et al. 2020 [23]	Hate speech	Incremental	AIBERTo	0,688
N. Gozuacik et al. 2021 [27]	Sentiment prediction		Word2Vec	0,632
N. Gozuacik et al. 2021[27]	Opinion detection		Word2Vec	0,677
A. Fahfouh et al. 2022 [28]	Deceptive Opinion detection	Mixed Domain	BERT	0,764
A. Fahfouh et al. 2022 [28]	Deceptive Opinion detection	Mixed Domain	T5	0,737
A. Fahfouh et al. 2022 [28]	Deceptive Opinion detection	Mixed Domain	CRM	0,876
Our model	Anti-vax categorization	Base model	huBERT	0,785
Our model	Anti-vax categorization	First annotation round—Fine tuned—09/2020–01/2021	huBERT	0,835
Our model	Anti-vax categorization	First annotation round—Fine tuned—09/2020–12/2021	huBERT	0,835
Our model	Anti-vax categorization	Second annotation round—Fine tuned—09/2020–12/2021	huBERT	0,605
Our model	Anti-vax categorization	Second annotation round—Fine tuned—09/2020–12/2021 + second annotation round (random)	huBERT	0,765

Our results confirm the general result found in the literature on fine-tuning. Domain-based fine-tuning of transformer models can significantly improve their performance. Considering the temporal aspect of fine-tuning, the results clearly show that the best solution is to use all available unlabeled annotations in tuning the models. Presumably, the more/better principle that is generally observed in larger language models applies here. This result partly contradicts previous recommendations in the literature [31].

However, fine-tuning does not protect against model performance loss, but at least slows down the performance loss. In a rapidly changing linguistic environment, it is not possible to maintain model performance without adding new annotations. The timing of re-annotation is not generalizable, but indicators such as the weirdness index can help to shed light on corpus change dynamics. An important result of our analysis is that it is not advisable to select those comments that were less likely encountered by our model previously for new comment annotations. A more efficient solution is to randomly select the comments from the new period. The random sampling approach worked better than the specific weirdness-based sampling approach, even for classifying comments with high weirdness values. In the latter case, we saw that the model learned much worse on specific comments than on a more general training set.

It is also a critical result that it is difficult to say anything about the performance of the models without repeated annotation. The confidence of the BERT models we tested was similar for both the fine-tuned and the extended models, although the latter performed better in reality. This model confidence can easily mislead researchers when no re-annotated data is available.

When evaluating the results, it is important to stress that classifying comments as anti-vaxxers is a complex task. The task's difficulty is shown by the fact that annotators also struggled with the classification. In the first round of annotation, the Kappa value was 0.73, while in the second round, it was 0.63. Although the latter value is not particularly low, it is far from a perfect match. One of the main challenges in the classification was that the comments had to be classified without context. For example, detecting sarcastic content is particularly difficult in such cases. Members of the research team carried out the first round of annotation. We expect a high-quality classification from a team working on a topic, as they know the study and the concepts behind the study very well.

External annotators carried out the second round of annotation. Although we provided them with training and they had to do test annotation, we assume that in their case, the uncertainty factor was higher. This uncertainty level may explain the lower kappa value of the second round annotation. However, we should also consider that deciding which comments were anti-vaccine and which were not in the second annotation period was more difficult.

To compensate for the lower agreement in the second annotation round, we tried to ensure that the research team member, who was also the supervisor in the first round of annotation, made the final decision on the uncertain cases (where the two annotators differed). This supervision was expected to reduce the difference between the two annotation rounds.

Of course, the reliability of the annotation impacts the reliability of the classification. This is true even if we consider that, in the end, we decided on the classification by supervision for all uncertain annotations. We cannot expect a perfect machine classification when human annotators are uncertain.

Although our work on COVID-19 has focused on anti-vaxxer attitudes, the methodological problem we have been investigating is not specific to this topic. Language can change very quickly, especially when certain topics take on a new dimension. Even an unexpected event can redraw the discourse, as happened with the European dialogue on refugees after the outbreak of the Ukrainian-Russian war. However, the problem is even more general; concept drift can happen in almost any classification task, whether image recognition or fraud detection [7–9]. In supervised learning, if models are based on manual annotation and data to be predicted is generated after annotation, it is important to ensure the validity of our predictions.

High performance of transformer models based on reinforced learning can only be maintained if a regular reclassification phase is built into the classification process, where the model is adapted to the new incoming data. Without this, our language models can quickly lose their predictive power.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42979-024-03553-2.

Acknowledgements Not applicable.

Author Contributions Conceptualization, Z.K. and B.K.; methodology, Z.K. K.B and B.K.; formal analysis, Z.K. K.B and B.K.; data curation, K.B.; writing—original draft preparation, Z.K. and B.K.; writing—review and editing, Z.K. and B.K; visualization, Z.K.; funding acquisition, Z.K.

Funding The research was supported by the European Union within the framework of the RRF-2.3.1-21-022-00004 Artificial Intelligence National Laboratory Program. The work of Zoltán Kmetty was supported by the Bolyai Scholarship, grant number: BO/834/22.

Data Availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interest The authors declare no conflict of interest.

Ethics Approval As we did not include human subject in our research no ethical approval was needed.

References

1. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–4. https://doi.org/10.1038/nature07634.
2. Lazer D, Kennedy R, King G, Vespignani A. The parable of google flu: traps in big data analysis. Science. 2014;343:1203–5. https://doi.org/10.1126/science.1248506.
3. Bayram F, Ahmed BS, Kassler A. From concept drift to model degradation: an overview on performance-aware drift detectors. Knowl-Based Syst. 2022;245:108632.
4. Ditzler G, Roveri M, Alippi C, Polikar R. Learning in Non-stationary Environments: A Survey. IEEE Comput Intell Mag. 2015;10:12–25. https://doi.org/10.1109/MCI.2015.2471196.
5. Iwashita AS, Papa JP. An overview on concept drift learning. IEEE Access. 2019;7:1532–47. https://doi.org/10.1109/ACCESS.2018.2886026.
6. Žliobaitė I, Pechenizkiy M, Gama J. An Overview of Concept Drift Applications. In: Japkowicz N, Stefanowski J, editors. Big Data Analysis: New Algorithms for a New Society, vol. 16. Cham: Springer International Publishing; 2016. pp. 91–114. https://doi.org/10.1007/978-3-319-26989-4_4.
7. Rabiu I, Salim N, Dau A, Osman A. Recommender system based on temporal models: a systematic review. Appl Sci. 2020. https://doi.org/10.3390/app10072204.
8. Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontempi G (2015). Credit card fraud detection and concept-drift adaptation with delayed supervised information. 2015 International Joint Conference on Neural Networks (IJCNN). Killarney Ireland IEEE https://doi.org/10.1109/IJCNN.2015.7280527.
9. Fenza G, Gallo M, Loia V, Petrone A, Stanzione C. Concept-drift detection index based on fuzzy formal concept analysis for fake news classifiers. Technol Forecast Soc Chang. 2023. https://doi.org/10.1016/j.techfore.2023.122640.
10. Kulkarni V, Al-Rfou R, Perozzi B and Skiena S (2015). Statistically significant detection of linguistic change. in Proceedings of the 24th international conference on world wide web 625–635.
11. Vaswani A. et al. (2017). Attention is all you need. Advances in neural information processing systems 30.
12. Ramponi A and Plank B (2020). Neural Unsupervised Domain Adaptation in NLP—A Survey. in The 28th International Conference on Computational Linguistics (Association for Computational Linguistics, 2020).
13. Koh PW et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. in International Conference on Machine Learning 5637–5664 (PMLR, 2021).
14. Larson S, Singh N, Maheshwari S, Stewart S and Krishnaswamy U (2021). Exploring Out-of-Distribution Generalization in Text Classifiers Trained on Tobacco-3482 and RVL-CDIP. in Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16 416–423 (Springer, 2021).
15. Ma X, Xu P, Wang Z, Nallapati R and Xiang B (2019). Domain Adaptation with BERT-based Domain Classification and Data Selection. in Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019) 76–83 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/D19-6109.
16. Wang J et al. (2022) Generalizing to unseen domains: a survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering.
17. Devlin J, Chang M.-W., Lee K and Toutanova K (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in NAACL-HLT (1) (eds. Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019).
18. Lee J, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36:1234–40.
19. Beltagy I, Lo K and Cohan A (2019). SciBERT: A Pretrained Language Model for Scientific Text. in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 3615–3620 (2019).
20. Araci D (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. Preprint at http://arxiv.org/abs/1908.10063.

21. Lazaridou A, et al. Mind the gap: assessing temporal generalization in neural language models. Adv Neural Inf Process Syst. 2021;34:29348–63.
22. Luu K, Khashabi D, Gururangan S, Mandyam K and Smith NA (2022). Time Waits for No One! Analysis and Challenges of Temporal Misalignment. in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 5944–5958 (Association for Computational Linguistics, 2022). https://doi.org/10.18653/v1/2022.naacl-main.435.
23. Röttger P and Pierrehumbert J (2021). Temporal Adaptation of BERT and Performance on Downstream Document Classification: Insights from Social Media. in Findings of the Association for Computational Linguistics: EMNLP 2021 2400–2412 (Association for Computational Linguistics, 2021). https://doi.org/10.18653/v1/2021.findings-emnlp.206.
24. Goel R. et al. (2016) The social dynamics of language change in online networks. in Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11–14, 2016, Proceedings, Part I 8 41–57 (Springer, 2016).
25. Huang X. and Paul M. J (2019). Neural Temporality Adaptation for Document Classification: Diachronic Word Embeddings and Domain Adaptation Models. in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics 4113–4123 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/P19-1403.
26. Huang X. and Paul M. J (2018). Examining Temporality in Document Classification. in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 694–699 (Association for Computational Linguistics, 2018). https://doi.org/10.18653/v1/P18-2110.
27. Agarwal O, Nenkova A. Temporal effects on pre-trained models for language processing tasks. Transact Assoc Comput Ling. 2022;10:904–21.
28. Rijhwani S and Preotiuc-Pietro D (2020). Temporally-Informed Analysis of Named Entity Recognition. in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 7605–7617 (Association for Computational Linguistics, 2020). https://doi.org/10.18653/v1/2020.acl-main.680.
29. Chen S, Neves L and Solorio T (2021). Mitigating Temporal-Drift: A Simple Approach to Keep NER Models Crisp. in Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media.
30. Lukes J & Søgaard, A. Sentiment analysis under temporal shift. in Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis 65–71 (Association for Computational Linguistics, 2018). https://doi.org/10.18653/v1/W18-6210.
31. Florio K, Basile V, Polignano M, Basile P, Patti V. Time of your hate: the challenge of time in hate speech detection on social media. Appl Sci. 2020;10:4180.
32. Nemeskey DM (2021) Introducing huBERT. in XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021) 3–14.
33. Basile V (2020). Domain Adaptation for Text Classification with Weird Embeddings. in Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 (eds. Dell'Orletta, F., Monti, J. & Tamburini, F.) 37–43 (Accademia University Press, 2020). https://doi.org/10.4000/books.aaccademia.8250.
34. Zsibrita J, Vincze V, Farkas R. magyarlanc: a tool for morphological and dependency parsing of hungarian. Proc Int Conf Rec Adv Nat Lang Process RANLP. 2013;2013:763–71.
35. Gozuacik N, Sakar CO, Ozcan S. Social media-based opinion retrieval for product analysis using multi-task deep neural networks. Expert Syst Appl. 2021;183: 115388.
36. Fahfouh A, Riffi J, Mahraz M. A., Yahyaouy A. and Tairi H. (2022). A Contextual Relationship Model for Deceptive Opinion Spam Detection. IEEE Transactions on Neural Networks and Learning Systems.
37. Stefkovics Á, Krekó P, Koltai J. When reality knocks on the door. The effect of conspiracy beliefs on COVID-19 vaccine acceptance and the moderating role of experience with the virus. Soc Sci Med. 2024;356:117149.