Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 1206 – 1222

doi:10.55730/1300-0632.4043

Exploring the impact of training datasets on Turkish stance detection

Muhammed Said Zengin, Berk Utku Yenisey, Mücahid Kutlu*
Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkiye

Received: 25.11.2022 • **Accepted/Published Online:** 11.08.2023 • **Final Version:** 30.11.2023

Abstract: Stance detection has garnered considerable attention from researchers due to its broad range of applications, including fact-checking and social computing. While state-of-the-art stance detection models are usually based on supervised machine learning methods, their effectiveness is heavily reliant on the quality of training data. This problem is more prevalent in stance detection task because the stance of a text is intimately tied to the target under consideration. While numerous datasets exist for stance detection, determining their suitability for a specific target can be challenging. In this work, we focus on Turkish stance detection and explore the impact of training data on the model performance. In particular, we fine-tune BERT model with various datasets and assess their performance when the test data is the same/different compared to the training data in terms of target and domain. In addition, given the scarcity of resources for Turkish stance detection, we investigate i) whether we can use existing datasets in other languages in a cross-lingual setup, and ii) the effectiveness of data augmentation with simple automatic labeling methods. In order to conduct our experiments, we also create new Turkish stance detection datasets for various targets in different domains. In our comprehensive experiments, our findings are as follows. 1) Using training data with multiple targets in the same domain yields high performance as the model is able to learn more characteristics of expressing stance with additional data. 2) The domain of the training data plays a crucial role in achieving high performance. 3) Automatically generated data enhances performance when combined with manually annotated data. 4) Training solely on Turkish data outperforms training with the combination of Turkish and English data. Overall, our study points out the importance of creating Turkish annotated datasets for different domains to achieve high performance in stance detection.

Key words: Stance detection, natural language processing, Turkish

1. Introduction

The task of detecting whether a text expresses a favorable or unfavorable viewpoint towards a given entity or idea has numerous practical applications. For example, in order to detect the veracity of a claim, we can check whether credible resources corroborate it [1]. In addition, we can determine stance of social media accounts towards a particular topic such as vaccine and political parties, and use it to forecast public opinion [2] or investigate social polarization [3]. Hence, the stance detection task attracted many researchers in recent years [4].

Prior work on stance detection investigated several approaches such as designing specific features [5, 6], building deep learning architectures [7, 8], and utilizing social network [9]. As the target plays a crucial role in stance detection methods, numerous researchers have also sought to develop models for topics not encountered

*Correspondence: mucahidkutlu@gmail.com

during training [10, 11]. However, the majority of the studies focus on English language. Unfortunately, the data resources and studies for Turkish stance detection are highly limited.

While having limited number of studies is a general problem for many Turkish natural language processing (NLP) tasks, the transformer models, e.g., BERT [12], enabled many researchers to easily develop models with a decent performance for various Turkish NLP tasks such as sentiment analysis [13], medical text classification [14], and keyphrase extraction [15]. Furthermore, multilingual BERT models enabled researchers to leverage non-Turkish data for Turkish tasks by cross-lingual training [16]. However, the quality of the training data utilized to fine-tune these models is crucial to attain high performance. This issue becomes more prevalent in stance detection, because we also have to consider targets in the training set. Thus, developers should judiciously assess whether existing datasets can be utilized to develop a stance detection model for a particular target and domain, or develop their own dataset.

In this paper, we focus on Turkish stance detection and explore the impact of training data on the classification performance. In particular, we explore the impact of i) target, ii) domain, iii) size, and iv) language of the training data on performance. Ghosh et al. [7] report that fine-tuned BERT model is superior to other stance detection studies they compare. Thus, we use fine-tuned BERT model as our reference and evaluate its performance when fine-tuned with several different datasets. We investigate its cross-target (i.e. the target in the train set is different than the test set), cross-domain (i.e. the domain in the train set is different than the test set), cross-lingual (i.e. the language in the train set is different than the test set) performance. Specifically, we seek answers for the following research questions.

RQ-1: How much is the performance of the model affected by the target used in the train set? We first extend the dataset constructed by Küçük and Can [5] which initially covers two Turkish football clubs, Fenerbahçe (FB) and Galatasaray (GS). In particular, we annotate tweets for two other Turkish football clubs, Beşiktaş (BJK) and Trabzonspor (TS). Our experimental analysis of this dataset with four distinct targets revealed that achieving high performance does not always require training the model with the same target. Thus, we can use datasets for other targets in the same domain. In addition, our results indicate that combining datasets for different targets yields a higher performance in most cases.

RQ-2: Is it beneficial to increase the training data size by automatically judged data? In our experiments, we find out that manually judged documents yield higher performance than automatically judged ones. However, we achieve the highest performance when manually and automatically judged documents are used together, highlighting the potential benefits of incorporating automatically judged data into the training process.

RQ-3: To what extent does the performance of the model vary when used in a different domain, and consequently for a different target? To address this research question, we constructed three Turkish datasets in the domains of health, economics, and politics. In our experiments, we observe that the domain is one of the main factors affecting the performance of the model. Furthermore, we achieve the highest performance when the domains of the training and test datasets are same. Our findings suggest that domain adaptation techniques may be necessary to enhance the model's performance when it is deployed in an unfamiliar domain.

RQ-4: Does utilizing existing annotated English dataset enhance the performance of Turkish stance detection? As there exist already many English datasets, we investigate whether we can use them for Turkish stance detection. Specifically, we utilize SemEval-2016 dataset [17] which consists of five different targets in various domains. In addition, we create two other English datasets in the health and economics domains. In our experiments, we observe that using only Turkish dataset yields the highest performance in all

domains. As anticipated, combining the Turkish and English datasets improves the performance. However, in some cases, using only English data yields a result comparable to the one that use Turkish and English data together.

The contributions of our work are as follows.

1. We create new datasets for Turkish stance detection for various targets in different domains. We share our datasets to reduce the shortcomings of Turkish NLP resources, following Twitter Inc.'s redistribution policy. Additionally, we share our code to ensure the reproducibility of our experimental results[1].

2. We investigate how to fine-tune transformer models for effective stance detection for Turkish using several different datasets. Our findings provide insights into the cross-domain, cross-lingual, and cross-target performance of the models, thereby contributing to the development of more robust and effective stance detection systems in the Turkish language.

3. By using the LIME method [18], we explore how the reasoning of the model changes when trained with various datasets.

The remaining of the paper is organized as follows. In Section 2, we discuss the stance detection studies in the literature. We discuss the important factors in fine-tuning transformer models for stance detection in Section 3. Section 4 describes the datasets we use in our experiments. In Section 5, we present and discuss experimental results. We draw our conclusions in Section 6.

2. Related work

Stance detection aims to classify the attitude of a given document towards a target [19]. It is worth noting that a text can contain multiple stances and targets. For instance, in the sentence "I am devastated that Trump won the elections! but I will continue to support Clinton." the stance is negative towards Trump and positive towards Clinton. Ideally, both the stance and the target should be detected, but existing datasets, such as the SemEval 2016 Dataset [17], provide stance labels for a predefined target, obviating the need for target detection. In our study, we follow the same approach since we leverage these existing datasets in our experiments.

A number of researchers focused on the stance detection problem to apply it for various tasks such as public opinion prediction [2] and fake news detection [20]. Prior work also developed stance detection methods for various document types such as news articles [20], tweets [9, 17], forum messages [21], and social media accounts [3, 22]. While the majority of studies have focused on English, there are also studies for Arabic [23], Chinese [24], Japanese [21], Spanish [25], and Turkish [5]. In our work, we focus on stance detection for Turkish tweets (i.e. short social media messages).

Annotated datasets play a crucial role in training and testing stance detection methods, and thus, many researchers have worked on developing such datasets. For instance, the popular SemEval-2016 dataset includes 4163 English tweets covering five topics with three different labels, i.e. favor, against, and neutral. Conforti et al. [26] constructed a large dataset covering English tweets about finance with four labels including supporting, refute, commenting, and unrelated. Allaway and McKeown [27] introduce VAST dataset which consists of annotated comments from The New York Times in a wide range of topics such as Palestinian, charter schools, and vaccination. In our work, we use Küçük and Can [5]'s dataset for experiments on Turkish and SemEval-2016

[1]The URL for the dataset will be shared upon acceptance of the paper

dataset [17] for English. Furthermore, we create additional datasets because the existing Turkish tweet datasets are not sufficient for our cross-domain and cross-topic experiments.

While there exist unsupervised stance detection methods in the literature [3, 22], the majority of the studies focus on supervised machine learning approaches [4]. Researchers have investigated several features for this task, such as n-grams [6], named entities [5], topic models [28], and character tri-grams [29]. The studies focusing on social media accounts also utilized social network-based features such as retweeted accounts [23] and followers [9]. A number of studies also employed deep learning models such as bi-LSTM [8] and bi-GRU [30]. Ghosh et al. [7] compare several stance detection methods and report that fine-tuned BERT model has the highest performance on the SemEval 2016 dataset. In our work, we also use BERT model and explore the impact of training data used for fine-tuning.

Previous studies have mostly evaluated their stance detection systems with the same target in both the training and test datasets. However, this experimental setup is not suitable for real-life scenarios where the target may vary. Therefore, prior work also investigated cross-target stance detection. Xu et al. [10] propose a neural network-based solution to learn domain-specific aspects of expressing stance instead of target-specific aspects. They specifically focus on targets from the same domain, leaving cross-domain models as future work. Liang et al. [11] utilize graph neural networks to identify target-dependent and target-independent roles of words for cross-target stance detection. Zhu et al. [31] propose incorporating domain background knowledge about the target from Wikipedia into the model to increase cross-target performance of models. Zhang et al. [32] extend BiLSTM model to integrate the semantic and emotion lexicons for cross-target stance detection. Wei and Mao [19] use latent topics in text representation to increase their model's cross-target performance. Liu et al. [33] utilize ConceptNet to consider commonsense knowledge in cross-target stance detection. Sun et al. [34] first detect the topic and sentiment of texts to collect topic-invariant information. Next, they use an attention mechanism to learn the correlation between similar topics and sentiments. In our work, we follow a data-centric approach and investigate the impact of training data on cross-target, cross-domain, and cross-lingual setups.

Studies that investigate the effect of training data on stance detection are closely related to our work. For example, Li et al. [35] report that training a model with multiple datasets in different domains yields higher performance than training a separate model for each dataset. Similarly, Reuver et al. [36] show that cross-topic performance of BERT model varies across topics, and hypothesize that BERT model relies on topic-specific words instead of words related to argumentation. In our work, we specifically focus on Turkish stance detection and explore the impact of various training data factors, such as target, domain, data size, and language.

Regarding studies on Turkish stance detection, Küçük [37] first creates a dataset consisting of 700 tweets with binary labels for two football clubs. Next, she evaluates an SVM model with unigram, bigram, and hashtag features. Küçük and Can [5] extend the dataset of Küçük [37] and explore additional features such as emoticons, named entities, and external links. Rashed et al. [3] focus on user stance detection and represent users in an embedding space using Convolutional Neural Network (CNN) based multilingual universal sentence encoder. Subsequently, they project users onto a lower dimensional space and cluster them. Their method can be applied to any topic by filtering tweets based on predefined keywords. Bayrak and Kutlu [2] propose a retweet-based semisupervised method to detect political stance of users and use their method to predict results of Turkish presidential elections in 2018. Polat et al. [38] create a dataset consisting of Turkish online forum messages covering six topics. They investigate representing the text using n-grams and word embeddings, and report that XGBoost and CNN yields the highest performance. In our work, we select BERT model as reference and

explore its performance in cross-topic, cross-domain, and cross-lingual setups. In addition, we experimentally analyze the impact of data augmentation.

3. Fine-tuning transformer models for stance detection

Ghosh et al. [7] compare several stance detection methods and find out that the fine-tuned BERT model yields the highest result on the SemEval-2016 dataset. Therefore, we have also chosen to concentrate on BERT models in our research. Pretrained transformer models, such as BERT, are designed to learn the general linguistic features of a language by processing large-scale datasets with an enormous number of parameters. These models can then be fine-tuned for specific tasks using labeled data. Therefore, the characteristics of labeled datasets such as data size and label quality have great impact on the models trained with them. In this work, we explore the impact of training data for stance detection based from four different aspects.

3.1. Data domain

Statements and words to express a stance towards a particular target might vary depending on the domain of the target. For instance, if the task is to detect stance towards people (e.g. politicians), we might use words like "fan" and "enemy". However, if the target is an environmental issue (e.g., climate change), it is unlikely to use these words to express stance. Similarly, in Turkish, expressions used in different domains vary. For instance, "tutmak" (it literally means "to hold", but, in figurative speech, it means "to support") and "taraftar" (fan) are common phrases to express support for sport clubs. However, in the political domain, these phrases imply a very strong stance. Therefore, phrases like "seçmen" (voter) and "desteklemek" (to support) are more commonly used in the political domain. In our work, we utilize data from different domains including sports, health, politics, and economy to explore the impact of text domain on the stance detection performance.

3.2. Target of the stance

As mentioned before, many existing datasets have a predefined target and the task is not involved in detecting the target. However, this raises concerns when applying existing datasets to detect stances towards different targets, as certain expressions used to convey stances may be specific to the target being considered. For example, fans of Liverpool Football Club call themselves "Kopites"[2] while the fans of Manchester United are usually called "The Citizens"[3]. Similar target-specific phrases exist in Turkish, such as "Çarşı" (bazaar) for fans of Beşiktaş Football Club. Therefore, a fine-tuned model might mistakenly learn target-specific phrases. In our study, we investigate the impact of training models with the same or different targets in the test set.

3.3. Data size

While we need large amounts of labeled data for effective training, obtaining human annotations can be a time-consuming and costly process. Therefore, prior work investigated several methods to optimize annotation budget such as crowd-sourcing [39] and active learning [40]. However, these approaches still require substantial amount of money and time. Another popular approach to increase labeled data size is weak-supervision in which data are labeled automatically using another model. In weak-supervision methods, it is expected that having a large amount of data with noisy labels can be more advantageous than having a limited amount of

[2]https://en.wikipedia.org/wiki/Liverpool_F.C.
[3]https://www.mancity.com/fans

data with correct annotations. Similarly, in our work, we explore the impact of automatically labeled data on fine-tuning stance detection models.

3.4. Language

NLP solutions are often tailored to a specific language. However, multilingual transformer models such as MBERT provide an exciting alternative approach to develop solutions in which train and test datasets can be in different languages. This enables the utilization of the abundant resources of a language, e.g., English, when resources are limited for the target language, thus addressing one of the main challenges in developing NLP models for low and medium-resource languages. In our work, we explore the feasibility of employing English datasets for Turkish stance detection tasks.

4. Data

We need various datasets covering different topics, languages and targets in order to conduct our experiments. However, the available dataset for Turkish stance detection is very limited. As we focus on Turkish tweets, to our knowledge, there is only one dataset constructed by Küçük and Can [5] which contains 1065 tweets with binary labels (i.e. favor or against) for two popular football clubs in Türkiye, namely Fenerbahçe and Galatasaray. Furthermore, due to Twitter's prohibition in redistribution of tweet contents, Küçük and Can [5] share only tweet IDs for their labeled data. Consequently, we have to recrawl tweets using their IDs. However, we could crawl only 454 tweets due to deleted tweets and closed accounts. Therefore, we created new datasets for stance detection, covering various topics and targets. Now we explain our annotation procedure (Section 4.1) and the datasets used in this work (Section 4.2).

4.1. Annotation procedure

We use binary labels (i.e. favor and against) as Küçük and Can [5]. We first crawled tweets for annotation and removed the duplicates. Each tweet has been labeled by two annotators. We filter out tweets that annotators disagree. In addition, we also eliminated tweets with multiple targets (e.g., "I hate FB, but love GS"), contradicting stances, and a neutral stance. Note that these filtering processes are essential to ensure label quality. For instance, we initially crawled 1741 tweets including duplicates for the domain of football, however, we eliminated %52.26 of the initial crawl due to our elimination process explained above. The annotation process was carried out by four graduate students, who are native speakers of Turkish and also fluent in English.

4.2. Datasets

Now we present the datasets we used in this study and provide brief statistics about them. To provide clear identification, we adopt the following notation for the dataset names: $D_{<DOMAIN>\_<LANGUAGE>}$.

4.2.1. Manually-labeled Turkish datasets

We have created four datasets in different domains, including football, health, economics, and politics. To obtain these datasets, we utilized the Twint library[4] and crawled tweets using different sets of keywords relevant to each domain. Subsequently, we annotated them following the procedure explained in Section 4.1.

[4]https://pypi.org/project/twint/

For the dataset with tweets about football ($D_{Football\_TR}$), we focus on tweets with known targets. Specifically, in an effort to broaden the coverage of the existing dataset introduced by Küçük and Can [5], we manually annotated additional tweets pertaining to other football clubs. Specifically, the four football teams targeted in this study include Fenerbahçe (FB), Galatasaray (GS), Beşiktaş (BJK), and Trabzonspor (TS). To achieve this, we utilized team names as tracking keywords and crawled 1000 tweets for each of the targeted teams. We combined our labels with those obtained by Küçük and Can to construct this dataset. **Table 1** presents data size for each football club.

Table 1. Label distribution in $D_{Football\_TR}$. The dataset contains 247 and 297 annotations for Galatasaray and Fenerbahçe Clubs from Küçük and Can [5]'s dataset, respectively.

Target	Favor	Against	Total
Galatasaray	201	152	353
Fenerbahçe	126	150	276
Beşiktaş	52	48	100
Trabzonspor	37	64	101

Table 2. Turkish data and label distribution in different domains. English translations are given in parentheses when necessary.

Name	Domain	Keywords used for crawling	Favor	Against	Total
$D_{Football\_TR}$	Football	beşiktaş, galatasaray, trabzonspor, fenerbahçe	416	414	830
D_{Health_TR}	Health	covid, aşı (vaccine), maske (mask), fahrettin koca (i.e. the name of the Minister of Health of Türkiye)	475	765	1240
$D_{Economics\_TR}$	Economics	bitcoin, kripto para (cryptocurrency), dolar (dollar), ekonomi (economics)	620	1215	1835
$D_{Politics\_TR}$	Politics	akp, mhp, chp (i.e. abbreviations for three political parties in Türkiye)	276	300	576

The datasets pertaining to the domains of health, politics, and economics do not focus on a particular target; rather, our aim was to ensure that each tweet had a single target. **Table 2** shows the keywords used to crawl tweets for each dataset and the number of each label. In particular, we used COVID-19-related keywords for the health topic. Regarding tweets about economics, we used the word "ekonomi" (economics) to capture individuals' views on Türkiye's economic status. Additionally, we used keywords about different currencies as there are heated discussions about the use of US dollars and crypto-currencies on social media platforms. Lastly, we employed common abbreviations for three political parties to retrieve tweets expressing individuals' stances on various political issues. The number of labeled tweets for each topic varies because of elimination during the manual judging process, as explained in Section 4.1.

4.2.2. Manually-labeled English datasets

In our cross-lingual experiments, we utilize the SemEval-2016 dataset, which covers five different topics including atheism, climate change, Hillary Clinton, abortion, and feminism. However, since the SemEval-2016 dataset mainly pertains to political topics and does not cover health and economics domains, we created two small datasets on economics ($D_{Economics\_EN}$) and health (D_{Health_EN}) domains as our Turkish datasets. To obtain

tweets in the health domain, we crawled tweets tracking "vaccine" and "covid" keywords. Additionally, we used "bitcoin" and "cryptocurrency" keywords to crawl tweets about economics. **Table 3** presents detailed information on the English datasets we use.

Table 3. Label distribution in English datasets. The last five rows show SemEval2016 dataset. We use the following abbreviations for topics in SemEval2016 dataset: At (atheism), HC (Hillary Clinton), Fe (Feminism), CC (climate change), Ab (abortion).

Name	Topic	Favor	Against	Total
D_{Health_EN}	Health	20	15	35
$D_{Economics\_EN}$	Economy	52	13	65
$SE16_{At}$	Atheism	124	464	588
$SE16_{HC}$	Hillary Clinton	305	832	1137
$SE16_{Fe}$	Feminism	268	511	779
$SE16_{CC}$	Climate change	335	26	361
$SE16_{Ab}$	Abortion	124	464	588

4.2.3. Automatically-labeled Turkish datasets

In order to create an automatically annotated dataset, we could first train a sophisticated model and then apply it to each data item. However, this approach also requires a substantial amount of labeled data. As available datasets are limited in Turkish, we use a simple, rule-based tagging method which can be easily applied to many different topics. In particular, for each football team, we use as a target in $D_{Footbal\_TR}$ dataset (i.e. Fenerbahçe, Beşiktaş, Trabsonspor, and Galatasaray), we crawl tweets in which the team name appears right after the word "şikeci" (flammer) or "şampiyon" (champion). We label tweets with the word "şikeci" as against the respective team and those containing the word "şampiyon" as supporting the team. We select these words because they are frequently used by Turkish football fans. For each team and label, we crawled 5000 tweets, yielding a total of 40,000 tagged tweets.

5. Experiments

In this section, we describe experimental setup (Section 5.1), present and discuss results (Section 5.2), and qualitatively analyze the impact of training data (Section 5.3).

5.1. Experimental setup

The pretrained models are obtained from Huggingface [41]. We use k-train library [42] for fine-tuning models. We use BERTurk[5] for Turkish data unless stated otherwise, and M-BERT [12] for English data and cross-lingual experiments. Among the models we compare, we use exactly the same parameters to make a fair comparison. We report macro-average F_1 score. The train and test set ratio is set to 90:10. Throughout the paper, we use dataset names for their train and test subsets rather than explicitly indicating whether they are part of the train or test set, to facilitate clarity of exposition.

[5]https://zenodo.org/record/3770924 [accessed 09/06/2023].

5.2. Experimental results

Cross-target experiments. In our first experiment, we explore the impact of target in stance detection. Therefore, we use $D_{Footboll\_TR}$ which consists of manually labeled Turkish tweets about various football teams (i.e. GS, FB, BJK, and TS). We partition the training data of $D_{Footboll\_TR}$ into different subsets based on the targets and fine-tune BerTurk model with each subset separately. Subsequently, we evaluate each fine-tuned model on the test data. In order to calculate their cross-target performance, we also report results for each target. The results are shown in **Table 4.**

Table 4. F_1 score of BerTurk model on various datasets when fine-tuned with datasets with different targets. The highest score for each case is written in **bold**.

Target in training data	**Single target test**				**Multi target test**
	GS	FB	BJK	TS	GS+FB+BJK+TS
GS	0.633	0.695	0.650	0.587	0.640
FB	0.631	0.589	0.413	0.492	0.585
BJK	0.245	0.294	0.576	0.492	0.384
TS	0.419	0.711	0.529	0.689	0.567
GS + FB	0.834	0.639	0.700	0.487	0.744
BJK + TS	0.705	0.677	0.740	**0.849**	0.708
GS+FB+BJK+TS	**0.890**	**0.750**	**0.879**	0.707	**0.823**

The findings of our cross-target experiments, as presented in Table 4, reveal several important observations. Firstly, as expected, training with more data covering all targets yields the highest performance in almost all cases. Similarly, models trained with two targets outperform models trained with a single target in most of the cases. Secondly, when we use a single target to train the model, training with GS targets yields highest performance when the target in the test set is GS and BJK. Likewise, training with TS target outperforms other models trained with a single target when the target in the test set is FB and TS. These results suggest that when the training data is limited, it may be beneficial to use datasets for different targets. However, it should be noted that, when training models with two targets, the best performing model for each target contains that target in the training set, highlighting the importance of training data with the same target. Lastly, we observe that among models trained with a single target, the model trained with the GS target achieves the highest F1 score when there are four targets in the test set. This might be because the number of training data for GS is higher than that for other targets, as shown in Table 1).

Overall, the experimental results indicate the crucial role of the size of the training dataset and highlight the potential benefits of merging datasets with different targets in the same domain to enhance classification accuracy. Furthermore, the findings suggest that in cases where training data for a particular target is unavailable, leveraging existing data for a different target can still be beneficial for training models.

Automatic labelling. In this experiment, we evaluate the impact of utilizing automatically labeled data to fine-tune models. Specifically, we fine-tune BerTurk model using datasets that are automatically labeled for each football club. In addition, we combine the manually labeled data for each football club (i.e. corresponding training set of $D_{Footboll\_TR}$) with the automatically labeled ones and fine-tune the BerTurk model accordingly. Additionally, we combine all automatically labeled data (i.e. All_A) with the training set of $D_{Footboll\_TR}$ (i.e. manually labeled tweets about football) and train the model accordingly. We present F_1 scores of each model trained with a different dataset in **Table 5**. We again report results for each target separately to explore their cross-target performance.

Table 5. F_1 scores of BerTurk model fine-tuned with varying datasets with data items automatically labeled. X_A stands for dataset with automatically labeled tweets for the team X. X_M stands for the subset of $D_{Footboll\_TR}$) for the team X. ALL_A is the combination of all automatically labeled data. The highest score for each case is written in **bold**.

Training data	**Single target test**				**Multi target test**
	GS	FB	BJK	TS	GS+FB+BJK+TS
GS_A	0.444	0.714	0.607	0.750	0.596
FB_A	0.474	0.714	0.560	0.654	0.569
BJK_A	0.289	0.369	0.354	0.514	0.411
TS_A	0.303	0.552	0.457	0.786	0.544
$GS_M + GS_A$	0.638	0.428	0.777	0.776	0.696
$FB_M + FB_A$	0.515	0.779	0.601	0.805	0.652
$BJK_M + BJK_A$	0.430	0.374	0.419	0.196	0.301
$TS_M + TS_A$	0.572	0.576	0.798	0.823	0.668
All_A	0.303	0.294	0.422	0.823	0.535
$D_{Footboll\_TR} + All_A$	**0.860**	**0.818**	**0.870**	**0.833**	**0.870**

The results demonstrate that utilizing manually labeled data in addition to automatically labeled data improves the performance of the model in almost all cases. In addition, comparing the results presented in Table 4 vs. Table 5, we observe that incorporating automatically labeled data leads to increased model performance in most cases. For instance, the model trained with GS_M (i.e. the subset of $D_{Footboll\_TR}$ for GS) and GS_A outperforms the model trained with GS_M on the whole test set, i.e. multi target test case (0.696 vs. 0.640). Regarding automatically labeled data vs. manually labeled data, we observe that the models trained with a single target and manual annotations (Table 4) outperform the models trained with a single target and automatic labels in most cases. For example, the model trained with GS_M (in Table 4) outperforms the model trained with GS_A (in Table 5) in multi target set case (0.696 vs. 0.596). However, in some cases, the models trained solely with automatic labels outperform those trained with only manually labeled data, such as when the target in the test set is TS.

Overall, our experiments point out the importance of manually labeled data. However, our experiments also show that incorporating additional training data labeled by a simple automatic method can improve model performance. Furthermore, we observe that in some cases, the performance of models trained with automatically labeled data is higher than those trained with manually labeled data.

Cross domain experiments. In this experiment, we use manually labeled datasets presented in Table 2 in order to investigate the impact of the domain of the training data. We employ MBERT as our underlying model in this experiment to facilitate comparison with our subsequent cross-lingual experiments. The results are shown in **Table 6**.

We observe that the models fine-tuned with datasets in different domains exhibit lower classification accuracy as compared to the models trained with data in the same domain, irrespective of the size of the training data. These results suggest that the availability of training data in the same domain as the test data is a crucial factor in achieving high classification performance.

Cross lingual experiments. In this experiment, we investigate the transferability of a pretrained MBERT model from English to Turkish tweets. Specifically, we fine-tune MBERT using various datasets containing

Table 6. F_1 score of MBERT model fine-tuned with datasets in different domains. The highest score for each case is written in **bold**.

Domain of the training data	**Domain of the test data**			
	Football	Health	Economics	Politics
Football	**0.793**	0.610	0.556	0.476
Health	0.548	**0.864**	0.700	0.665
Economics	0.597	0.666	**0.912**	0.590
Politics	0.511	0.504	0.597	**0.771**

English tweets and evaluate their performance on various Turkish datasets. We train the model using each subset of the SemEval16 dataset separately, yielding five different models. Additionally, we fine-tune the MBERT model by combining English and Turkish datasets in the same domain. In the politics domain, we use the Hillary Clinton topic from the SemEval 2016 dataset, as it is directly related to politics. The results are shown in **Table 7**.

Table 7. F_1 score of MBERT model fine-tuned with varying datasets. Note that the train and test splits of each data is used to train and test the models, respectively. The highest score for each case is written in **bold**.

Training data	**Test data**			
	$D_{Football\_TR}$	D_{Health_TR}	$D_{Economics\_TR}$	$D_{Politics\_TR}$
$SE16_{HC}$	0.566	0.629	0.518	0.466
$SE16_{Fe}$	0.381	0.486	0.582	0.296
$SE16_{CC}$	0.288	0.201	0.133	0.372
$SE16_{Ab}$	0.408	0.454	0.582	0.296
$SE16_{At}$	0.483	0.515	0.569	0.380
$D_{Health\_TR} + D_{Health\_EN}$	0.503	**0.765**	**0.657**	0.358
$D_{Economics\_TR} + D_{Economics\_EN}$	0.556	0.381	0.588	0.296
$D_{Politics\_TR} + SE16_{HC}$	**0.680**	0.605	0.583	**0.756**

Our observations are as follows. Firstly, the topical similarity between train and test sets again plays a crucial role in achieving high performance. Specifically, the models trained with health and politics dataset outperform others when the domain of the test set is also health and politics, respectively. However, interestingly, the model trained with health data outperforms the one trained with economics data when the test set is $D_{Economics\_TR}$. Secondly, models trained with Turkish and English datasets outperform models trained with subsets of SemEval16 dataset in most of the cases, showing the importance of using Turkish dataset. Thirdly, among SemEval16 datasets, $SE16_{HC}$ yields the highest performance on $D_{Football\_TR}$, D_{Health_TR}, and $D_{Politics\_TR}$. The superior performance we achieve by using $SE16_{HC}$ might be because its size is larger than other subsets of SemEval16 dataset (See Table 3). However, we note that $SE16_{Fe}$ and $SE16_{Ab}$ yield higher performance than others in $D_{Economics\_TR}$. Lastly, comparing models trained with only Turkish data (Table 6) and models trained with English and Turkish data (Table 7), we observe that using additional English data yields a lower performance in almost all cases. For instance, the model trained with the training set of $D_{Economics\_TR}$ outperforms the model trained with the training set of $D_{Economics\_TR}$ and $D_{Economics\_EN}$ on the test set of $D_{Economics\_TR}$ (0.912 vs. 0.588).

5.3. Qualitative analysis

Now we conduct a qualitative analysis to have a deeper understanding of how models' predictions vary with respect to the training data. Specifically, we train BERT models using different training sets and use Eli5 implementation[6] of LIME [18] to detect which words are more effective for the predicted labels and how they vary with respect to the training set.

In our first experiment, we focus on the results we presented before for cross-target, cross-domain, and cross-lingual performance of the BERT model. However, due to space limitations, we present results only for two cases for each experimental setup. The results are shown in **Table 8**.

Table 8. LIME results of BERT models fine-tuned with different datasets for various cases in our test sets. Green-colored and red-colored words are the words that have a positive and negative impact on the corresponding predicted label, respectively. The darker color represents a higher impact. The translation of each sentence is given in italics.

Row	Training Data	LIME Results	Predicted Label
1	GS_M	vizyonsuz baskan vizyonsuz teknik direktor. basarisiz takim. #galatasaray *(The visionless president, the visionless technical director, the unsuccessful team. galatasaray)*	Against (P = 0.999)
2	GS_M	yeter bu rezillik 👎 yeter yeter artık rahat,keyifli bir futbol izlemek istiyoruz. bu kadar basit değil fenerbahçe ruhu denen bir şey kalmamış *(Enough of this disgrace, enough! We just want to watch an enjoyable football. It's not that simple. There is nothing left of the so-called Fenerbahçe spirit.)*	Against (P = 0.991)
3	$D_{Football\_TR}$	ne yaptınız ki bir takım koskoca bir beşiktaş anca bu kadar dirayetsiz olur *(What have you done that a big team like Beşiktaş can only be this incapable?)*	Against (P = 0.997)
4	$D_{Football\_TR}$	isimizin basındayız çünkü biz sağlık çalışanıyız. sisli kapalı bir hava, aşılara devam *(We are on duty because we are healthcare workers. It's a foggy and cloudy weather, but we continue with the vaccinations.)*	Against (P = 0.887)
5	$SE16_{HC}$	gerçek ülkücü milliyetçi parti mhp dir yarın göreceğiz *(The real idealist nationalist party is MHP. We will see tomorrow.)*	Against (P = 0.545)
6	$SE16_{HC}$	@hillaryclinton made me proud today! nothing like reinforcing what I already knew! #semst	Supporting (P=0.550

The first two rows of Table 8 present results for cross-target experiments. In this case, we fine-tune BerTurk model using GS_M (i.e. manually annotated Turkish data where the target is Galatasaray) and evaluate its performance when the target is Galatasaray and Fenerbahçe, separately. In both cases, the model is able to predict the stance accurately. We see that the words with very negative sentiment such as "vizyonsuz" (visionless), "basarisiz" (unsuccessful), "rezillik" (disgrace) have positive impact on the predicted label, i.e. "against", as expected. On the other hand, the word "keyifli" (enjoyable) has a negative impact on the "against" label.

The third and fourth rows of Table 8 present results for cross-domain experiments. Specifically, we fine-tune BerTurk model using $D_{Football\_TR}$ and evaluate its performance on tweets about football and health, separately. The model's prediction is correct for the tweet about football but incorrect for the other one. In the tweet about vaccines, the word "devam" (continue) has a positive impact on the predicted "against" label while it is the main word that makes the tweet favoring vaccines.

[6] https://github.com/eli5-org/eli5

The fifth and sixth rows of Table 8 present results for cross-lingual experiments. We fine-tune MBERT model using $SE16_{HC}$ dataset and evaluate its performance on a Turkish and an English example. In the Turkish tweet, the prediction is incorrect and words unrelated to the stance such as "yarın" (tomorrow) seem influential in the prediction. In the English tweet, the prediction is correct and the word "proud" has a positive impact on the supporting label, as expected.

Table 9. LIME result of BerTurk which is fune-tuned with different datasets for a sample text: "Mali açıdan zor durumda olan takımımızın çok fazla sakat oyunucunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız." (We stand by our president, as having too many injured players will put our team which is in financial difficulties, in a tough situation before the presidential election). Green-colored and red-colored words are the words that have a positive and negative impact on the corresponding predicted label, respectively. The darker color represents a higher impact.

Training data	LIME results	Predicted label
$D_{Football\_TR}$	mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız.	Supporting (P = 0.968)
D_{Health_TR}	mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız.	Against (P = 0.611)
$D_{Economics\_TR}$	mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız.	Supporting (P = 0.751)
$D_{Politics\_TR}$	mali açıdan zor durumda olan takımımızın çok fazla sakat oyuncusunun olması başkanlık seçimi öncesinde takımı zor duruma sokacağı için başkanımızın yanındayız.	Against (P = 0.983)

In our next experiment, we train BerTurk model using four different Turkish datasets separately, and make a prediction for the same sentence. The sentence we use in the prediction and LIME results are shown in **Table 9**. While the correct label for the selected sentence is supporting, only two models could predict it correctly. As the sentence is about a sport team, the model trained with the football data correctly predicts the label with a very high prediction confidence. We also observe that the most effective words for the "supporting" label (shown in green) are the words explicitly expressing the support of the text's author: "yanındayız" (we stand by) and "başkanımızın" (our president). However, even though the model trained with economical data makes a correct prediction, the words affecting its prediction do not make sense. In particular, the word "yanındayız" (we stand by) is considered as related to the "against" label while the words "zor duruma sokacağı" (put in a tough situation) are aligned with the "supporting" label. Furthermore, while the model trained with health data could correctly align the phrase "başkanımızın yanındayız" (we stand by our president) with the supporting label[7], it seems that the model got confused with other words, causing a wrong prediction eventually. Overall, our analysis shows the importance of using training data from the same domain because it is the only model which could make a correct prediction with a reasonable explanation.

In our next qualitative analysis, we fine-tune MBERT model using $D_{Economics\_TR}$ and $D_{Economics\_EN}$ separately, yielding two models trained on the same domain in different languages. Subsequently, we get LIME results for the same Turkish and English sentences, for both models, yielding four different outputs. The results are shown in **Table 10**.

The stance of the text is against using cryptocurrencies. Only the model trained with Turkish data predicts it correctly when the text's Turkish version is used. As its explanation, we see that "batması" (the bankruptcy of) is correctly correlated with the "against" label. On the other hand, the model trained with

[7]As the predicted label is "against", the red colored words are the ones that have impact on the "supporting" label.

Table 10. LIME Result of two MBERT models which are fine-tuned with English and Turkish datasets in economics domain separately for the same sentence in English and Turkish: "Kripto paralara hiç yatırım yapmayan insan en mutlu insandır. FTX'in batması bunu bize bir kez daha göstermiş oldu." (The person who has never invested in cryptocurrencies is the happiest person. The bankruptcy of FTX has shown this to us once again). Green-colored and red-colored words are the words that have a positive and negative impact on the corresponding predicted label, respectively. The darker color represents a higher impact.

Training data	LIME results	Predicted label
$D_{Economics\_TR}$	kripto paralara hiç yatırım yapmayan insan en mutlu insandır. ftx'in batması bunu bize bir kez daha göstermiş oldu.	Against (P = 0.976)
$D_{Economics\_EN}$	kripto paralara hiç yatırım yapmayan insan en mutlu insandır. ftx'in batması bunu bize bir kez daha göstermiş oldu.	Supporting (P = 0.894)
$D_{Economics\_TR}$	the person who has never invested in cryptocurrencies is the happiest person. the bankruptcy of ftx has shown this to us once again.	Supporting (P = 0.523)
$D_{Economics\_EN}$	the person who has never invested in cryptocurrencies is the happiest person. the bankruptcy of ftx has shown this to us once again.	Supporting (P = 0.842)

English data makes an incorrect prediction with the Turkish statement and its explanations are not reasonable such that the words "daha" (again), "insandır" (it is the person), and "FTX" are correlated with the "against" label.

When MBERT model was tested on the English version of the statement, both models failed in prediction. However, some words that have an impact for "against" label (i.e. "never", "happiest", and "bankruptcy") are reasonable in the LIME explanation of both models. The words "never" and "bankruptcy" are shown in red dark color for the model trained with English data, suggesting that it was more successful at catching semantics of the words than the model trained with Turkish data. Overall, our qualitative analysis points out using training data that matches the language of the test data.

6. Conclusion

In this paper, we focused on Turkish stance detection and explored how to select and/or create a dataset to train a stance detection model. We use fine-tuned BERT model as our reference model because Ghosh et al. [7] report that it outperforms other stance detection models. We first create Turkish datasets which cover multiple targets in several domains. Subsequently, we evaluate how the classification performance of BERT model changes when fine-tuned with different datasets. In particular, we assess its cross-target, cross-domain, cross-lingual performance. Furthermore, we investigate if automatically labeled data is beneficial to train models.

In our comprehensive experiments, we have the following observations. Firstly, combining data for different targets in the same domain yields higher performance than training with the dataset which has the same target with the test data. Secondly, the domain of the train data plays an important factor in the performance of the model. Thirdly, manually annotated data yields higher performance than automatically judged data. However, using automatically judged data together with the manually annotated ones improves the classification performance. Lastly, while we can build models with English data using multilingual models, we achieve the highest classification performance when we use only Turkish data.

Our work can be extended in several directions. Firstly, we plan to extend the datasets we use in our experiments and conduct experiments with larger datasets covering more targets and domains. Secondly, we

used a very simple way to automatically label the tweets in the respective experiment. We plan to develop more sophisticated ways to automatically label data and conduct experiments accordingly. Lastly, we use only one state-of-the-art model in our experiments. Hence, we plan to implement other models and assess their performance with various datasets.

Acknowledgment

This study was funded by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) ARDEB 3501 Grant No 120E514. The statements made herein are solely the responsibility of the authors.

References

[1] Hardalov M, Arora A, Nakov P, Augenstein I. A survey on stance detection for mis-and disinformation identification. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1259–1277, 2022.

[2] Bayrak C, Kutlu M. Predicting election results via social media: A case study for 2018 turkish presidential election. IEEE Transactions on Computational Social Systems, 2022.

[3] Rashed A, Kutlu M, Darwish K, Elsayed T, Bayrak C. Embeddings-based clustering for target specific stances: The case of a polarized turkey. In Proceedings of the International AAAI Conference on Web and Social Media, volume 15, pages 537–548, 2021.

[4] AlDayel A, Magdy W. Stance detection on social media: State of the art and trends. Information Processing & Management 2021; 58 (4):102597

[5] Küçük D, Can F. Stance detection on tweets: An svm-based approach. arXiv preprint arXiv:1803.08910, 2018.

[6] Mohammad SM, Sobhani P, Kiritchenko S. Stance and sentiment in tweets. ACM Transactions on Internet Technology (TOIT) 2017;17 (3):1–23

[7] Ghosh S, Singhania P, Singh S, Rudra K, Ghosh S. Stance detection in web and social media: a comparative study. In International Conference of the Cross-Language Evaluation Forum for European Languages, pages 75–87, 2019.

[8] Siddiqua UA, Chy AN, Aono M. Tweet stance detection using an attention based neural ensemble model. 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) 2019; 1868–1873

[9] Aldayel A, Magdy W. Your stance is exposed! analysing possible factors for stance detection on social media. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–20

[10] Xu C, Paris C, Nepal S, Sparks R. Cross-target stance classification with self-attention networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 778–783, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Liang B, Fu Y, Gui L, Yang M, Du J et al. Target-adaptive graph for cross-target stance detection. In Proceedings of the Web Conference 2021, pages 3453–3464, 2021.

[12] Devlin J, Chang M, Lee K, Toutanova K. 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), pages 4171–4186, 2019.

[13] Acikalin UU, Bardak B, Kutlu M. Turkish sentiment analysis using bert. In 2020 28th Signal Processing and Communications Applications Conference (SIU) 2020; 1–4

[14] Çelıkten A, Bulut H. Turkish medical text classification using bert. In 2021 29th Signal Processing and Communications Applications Conference (SIU) 2021; 1–4

[15] Ayan ET, Arslan R, Zengin MS, Duru HA, Salman S et al. Turkish keyphrase extraction from web pages with bert. In 2021 29th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2021.

[16] Zengin MS, Kartal YS, Kutlu M. Tobb etu at checkthat! 2021: Data engineering for detecting check-worthy claims. In CLEF (Working Notes) 2021;670–680

[17] Mohammad S, Kiritchenko S, Sobhani P, Zhu X, Cherry C. SemEval-2016 task 6: Detecting stance in tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 31–41, San Diego, California, June 2016.

[18] Ribeiro MT, Singh S, Guestrin C. "why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

[19] Wei P, Mao W. Modeling transferable topics for cross-target stance detection. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1173–1176, 2019.

[20] Umer M, Imtiaz Z, Ullah S, Mehmood A, Choi GS et al. Fake news stance detection using deep learning architecture (cnn-lstm). IEEE Access, 8:156695–156706, 2020.

[21] Murakami A, Raymond R. Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 869– 875, 2010.

[22] Darwish K, Stefanov P, Aupetit M, Nakov P. Unsupervised user stance detection on twitter. In Proceedings of the International AAAI Conference on Web and Social Media, volume 14, pages 141–152, 2020.

[23] Darwish K, Magdy W, Zanouda T. Improved stance prediction in a user similarity feature space. In Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017, pages 145–148, 2017

[24] Xu R, Zhou Y, Wu D, Gui L, Du J et al. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In Natural language understanding and intelligent applications, pages 907–916. 2016.

[25] Taulé M, Martí MA, Rangel FM, Rosso P, Bosco C et al. Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In 2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017, volume 1881, pages 157–177. CEUR-WS, 2017.

[26] Conforti C, Berndt J, Pilehvar MT, Giannitsarou C, Toxvaerd F et al. Will-they-won't-they: A very large dataset for stance detection on twitter. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1715–1724, 2020.

[27] Allaway E, Mckeown K. Zero-shot stance detection: A dataset and model using generalized topic representations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8913–8931, 2020.

[28] Elfardy H, Diab M. Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 434–439, 2016.

[29] Bøhler H, Asla P, Marsi E, Sætre R. Idi@ ntnu at semeval-2016 task 6: Detecting stance in tweets using shallow features and glove vectors for word representation. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 445–450, 2016.

[30] Li Y, Caragea C. Multi-task stance detection with sentiment and stance lexicons. 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), pages 6299–6305, 2019.

[31] Zhu Q, Liang B, Sun J, Du J, Zhou L et al. Enhancing zero-shot stance detection via targeted background knowledge. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2070–2075, 2022.

[32] Zhang B, Yang M, Li X, Ye Y, Xu X et al. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020; 3188–3197

[33] Liu R, Lin Z, Tan Y, Wang W. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021;3152–3157

[34] Sun Q, Xi X, Sun J, Wang Z, Xu H. Stance detection with a multi-target adversarial attention network. Transactions on Asian and Low-Resource Language Information Processing, 2022.

[35] Li Y, Zhao C, Caragea C. Improving stance detection with multi-dataset learning and knowledge distillation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing 2021;6332–6345

[36] Reuver M, Verberne S, Morante R, Fokkens A. Is stance detection topic-independent and cross-topic generalizable?-a reproduction study. In Proceedings of the 8th Workshop on Argument Mining 2021;46–56

[37] Küçük D. Stance detection in turkish tweets. In Workshop on Social Media World Sensors (SIDEWAYS), 2017.

[38] Polat KK, Bayazıt NG, YILDIZ OT. Türkçe duruş tespit analizi. Avrupa Bilim ve Teknoloji Dergisi 2021; (23):99–107

[39] Sabou M, Bontcheva K, Derczynski L, Scharl A. Corpus annotation through crowdsourcing: Towards best practice guidelines. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 859–866, 2014.

[40] Ren P, Xiao Y, Chang X, Huang PY, Li Z et al. A survey of deep active learning. ACM Computing Surveys (CSUR) 2021;54 (9):1–40

[41] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C et al. Huggingface's trans- formers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, 2019.

[42] Maiya AS. ktrain: A low-code library for augmented machine learning. arXiv preprint arXiv:2004.10703, 2020.