PoliStance-TR: A Dataset for Turkish Stance Detection in Political Domain

Muhammed Cihat Ünal
EPAM Systems
muhammedcihat.unal@epam.com

Yasemin Sarkın
Atılım University
sarkin.yasemin@student.atilim.edu.tr

Alper Karamanlıoğlu
TURKSAT Inc.
alper.karamanlioglu@turksat.com.tr

Berkan Demirel
TURKSAT Inc.
berkan.demirel@turksat.com.tr

Abstract

Stance detection in NLP involves determining whether an author is supportive, against, or neutral towards a particular target. This task is particularly challenging for Turkish due to the limited availability of data, which hinders progress in the field. To address this issue, we introduce a novel dataset focused on stance detection in Turkish, specifically within the political domain. This dataset was collected from X (formerly Twitter) and annotated by three human annotators who followed predefined guidelines to ensure consistent labeling and generalizability. After compiling the dataset, we trained various transformer-based models with different architectures, showing that the dataset is effective for stance classification. These models achieved an impressive Macro F1 score of up to 82%, highlighting their effectiveness in stance detection.

1 Introduction

Stance detection involves determining whether the expressed opinion in a text is supportive (*Favor*), opposing (*Against*), or neutral (*Neutral*) toward a specific target, such as an organization, movement, product, or individual. Unlike sentiment detection, which focuses on emotional tone (*Positive*, *Negative*, *Neutral*), stance detection seeks to identify viewpoints on particular issues, making annotation more complex.

This study introduces a benchmark dataset for Turkish stance detection by collecting 8,000 tweets from X (formerly Twitter) using targeted political keywords. Social media platforms like X offer valuable insights into public perspectives on topics such as elections and reforms due to their large volume of user-generated content. Details of the keyword-based data collection process are provided in Section 3.1.

Previous work by Küçük (2017) focused on Turkish stance detection within tweets about two football clubs, initially labeling 700 tweets without the *Neither* category. This dataset was later expanded to 1,065 tweets in Küçük and Can (2018). In contrast, our dataset broadens the scope by avoiding domain-specific labels to create a generalized stance detection dataset. All common stance detection labels (*Favor*, *Against*, *Neutral*) were utilized in our experiments.

The annotation process focused on identifying stance (*Favor*, *Against*, or *Neutral*) without relying on pre-defined targets like specific individuals or organizations. This approach minimizes bias and ensures flexibility across diverse topics. Details are provided in Section 3.3.

To evaluate the quality and utility of our dataset, we used two approaches. First, we tested the Llama 3 model (AI@Meta, 2024) in a zero-shot setting to assess its ability to detect stance without specialized training. Second, we fine-tuned six pre-trained BERT models (Devlin et al., 2019) and evaluated their performance on test samples (details in Section 4). Model performance was measured using both accuracy and Macro F1 scores to address potential class imbalance.

Figure 1 outlines the complete methodology, from data collection to evaluation, as detailed in subsequent sections.

Our contributions are as follows: 1) We present a new Turkish Stance Detection dataset with 8,000 tweets, sampled from over 5.5 million tweets on X. This dataset, over eight times larger than the previous benchmark (Küçük and Can, 2018), introduces additional challenges such as linguistic complexity, which we describe and analyze comprehensively. 2) We fine-tuned Transformer-based models, pre-trained on Turkish corpora, for stance detection. Our structured annotation rules enable these models to handle challenging linguistic phenomena, such as sarcasm, while reducing reliance on sentiment cues that often lead to classification

Proceedings of Recent Advances in Natural Language Processing,pages 1282–1288
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-148

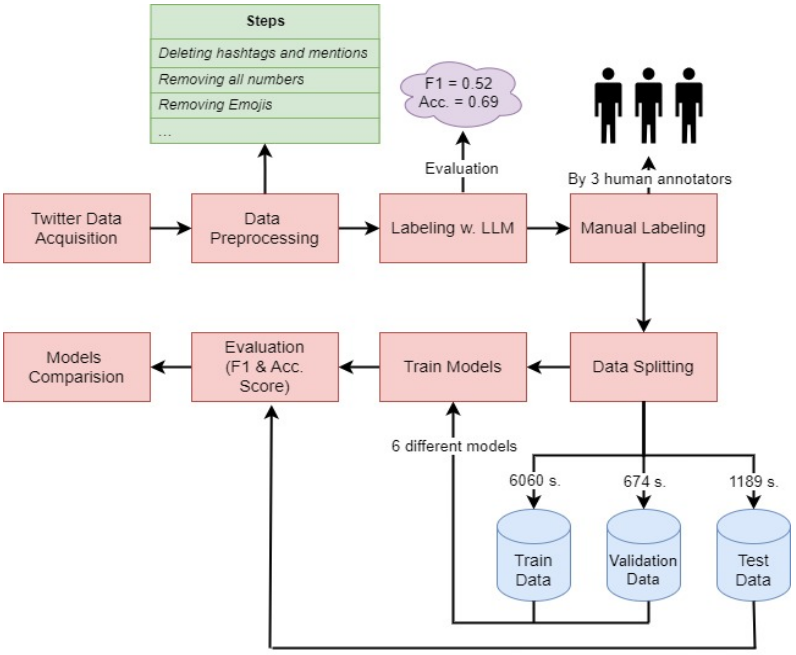

Figure 1: Overview of the data processing pipeline, including data collection, preprocessing, labeling, model training, and evaluation. Values (e.g., 6060, 674, 1189) represent the number of samples (s.) in each dataset split.

errors. To facilitate reproducibility and promote further research, we publicly release the Turkish stance detection dataset[1].

2 Related Work

In recent years, analyzing politically oriented social media data for stance detection has gained significant attention, with efforts focused on determining contributors' political leanings. During this time, various stance detection datasets and methods have been developed. In this section, we review key datasets, highlight their differences, and compare them with our work.

Küçük (2017) introduced the first Turkish stance detection dataset, collected from X, using two labels (*Favor*, *Against*) and focusing on tweets about two popular Turkish football teams. This dataset initially included 700 tweets, later expanded to 1,065 (Küçük and Can, 2018). Unlike our dataset, which includes 8,000 tweets with three labels (*Favor*, *Against*, *Neutral*), their dataset annotated both domain names and stance labels during the process. While their work remains a seminal resource, our larger dataset complements it by addressing broader topics. Joint training with both datasets is a promising direction for future research.

Mohammad et al. (2016) published the Stance Dataset, consisting of 4,870 tweet-target pairs across six topics (e.g., Atheism, Climate Change, Hillary Clinton). It was developed for SemEval-2016 Task 6 and includes sentiment labels, enabling the study of sentiment-stance relationships. This domain-specific dataset, which relied on tools like lexicons and n-grams, has been instrumental in advancing research in textual inference and stance analysis.

Li et al. (2021) and Sobhani et al. (2017) introduced benchmark datasets targeting political tweets. Li et al. (2021) created a 21,574-tweet dataset centered on challenging tasks like cross-target stance detection, with BertTweet achieving the highest F1 score (80%). Sobhani et al. (2017) developed a 4,455-tweet dataset from the 2016 U.S. elections for multi-target stance detection, incorporating innovative methods like self-attention and cascading classification. Both datasets offer unique challenges and insights for political stance detection.

3 Building the Dataset

In this section, we explain how the dataset was generated, the rules we followed during annotation, and the data distribution.

[1] https://anonymous.4open.science/r/polistance-tr-22C4/

3.1 Data Collection

We collected the dataset from X (formerly Twitter) using the *tweet-harvest* [2] library, which facilitates bulk data collection and analysis from the platform. The *tweet-harvest* library is designed to streamline and manage the data collection process more efficiently by utilizing the X API. It connects to X APIs to gather tweets based on specific users, hashtags, or keywords and allows for the application of time-frames or filters. The library stores the collected tweets in formats suitable for content analysis. The data is typically returned in JSON format, which can be processed with various programming languages.

When collecting the dataset, we selected tweets from the 2023 election period in Türkiye, extending back ten years. To filter the data, we used the official Twitter accounts of prominent political figures and institutions, including "Recep Tayyip Erdoğan" (President of Türkiye), "Ak Parti" (Justice and Development Party), "CHP" (Republican People's Party), and "İçişleri Bakanlığı" (Ministry of Interior), among others. These keywords were chosen to capture a broad spectrum of politically relevant discourse, including both party-specific and institutional narratives. For instance, the Ministry of Interior was included due to its central role in election security and public administration, which often makes it a focal point in political discussions.

As a result, the dataset predominantly consists of politically inclined content, but this does not imply that it entirely lacks non-political data since we removed hashtags and account names from the text. Consequently, there are instances where the model may not be able to determine the stance, if any, towards a specific target. Thus, we aim to have a more generalized model after training.

3.2 Data Preprocessing

After collecting the raw data with *tweet-harvest*, we observed that some samples included noisy or inappropriate text, which could negatively affect the performance of the models. To address this issue, we implemented several filtering steps, which are as follows:

- Converting all tweets to lowercase, depending on the model,
- Removing emojis,
- Expanding abbreviations to their full forms,
- Retaining punctuation marks that may influence stance evaluation (e.g., !, ?, ..., %) while reducing multiple instances of the same punctuation mark to a single occurrence,
- Removing other punctuation marks (e.g., *, +, /),
- Deleting hashtags and mentions from the tweets,
- Removing all numbers,

Additionally, we performed rare-word and common-word analysis to reduce noise further and enhance overall model performance while making the data more manageable. Finally, we deleted any resulting null texts, as some tweets consisted solely of emojis and became empty following the preprocessing steps mentioned above.

3.3 Data Annotation

The annotation process was conducted by three human annotators, each working independently on a distinct batch of tweets with no overlap. The dataset was labeled with "Favor," "Against," and "Neutral," where "Neutral" encompasses both No-Stance and Neutral Stance.

Annotators assessed the text without relying on predefined targets (e.g., specific individuals, organizations, or topics). Instead, they focused on linguistic cues in the text to determine the author's overall stance. This approach ensured objectivity and generalizability across diverse topics, including politics, terrorism, sports, and the economy.

Prior to the annotation task, eight specific guidelines were provided to the annotators to reduce bias towards the tweets, ensure consistency and minimize errors during the tagging process. These guidelines served as critical references to address potential discrepancies among annotators. The rules followed by the annotators are outlined as follows:

1. Texts with a news value, questions, requests, proverbs or idioms, or advice are categorized as *neutral*,
2. If, after processing, the text contains only the name of an institution or organization or includes predicates that do not convey a stance, it is classified as *neutral*,
3. Texts expressing goodwill, prayers, or respect are categorized as *favor*,

[2] github.com/helmisatria/tweet-harvest

4. Sarcastic texts are closely analyzed, and if the stance remains unclear, they are labeled as *neutral*,
5. Texts that clearly express goodwill are labeled as *favor*, while those expressing malice are labeled as *against*,
6. Texts containing complaints, grievances, or warnings are labeled as *against*,
7. Texts containing references to terrorism or support for terrorism are classified as *against*,
8. For texts with multiple stances, the stance is determined based on the segment following the conjunction.

For further details, check the appendix, which presents text samples and their corresponding labels aligned with specific annotation guidelines.

Upon completing the annotations, the annotators collectively reviewed their work to ensure consistency and reliability, reaching a consensus on the final labels. In cases where discrepancies arose, the annotators applied majority voting, as recommended by (Li et al., 2021).

3.4 Data Distribution

Our data is split into three parts: 6060 train data, 674 validation data, and 1189 test data. While doing so, we have kept the percentage of labels in each set of data splits the same. The total number of labels and their distribution is as follows: 2898 Favor, 2858 Against and 2167 Neutral.

Label	Train	Test	Validation
Favor	2216	435	247
Against	2186	429	243
Neutral	1658	325	184
Total	6060	1189	674

Table 1: The distribution of "Favor," "Against," and "Neutral" categories is shown for the Train, Test, and Validation sets, along with the totals for each category.

4 Methodology

To conduct our experiments, we employed six transformer-based models—BERT, DistilBERT, ALBERT, ConvBERT, ELECTRA, and XLM-RoBERTa—chosen for their proven effectiveness in classification tasks. We focused on transformer architectures due to their ability to handle Turkish agglutination via sub-word tokenization and capture long-range dependencies through self-attention, both of which are crucial for stance detection. To ensure language-specific relevance, we used pre-trained Turkish versions of these models, sourced from publicly available repositories, including the *Turkish BERT* project (Schweter, 2020) and other community-contributed models.

We excluded larger variants like BERT-large to emphasize architectural differences rather than model size. The models were grouped by complexity: ConvBERT, ELECTRA, and XLM-RoBERTa as heavyweight; BERT as middleweight; and the rest as lightweight.

There are established benchmark datasets for stance detection in Turkish, notably the one introduced by Küçük (2017), discussed in Section 1. This domain-specific dataset contains 700 manually labeled tweets, each referencing either *Galatasaray* or *Fenerbahçe*, with stance labels evenly split between 'Favor' and 'Against'. The authors reported an average F1 score of 77.5 using SVM classifiers with unigrams and hashtag features, and 76.4 using unigram-only SVMs. Although we intended to compare our dataset with this benchmark, we were unable to access the full tweet content, limiting direct evaluation.

4.1 Training Details

We previously discussed the data split into training, validation, and test sets. Based on this split, we began training the baseline methods mentioned earlier. Due to memory limitations, all models were trained with a batch size of 32, utilizing two T4 GPUs. The number of epochs was set to 10 for both training and validation. We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 5e-5. Given that our dataset contains three labels, we selected CrossEntropyLoss as the appropriate loss function. For evaluation metrics, we opted for Macro-F1 to obtain a comprehensive understanding of the performance of the models.

5 Experimental Results

Following the completion of the annotation process, we evaluated the dataset labels against predictions generated by Llama 3.0. This zero-shot evaluation yielded an F1 score of 0.51 and an accuracy of 0.69, establishing a baseline for subsequent model training. The zero-shot inference parameters for this experiment are presented in Table 3.

Scores	Llama3.0	BERT
F1	0.52	0.83
Accuracy	0.69	0.82

Table 2: Comparison of standalone Llama3.0 and our best fine-tuned model based on F1 and Accuracy scores.

Parameter	Value
Model variant	Meta-Llama-3-8B
Temperature	0.0 (greedy)
Max tokens	128
Context window	4 096
Hardware	1 × RTX 4090 40 GB

Table 3: Zero-shot inference settings for the Llama-3.0 baseline.

Fine-tuned models significantly outperformed this baseline. Among them, BERT and ELECTRA achieved the highest scores (Macro F1 and accuracy of 0.82), demonstrating the reliability of our dataset for stance detection. ConvBERT followed closely with strong performance, while DistilBERT delivered competitive results despite its lightweight design. XLM-RoBERTa and RoBERTa also performed well, illustrating the dataset's adaptability across multilingual and general-purpose models. In contrast, ALBERT produced the lowest scores, suggesting limited effectiveness in this context. Table 4 provides a summary of the performance metrics for each model.[3]

Figure 2 highlights that BERT performs well on the *Favor* and *Against* labels but struggles with the *Neutral* label, often misclassifying texts referencing emotionally charged events. Addressing this challenge may require more extensive training data and targeted optimization strategies for the *Neutral* category.

5.1 Neutral difficulty

The confusion matrix (Figure 2) confirms that most residual errors involve the *Neutral* label, reinforcing our qualitative observation that irony, hedging, and concessive markers (e.g. "ama", "gerçi") are harder for models to interpret than overt sentiment words.

5.2 Inter-Annotator Agreement

Inter-Annotator Agreement (IAA) measures the consistency among annotators when labeling data, ensuring the reliability of annotations—especially critical in machine learning and NLP tasks.

[3]Details about these models can be found on Hugging Face's model hub: huggingface.co/byunal/models.

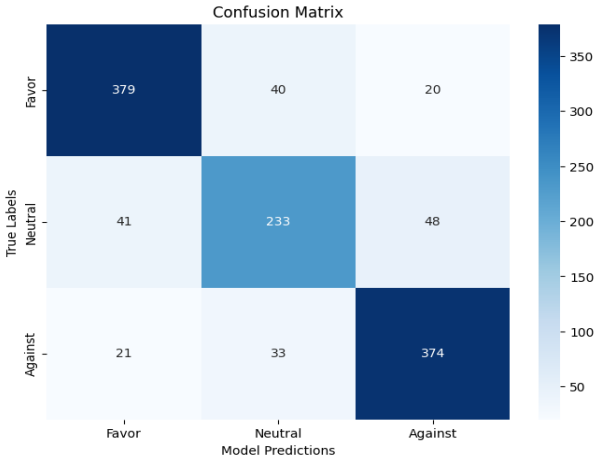

Figure 2: Confusion matrix summarizing the model's performance, including correct and incorrect classifications for the *Favor*, *Neutral*, and *Against* labels.

We used Fleiss's kappa coefficient to assess agreement among multiple annotators. This metric evaluates whether observed agreement exceeds chance levels. Based on the interpretation by Nichols et al. (2010), our score of 0.78 falls within the 0.61–0.80 range, indicating substantial agreement:

- 0.81–1.00: Almost perfect,
- 0.61–0.80: Substantial,
- 0.41–0.60: Moderate,
- 0.21–0.40: Fair,
- 0.00–0.20: Slight,
- <0.00: Poor or none,

This result reflects the annotators' careful adherence to labeling guidelines and supports the overall reliability of the dataset.

6 Conclusion and Future Work

In this paper, we introduced a new Turkish stance detection dataset comprising 8,000 tweets with balanced stance distributions. Our experiments demonstrated that transformer-based models effectively learn stance detection patterns from our dataset, with rigorous preprocessing and annotation guidelines contributing significantly to model performance. Notably, even without fine-tuning, Llama 3.0 demonstrated reasonable capability on our test set, suggesting promising directions for zero-shot stance detection.

Models	Macro F1	Precision	Recall	Accuracy
BERT	**0.82**	**0.83**	**0.82**	**0.82**
ALBERT	0.75	0.76	0.76	0.75
DistilBERT	0.77	0.77	0.77	0.77
ConvBERT	0.81	0.82	**0.82**	**0.82**
ELECTRA	**0.82**	0.82	**0.82**	**0.82**
XLM-RoBERTa	0.77	0.77	0.77	0.77
RoBERTa	0.78	0.79	0.78	0.78

Table 4: Evaluation scores across used models. The best results are highlighted in bold.

A key strength of our dataset is its potential transferability. Although the tweets were harvested with political keywords, many instances appear domain-independent, hinting that models trained on POLISTANCE-TR could generalise beyond politics. Quantitatively verifying this cross-domain capability—e.g., by measuring lexical drift or testing on non-political stance tasks—remains important future work.

For future work, we plan to expand the dataset to include emerging political discourse while investigating cross-domain generalization through evaluation on non-political stance detection tasks. We also aim to explore specialized architectures better suited to capturing the nuanced linguistic markers of stance in Turkish text. This dataset represents a valuable contribution to Turkish NLP resources, addressing the scarcity of stance detection benchmarks in languages beyond English.

7 Ablation Study

We conducted an ablation study to evaluate the impact of two key components in our pipeline: data preprocessing and annotation guidelines. Removing preprocessing steps reduced model performance by 2% - 3%, with numeric content normalization having the largest impact. This highlights the importance of preprocessing for improving stance detection accuracy in Turkish social media text.

For annotation, we also tested a scenario where annotators performed labeling without the predefined guidelines. In this case, inter-annotator consistency dropped by up to 15%, showing that structured annotation rules are essential for creating reliable stance detection datasets. These findings underline the significance of both careful preprocessing and systematic annotation in our pipeline.

8 Limitations

This study has several limitations. First, the focus on Turkish-language content (8,000 tweets) introduces unique linguistic challenges but limits generalizability to other languages. Despite a reasonable class distribution (see Table 1), the dataset size is modest.

Second, stance annotation, even with high inter-annotator agreement (see Section 5.2) and clear guidelines, remains subjective, particularly for texts with sarcasm or implicit stance markers. Classifying certain content types, such as news items and questions, as neutral may oversimplify their stance representation.

Lastly, the scope was restricted to transformer-based models (see Section 5), leaving alternative approaches unexplored. Rapidly evolving social media discourse may further degrade model performance over time, particularly for systems relying solely on textual content without broader contextual understanding.

Ethics / Legal Statement

Tweets were gathered with the v2 *Academic Research API* and are released *only* as Tweet IDs, fully compliant with the X/Twitter Developer Policy (X Corp., 2024b). All IDs are "rehydrated" prior to distribution (X Corp., 2024a) and any deleted or protected posts are removed, so withdrawn content is never redistributed. No tweet text or profile metadata is shared, keeping re-identification risk minimal under GDPR Recital 26 (European Parliament and Council, 2018). The annotation layer (stance labels, splits, guidelines) is released under CC-BY-NC-4.0 (Creative Commons, 2013); any commercial use of tweet content must separately meet X/Twitter's terms [4].

[4]IRB approval: protocol #2024-SOC-017.

References

AI@Meta. 2024. Llama 3 model card.

Creative Commons. 2013. Creative commons attribution–noncommercial 4.0 international. https://creativecommons.org/licenses/by-nc/4.0/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

European Parliament and Council. 2018. General data protection regulation—recital 26. https://gdpr-info.eu/recitals/no-26/.

Dilek Küçük. 2017. Stance detection in turkish tweets. *arXiv preprint arXiv:1706.06894*.

Dilek Küçük and Fazli Can. 2018. Stance detection on tweets: An svm-based approach. *arXiv preprint arXiv:1803.08910*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Thom Nichols, Paola Wisner, and Gary Gulabchand. 2010. Putting the kappa statistic to use. *Quality Assurance Journal*, 13:57–61.

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

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

X Corp. 2024a. Guide to rehydrating tweets with the twitter api v2. https://developer.twitter.com/en/docs/twitter-api/tweets/lookup/introduction. Accessed 2025-05-25.

X Corp. 2024b. X developer agreement and policy: Content redistribution. https://developer.x.com/en/developer-terms/agreement-and-policy. Accessed 2025-05-25.