A Dataset and BERT-based Models for Targeted Sentiment Analysis on Turkish Texts

M. Melih Mutlu
Department of Computer Engineering
Boğaziçi University
melih.mutlu@boun.edu.tr

Arzucan Özgür
Department of Computer Engineering
Boğaziçi University
arzucan.ozgur@boun.edu.tr

Abstract

Targeted Sentiment Analysis aims to extract sentiment towards a particular target from a given text. It is a field that is attracting attention due to the increasing accessibility of the Internet, which leads people to generate an enormous amount of data. Sentiment analysis, which in general requires annotated data for training, is a well-researched area for widely studied languages such as English. For low-resource languages such as Turkish, there is a lack of such annotated data. We present an annotated Turkish dataset suitable for targeted sentiment analysis. We also propose BERT-based models with different architectures to accomplish the task of targeted sentiment analysis. The results demonstrate that the proposed models outperform the traditional sentiment analysis models for the targeted sentiment analysis task.

1 Introduction

The increasing availability of the Internet and the growing number of online platforms allowed people to easily create online content. Because of the value of mining the people's opinions, the sentimental information contained in this online data makes sentiment analysis (SA) an interesting topic. It is an area that is attracting the attention not only of academic researchers, but also of businesses and governments (Birjali et al., 2021) and has become a rapidly growing field, as evidenced by the number of recent SA papers published (Mäntylä et al., 2018).

The problem with traditional sentiment analysis is that it cannot capture the different attitudes toward multiple aspects in a given text. For example, if the given text is *"Phones from this brand are great, but I don't really like their laptops"*, the sentiment towards the two targets *"phone"* and *"laptop"* are positive and negative, respectively. Traditional sentiment analysis methods would not be able to detect this opposing sentiment for *"phone"* and *"laptop"*, but would assign an overall sentiment for the text. Targeted Sentiment Analysis (TSA) aims to overcome this challenge and extracts sentiment from a given text with respect to a specific target. One of the challenges of TSA is the lack of available datasets. Both TSA and SA require labeled datasets. Collecting data from various sources and labeling them, which is mostly done manually, is an expensive process. Although the number of datasets suitable for SA has recently increased due to new studies in the SA area, not all SA datasets are usable for TSA (Pei et al., 2019). TSA requires more refined datasets. The labels should reflect the sentiment toward targets rather than the overall sentiment of the sentences.

English is the most studied language for sentiment analysis (Dashtipour et al., 2016). SA models that perform satisfactorily for English do not seem to always work with similar performance for Turkish (Kaya et al., 2012). In this work, we create a manually annotated dataset from Twitter specifically labeled for both traditional and targeted sentiment analysis in Turkish. Then, we experiment with different model architectures for the Turkish TSA task. Experimental results demonstrate that our techniques outperform traditional sentiment analysis models.

1.1 Problem Definition

Let E denotes all entities in a given document D such that:

$D = \{w_1, \ldots, w_k\}$ each w is a word; $k \in Z^+$
$E = \{e_1, \ldots, e_l\}$ each e is an entity; $l \in Z^+$
$T = \{t_1, \ldots, t_m\}$ t_i is a target; $t_i \in E$; $m, i \in Z^+$

The objective of targeted sentiment analysis is to find all sentiment (s_i, t_i) pairs in document D where t_i is a target from T and s_i is the sentiment toward t_i.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 467 - 472

Tweet	Sentence Sentiment	Targeted Sentiment
coca cola daha iyi lezzet olarak (*coca cola*'s taste is better)	positive	positive
whatsapp çöktü de biraz rahatladım bildirimlerden kurtuldum (*whatsapp* is crashed so I'm little relieved, got rid of notifications)	positive	negative

Table 1: Sample tweets from the dataset. Targets are shown in italics. Sentences are annotated with respect to overall sentence sentiment and targeted sentiment which represent the sentiment towards the target. English translations are provided in parenthesis.

2 Related Work

One of the challenges of targeted sentiment analysis is identifying contexts associated with target words in the sentiment classification. Early methods for understanding the relationship between the target and the rest of the sentence rely on handcrafted feature extractions and rule-based techniques (Ding et al., 2008; Jiang et al., 2011). Recurrent neural networks (RNN) have been implemented for sentiment analysis in the recent years. It achieved improved results compared to earlier methods (Dong et al., 2014; Nguyen and Shirai, 2015; Baktha and Tripathy, 2017). Two RNNs are used to obtain the context from both left and right and combine the context knowledge in (Tang et al., 2016). Attention mechanisms are recently added into RNN-based methods to model the connection between each word and the target (Wang et al., 2016; Ma et al., 2017; Zhang et al., 2020).

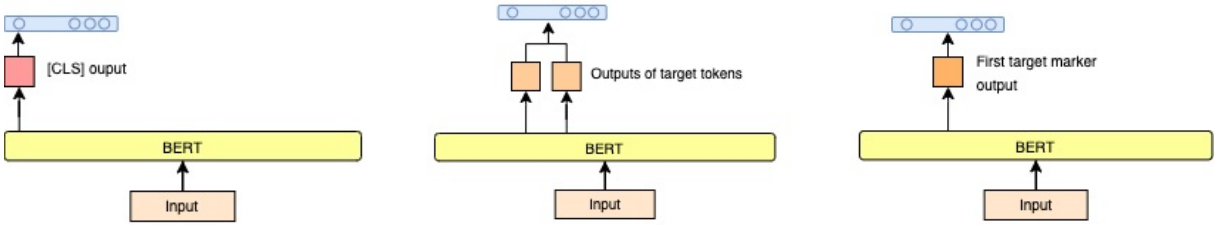
Vaswani et al. (2017) introduced the transformer architecture consisting of encoder and decoder blocks based on self-attention layers. Bidirectional Encoder Representations from Transformers (BERT) has been introduced and shown to achieve the state-of-the-art in various NLP tasks in (Devlin et al., 2019). BERT has recently become a widely used approach for sentiment analysis in many languages (Sun et al., 2019; Li et al., 2019). Köksal and Özgür (2021) provide a Twitter dataset in Turkish for sentiment analysis called BounTi. It consists of Twitter data which are about predefined universities and manually annotated by considering sentimental polarities towards these universities. They propose a BERT model fine-tuned using the BounTi dataset to identify sentiment in Turkish tweets.

3 Dataset

Twitter is a commonly used source of sentiment classification dataset in the literature (Jiang et al., 2011; Severyn and Moschitti, 2015; Kruspe et al., 2020). In this study, we also create a Twitter dataset with 3952 tweets whose timestamps span a six-month period between January 2020 and June 2020. The tweets are collected via the official Twitter API by separately searching our 6 targets selected from famous companies and brands.

This dataset is manually annotated with three labels, positive, negative, and neutral. Two factors are considered in the annotation process, namely sentence sentiment and targeted sentiment. Each tweet has the following two labels. The sentence sentiment label expresses the overall sentiment of the sentence, regardless of the target word, as in traditional sentiment analysis techniques. On the other hand, the targeted sentiment label reflects the sentiment for the target in that sentence. The collected tweets are annotated separately by two annotators (one of the authors and a volunteer annotator) who are native Turkish speakers. Cohen's κ (Cohen, 1960) is used to demonstrate inter-annotator agreement and is calculated as 0.855. In case of conflict between annotators, they re-evaluated the conflicting tweets. After re-evaluation, tweets on which the annotators agree are retained and conflicting tweets are removed from the dataset.

Table 1 shows example sentences from the dataset. The first tweet is a positive comment about the target and the sentence is also positive overall. The second tweet indicates a negative opinion about the target, since it has stated as crashed, although the sentence expresses a positive situation overall. Both sentence and targeted sentiment are the same for most of the tweets as in the first example. Only in 21% of the tweets, targeted sentiment differs from the overall sentence sentiment. This means that the rest of the dataset is similar to a standard sentiment analysis dataset. The number of negative tweets in the dataset is significantly higher than the number of positive and neutral tweets for



(a) [CLS] output

(b) Max-pooling outputs of target tokens

(c) The output of the first target marker

Figure 1: An overview of architectures to get and handle outputs from BERT

each target. The strikingly high number of negative tweets may be caused by the tendency of customers to write a review when they have had a bad experience. The total percentages of positive, negative and neutral classes are 19%, 58% and 23%, respectively. The dataset is randomly divided into train, test, and validation sets by 65%, 20% and 15%, respectively. The distribution of labels for each subset is kept similar to the distribution of labels for the entire dataset.

The dataset contains ungrammatical text, slang, acronyms, as well as special Twitter characters. During pre-processing URLs and mentions (@) are deleted. Hashtag signs (#) are removed, but hashtags are kept for two reasons: hashtags have been shown to express sentiment (Alfina et al., 2017; Celebi and Özgür, 2018) and some tweets contain the targets as hashtags.

4 Methodology

Baldini Soares et al. (2019) has introduced a novel method with transformer structure in the field of relation extraction. The key idea behind this work is to tag the entities with additional tokens before feeding the BERT model with the input. Different combinations of input and output types are evaluated. The best results are obtained when entity markers are added to the input and when the output of the starting entity markers are taken as the output from BERT. Motivated by the results of Baldini Soares et al.'s work, this paper evaluates several BERT architectures with different input and output techniques for the targeted sentiment analysis task.

Two input representation techniques are investigated. In the standard input representation, the inputs are simply entered into the model without modification. In the second input representation approach, the targets are highlighted by adding additional special target tokens [TAR] at the beginnings and ends of targets, as shown in Table 2. These target tokens are expected to collect information about the target, just as the [CLS] token collects knowledge about the entire sentence. The three approaches for outputs explored in this study are shown in Figure 1. The [CLS] output approach uses only the output of the first token from the last hidden state of BERT, as proposed for classification in the original paper (Devlin et al., 2019). In the second approach, the outputs of the tokens originating from the target, including the outputs of the [TAR] tokens, are max-pooled. The first target marker approach considers only the output of the first [TAR] token in the input instead of the output of the standard [CLS]. All output approaches utilize a softmax layer at the end for classification.

Tweets with [TAR] tokens
[TAR]*whatsapp*[TAR] çöktü de biraz rahatladım bildirimlerden kurtuldum
([TAR]*whatsapp*[TAR] is crashed so I'm little relieved, got rid of notifications)
[TAR]*coca cola*[TAR] daha iyi lezzet olarak
([TAR]*coca cola*[TAR]'s taste is better)

Table 2: Example tweets with target marker representation

4.1 Model Descriptions

First, two baseline models are defined in order to show the drawbacks of the traditional SA models. One baseline is the BERT-based BounTi model (Köksal and Özgür, 2021). The second baseline is also a BERT-based traditional SA model, but fine-tuned with our new dataset using sentence sentiment. Both have similar architectures and use the [CLS] output for sentiment classification.

Four other variants of BERT-based models are proposed for targeted sentiment analysis. **T-BERT** is a model with a similar architecture to our base-

Model	F1-Score
Baseline Model	0.591
BounTi Model	0.498
T-BERT	0.610
T-BERT$_{\text{marked}}$	0.659
T-BERT$_{\text{marked}}$-TS	0.653
T-BERT$_{\text{marked}}$-MP	0.669

Table 3: Performance of all models for TSA with test dataset against targeted sentiment labels

Model	F1-Score
Baseline Model	0.256
BounTi Model	0.233
T-BERT	0.401
T-BERT$_{\text{marked}}$	0.428
T-BERT$_{\text{marked}}$-TS	0.459
T-BERT$_{\text{marked}}$-MP	0.444

Table 4: Performance of all models for TSA with data whose targeted and sentence sentiment are different.

line models. It makes no changes to the input and takes its output from the [CLS] token. The main difference is that targeted sentiment labels are used in the training phase. Therefore, the model is trained to learn targeted sentiment, whereas the baseline models are not aware of the target. **T-BERT$_{\text{marked}}$** employs only the target marker representation on top of T-BERT and adds [TAR] tokens into the input. [TAR] token is introduced to BERT's tokenizer and the vocabulary is resized. Hence, the tokenizer accepts [TAR] as one of its special tokens such as [SEP]. **T-BERT$_{\text{marked}}$-MP** is another model with target marker representation, additionally it max-pools all outputs of target tokens. **T-BERT$_{\text{marked}}$-TS** also utilizes target markers. However, it takes its output only from the first target token [TAR] unlike T-BERT$_{\text{marked}}$-MP.

In the training phase of all models, BERTurk (Schweter, 2020) is chosen as the base BERT model. Class weights are set inversely proportional to the class distribution to reduce the effects of an unbalanced data set. The batch size is chosen as 24. Hyperparameters like weight decay, learning rate, and warm-up steps are selected as 0.1, $1e-5$, and 300 respectively. As optimizer, AdamW is used.

5 Results

All proposed BERT variants and baselines are evaluated for targeted sentiment analysis over our introduced dataset. Macro averaged F1-Score is used as the evaluation metric in these experiments. The results are presented in Table 3. All targeted BERT variants outperform both baseline models for TSA. T-BERT$_{\text{marked}}$-MP achieves the best results with 67% F1-score, while T-BERT is relatively the worst performing targeted model with 61% F1-score. T-BERT$_{\text{marked}}$-TS and T-BERT$_{\text{marked}}$ obtain performance quite close to each other, the difference between those models is insignificant. They both have approximately 65% F1-scores.

Only 21% of the dataset has different sentence and targeted sentiment. These portion of data can demonstrate the distinction between targeted and sentence sentiment classification better. If both labels are the same, then traditional SA models may seem to accurately predict targeted sentiment. However, such sentences do not show how accurate the predictions from neither TSA nor SA models are. For this reason, a subset of our dataset such that all sentences have different targeted and sentence sentiment is used for another round of experiments. Table 4 shows the results for the TSA task with this subset. Baseline models' F1-score decreases dramatically to 25%, and it's 23% for BounTi model. Targeted BERT model with the lowest score (40% F1-score) outperforms both models. T-BERT$_{\text{marked}}$-TS achieves better targeted sentiment predictions with 46% F1-score. T-BERT$_{\text{marked}}$-TS improves the baseline performance by 79% on F1-score.

6 Discussion

Our results suggest that target oriented models can significantly improve the performance for targeted sentiment analysis. BERT architectures that perform successfully in the relation extraction field are shown to be successful for the targeted sentiment analysis task. Target markers make BERT models understand target related context better compared to the [CLS] token. All three models with target markers outperform the baselines and T-BERT. Hence, adding target markers is an effective approach for improving TSA performance.

T-BERT$_{\text{marked}}$-TS and T-BERT$_{\text{marked}}$-MP are shown to perform slightly better than the other target oriented models. The common aspect of these models, apart from the target tokens, is that they both focus on the outputs of the target-related tokens rather than the [CLS] tokens. Therefore, it can be concluded that target outputs improves the performance for the TSA task.

We only considered one target in each sentence and annotated according to that target. Other targets in the sentence, if any, are ignored. Multiple targets with conflicting targeted sentiment in the same sentence can be a problem to consider. There are cases where a sentence has more than one target, and each target has a different targeted sentiment. For example, in a comparison, the sentiment toward one target may actually depend on the sentiment of another target in the same sentence. In this work, the scope is limited to only one target in each sentence. Target markers are also used only for this one target in the sentence and other possible targets are ignored. The lack of proper treatment of such cases in this work may affect the performance of all models.

Sentence and targeted sentiment are identical for 79% of the dataset. Thus, if a traditional SA model, which is designed to predict the overall sentence sentiment, is used for the TSA task, its success for this task would be overestimated. The results demonstrate that targeted sentiment analysis models perform significantly better than traditional sentiment analysis models on the TSA task. However, the performance of the TSA models increases when they are tested on the entire test dataset, rather than on a subset containing only tweets with different sentence and targeted sentiment labels. This highlights that they may still be biased in favor of sentence sentiment to some extent.

7 Ethical Considerations and Limitations

The dataset contains public tweets in Turkish that are provided by the official Twitter API for research. Only tweet ID's and labels of the tweets are shared publicly to follow Twitter's terms and conditions. The annotators have no affiliation with any of the companies that are used as targets in the dataset, so there is no potential bias due to conflict of interest.

The models developed in this work are not yet satisfactory to use their results without human monitoring. It is recommended to manually check the predictions of these models before using them.

8 Conclusion and Future Work

We presented a manually annotated Turkish Twitter dataset specifically created for targeted sentiment analysis and is also suitable for the traditional sentiment analysis task. This allowed us to develop and evaluate novel models for targeted sentiment analysis in a low-resource language such as Turkish.

We adapted and investigated BERT-based models with different architectures for targeted sentiment analysis. Experiments show significant improvement on baseline performance.

As future work, we plan to expand our dataset so that it contains more sentences with different sentence and targeted sentiment. Moreover, novel methods for sentences with multiple targets will be investigated.

Acknowledgements

We would like to thank Abdullatif Köksal for helpful discussions and Merve Yılmaz Mutlu for annotations. GEBIP Award of the Turkish Academy of Sciences (to A.Ö.) is gratefully acknowledged.

References

Ika Alfina, Dinda Sigmawaty, Fitriasari Nurhidayati, and Achmad Nizar Hidayanto. 2017. Utilizing hashtags for sentiment analysis of tweets in the political domain. In *Proceedings of the 9th international conference on machine learning and computing*, pages 43–47.

Kiran Baktha and BK Tripathy. 2017. Investigation of recurrent neural networks in the field of sentiment analysis. In *2017 International Conference on Communication and Signal Processing (ICCSP)*, pages 2047–2050. IEEE.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, page 107134.

Arda Celebi and Arzucan Özgür. 2018. Segmenting hashtags and analyzing their grammatical structure. *Journal of the Association for Information Science and Technology*, 69(5):675–686.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, Portland, Oregon, USA. Association for Computational Linguistics.

Mesut Kaya, Guven Fidan, and Ismail Toroslu. 2012. Sentiment analysis of turkish political news. pages 174–180.

Abdullatif Köksal and Arzucan Özgür. 2021. Twitter dataset and evaluation of transformers for turkish sentiment analysis. In *2021 29th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Anna Kruspe, Matthias Häberle, Iona Kuhn, and Xiao Xiang Zhu. 2020. Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4068–4074.

Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514, Lisbon, Portugal. Association for Computational Linguistics.

Jiaxin Pei, Aixin Sun, and C. Li. 2019. Targeted sentiment analysis: A data-driven categorization. *ArXiv*, abs/1905.03423.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training deep convolutional neural network for Twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469, Denver, Colorado. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Ji Zhang, Chengyao Chen, Pengfei Liu, Chao He, and Cane Wing-Ki Leung. 2020. Target-guided structured attention network for target-dependent sentiment analysis. *Transactions of the Association for Computational Linguistics*, 8:172–182.