# MARASTA: A Multi-dialectal Arabic Cross-domain Stance Corpus

**Anis Charfi***, **Mabrouka Bessghaier***, **Andria Atalla***, **Raghda Akasheh***, **Sara Al-Emadi***, **Wajdi Zaghouani**†
*Carnegie Mellon University in Qatar, †Hamad Bin Khalifa University
Education City, Doha, Qatar

**Abstract**

This paper introduces a cross-domain and multi-dialectal stance corpus for Arabic that includes four regions in the Arab World and covers the main Arabic dialect groups. Our corpus consists of 4657 sentences manually annotated with each sentence's stance towards a specific topic. For each region, we collected sentences related to two controversial topics. We annotated each sentence by at least two annotators to indicate if its stance favors the topic, is against it, or is neutral. Our corpus is well-balanced concerning dialect and stance. Approximately half of the sentences are in Modern Standard Arabic (MSA) for each region, and the other half is in the region's respective dialect. We conducted several machine-learning experiments for stance detection using our new corpus. Our most successful model is the Multi-Layer Perceptron (MLP), using Unigram or TF-IDF extracted features, which yielded an F1-score of 0.66 and an accuracy score of 0.66. Compared with the most similar state-of-the-art dataset, our dataset outperformed in specific stance classes, particularly "neutral" and "against".

**Keywords:** Natural Language Processing, dataset, stance detection, polarization, Arabic language

## 1. Introduction

According to Biber and Finegan (1988), the word stance refers to the writer expressing personal feelings, assessments, and judgments about a certain message or topic. Stance detection is a sub-task that branches out from sentiment analysis, which aims to determine the author's attitude towards a target (Küçük and Can, 2021), often stating whether this author is against, in favor, or neutral towards it (Li and Caragea, 2019). Stance detection has emerged as a powerful tool for analyzing and interpreting the public's online opinions (Cao et al., 2022).

With the widespread use of social media, stance detection has become increasingly important in various fields. This includes conducting analytical research to measure the public opinion on social, religious, and political issues, as well as using it in veracity checking applications, and determining the level of controversy on social media platforms (AlDayel and Magdy, 2021). Moreover, stance detection has been used in scenarios where individuals rely on Web searches to gather information for making critical decisions, such as adopting a vegan lifestyle (Draws et al., 2023) or joining a specific social movement. These decisions can influence people's openness to different perspectives and prompt them to change their viewpoints completely (Mckay et al., 2020); (Flaxman et al., 2016).

Stance detection is also vital for detecting polarization, which happens when there is a strong predisposition towards one party and a strong dislike or prejudice towards the other (McCarty, 2019). Detecting stance for identifying polarization has been used in various applications, such as detecting polarization on climate change on social media (Upadhyaya et al., 2023). Determining stance could be a prerequisite for assessing the level of polarization. At a high level, polarization detection enables further understanding of social media groups in locating the presence of propaganda and radicalism (Bail et al., 2018). Once polarization is detected, one could raise the topic in the public's discussion to understand the different polarized sides and, in some situations, attempt to establish a middle ground between them.

In this paper, we present a novel cross-domain and multi-dialectal stance corpus for Arabic called MARASTA, which we developed in the context of a collaborative research project funded by the Qatar National Research Fund (QNRF). Our corpus includes over 4,500 sentences annotated concerning their stance toward a topic by at least two annotators. The corpus covers four important dialectal regions in the Arab world: Maghreb, Egypt, the Gulf, and the Levant. Arabic dialects vary significantly; each dialect has unique linguistic features, syntax, and vocabulary. For each region, we identified two controversial topics and collected sentences covering these topics. For each sentence, our annotators manually indicated if the sentence's stance favored the topic, against it, or neutral. For each topic, we ensured that half of the sentences were in their respective region's dialect and the other half were in Modern Standard Arabic (MSA). Our corpus is well-balanced concerning both stance and dialect.



*LREC-COLING 2024*, pages 11060–11069

The remainder of this paper is organized as follows—section 2 reports on existing stance corpora for Arabic. Section 3 presents our proposed stance corpus and the process we followed for data collection and annotation. Section 4 reports on the experiments we conducted for stance detection based on our corpus. We also evaluated the performance of our models using a similar stance corpus. Section 5 concludes this paper and outlines directions for future work.

## 2. Related Work

Several studies have focused on creating and analyzing Arabic corpora for various NLP tasks. Ahmed et al. (2022) and Zaghouani (2014) presented two surveys of freely available Arabic corpora. Rosso et al. (2018) provided a comprehensive survey on author profiling, deception, and irony detection for the Arabic language. Charfi et al. (2019) introduced a fine-grained annotated multi-dialectal Arabic corpus, while Rangel et al. (2020) conducted a fine-grained analysis of language varieties and demographics in Arabic. Additionally, Abbes et al. (2020) introduced a dialectal Arabic irony corpus extracted from Twitter.

Stance and polarization detection are closely related, as both involve identifying the attitude or opinion of a person, a group, or an entity towards a particular topic or issue. A requirement for automating stance detection is the availability of stance datasets. This section discusses existing stance and polarization datasets for Arabic.

Weber et al. (2013) introduced the first dataset for detecting polarization in Arabic. It comprises a collection of Twitter users, categorized based on their polarization towards Islamist or secular ideologies. The data collection process started by manually categorizing ”seed users” as either secularists or Islamists. Following retweet edges, the authors obtained a set of 7,088 Twitter profiles with their location set to Egypt and were automatically labeled as either ”Secularist” or ”Islamist”. The main objective of this dataset was to explore the bipolar political environment in Egypt after 2011. A limitation of this dataset is its focus on a single topic and a single country.

Darwish et al. (2017) used a dataset called *Islands Dataset* and proposed a method for stance detection using similarity between users as classification features. This dataset was developed by the same team that studied the attitudes of Arab Twitter users interested in Egyptian politics. Specifically, the authors selected tweets related to the transfer of ownership of the islands of Tiran and Sanafir from Egypt to Saudi Arabia in April 2016. They identified 48,445 tweets that 4,164 distinct users authored. These users and their corresponding tweets were then categorized either in favor of the islands' transfer (POS) or against it (NEG). Three annotators evaluated each user independently, and finally, 2,607 users (who authored 33,207 tweets) were kept, where all three annotators agreed on the same judgment. This corpus covers only one topic.

Another corpus created by Baly et al. (2018) unifies fact-checking and stance detection. It includes a set of annotated claim-article pairs, where each claim is associated with one or more articles. Initially, true and false claims were collected from two primary sources: the Syrian fact-checking website ”Verify” and the news Website ”Reuters”. After that, the authors used the Google Custom Search API to retrieve documents relevant to the collected claims. Each claim-document pair was assigned to 3-5 Arabic-speaking annotators for the stance annotation. This corpus includes 3,042 claim-document pairs annotated with the document's stance towards the claim, which can be ”agree”, ”disagree”, ”discuss” or ”unrelated”. The authors applied some stance detection models from the Fake News Challenge (FNC) [1] to their proposed corpus, achieving a weighted accuracy of 55.6% and an F1-score of 41.0%. This corpus is entirely in MSA and does not cover dialects. It only covers one topic: the Syrian War. It is also quite unbalanced, with most of the pairs labeled unrelated (around 2,000 pairs).

In Jaziriyan et al. (2021), the authors presented a target-based Arabic stance corpus called *ExaASC*, which focuses on tweets and their replies. The data collection involved extracting argumentative tweets and their corresponding replies using Twitter's stream API. Then, the target of the source sentence was determined by analyzing the content discussed in the replies, specifically related to the original tweet. The resulting dataset comprises 9,566 tweet replies that at least two native Arabic speakers annotated. These annotations are based on the stance of each tweet reply towards the target or an entity associated with the target, and they are assigned one of the following labels: ”favor”, ”against”, or ”none”. The authors used various implementations of pre-trained BERT to evaluate the proposed corpus and obtained a Macro F1-score of 70%. They stated that the models that only used the target performed better than the ones that used the whole tweet sentence. A limitation of this corpus is that it is unbalanced due to the random extraction of tweets from different countries.

[1] http://fakenewschallenge.org

Another target-based stance dataset was developed by Alturayeif et al. (2022). This dataset comprises 4,121 multi-dialectal Arabic sentences primarily gathered from Twitter. The dataset centers around three topics: ”COVID-19 vaccine,” ”digital transformation,” and ”women empowerment.” These targets are associated with three distinct labels: ”Favor,” ”Against,” and ”None.” The annotation process involved hiring Arabic annotators through the Appen crowdsourcing platform, ensuring linguistic accuracy and cultural relevance. Four BERT-based models were used with this dataset, and the best model achieved a macro-F1 score of 78.89%. While the overall dataset is balanced regarding the total number of sentences related to each target, it is unbalanced when considering the distribution of sentences per stance label.

Target-based datasets for stance detection in Arabic are currently limited to the works of Darwish et al. (2017), Jaziriyan et al. (2021), and (Alturayeif et al., 2022). The first dataset focuses only on one target: the controversial issue of transferring ownership of the Tiran and Sanafir islands from Egypt to Saudi Arabia. The two other datasets cover different dialects and targets but are not balanced. To address this gap, we aim to propose a balanced dataset for Arabic stance detection that spans a variety of topics and covers the main Arabic dialects.

## 3. Overview of Our Stance Corpus

This section introduces our novel cross-domain and multi-dialectal stance corpus, MARASTA. We also highlight the steps taken to collect and annotate the corpus data. Unlike most existing corpora, our corpus is well-balanced and covers several topics and Arabic dialects.

Given the numerous dialects that exist for Arabic, we had to dissect the data collection task into regions with similar dialects. We covered the Maghreb region (focusing on Tunisia), Egypt, the Levant Region (including Palestine and Jordan), and the Gulf region. In addition to these Arabic dialects, we also included the Modern Standard Arabic (MSA), which is used for official and generic communication. We collected at least 1,000 sentences for each of the four Arab regions, totaling 4657 sentences. Table 1 shows examples of sentences in our corpus and their English translation and stance.

### 3.1. Data Collection

Our research team consisted of individuals from diverse Arab countries, each deeply understanding the controversial topics in their respective regions. Each team member was tasked with compiling a list of these controversial topics and verifying their relevance through social media and Website searches. We used three criteria for choosing the topics. First, a topic should be sufficiently controversial to ensure plenty of discussions, with many posts and comments expressing different stances. Second, a topic can be either current or previously pertinent (within the past 15 years). Third, a topic can be relevant to one country or multiple countries.

After gathering controversial topic suggestions from all team members, we explored social media platforms to assess the availability of sentences related to these topics. Then, for each Arab region, we selected the two most highly voted topics. After that, we started collecting sentences related to these topics by using related seed keywords. Table 2 shows examples of the seed keywords used to collect relevant sentences for some topics.

A sentence must meet the following criteria to be included in our corpus: it must be written in Arabic script (either in MSA or in the dialect of the chosen region), it must be grammatically correct, related to the chosen topic, and must either express a stance (pro or against) or be neutral.

As collecting the sentences manually from multiple social media sources and websites became time-consuming, we resorted to social media platforms that provided API access and wrote Python scripts for data collection from Twitter and YouTube. The Python script for Twitter API requires the user to have a set of keywords related to the topic saved in a text file. The script then reads the text file and looks for tweets that include the keywords. Then, the script outputs a comma-separated value (.csv) file containing the relevant tweets. The script for YouTube API follows the traditional YouTube search method: The user searches for videos related to a keyword, and the links for videos related to that keyword are returned. The user is then asked to input the links of the relevant videos, and the comments under those videos will be compiled and saved into a .csv file. The user should inspect these files and move the appropriate sentences to the shared workbook containing our corpus data.

### 3.2. Corpus Balancing

To ensure the balance of our stance corpus, we established two rules. First, each topic from every region should have at least 504 sentences, 168 from each stance class (“pro” for agreeing, “neutral” for expressing no stance, and “against” for expressing opposition). Second, half of the sentences should

| Region's dialect | Topic | Example | Translation | Stance |
|---|---|---|---|---|
| Egyptian | New Administrative Capital in Egypt | أنا في صدمه من إننا صرفنا ستين مليار دولار في العاصمة الجديدة... يا رب أكون فهمت أو سمعت غلط . | I am shocked that We have spent sixty billion dollars on the new capital... Oh God, I hope I understood or heard that wrong. | Against |
| | Egypt's 2013 Political Transition | تحيا مصر الأبية تحت حكم الرئيس السيسي | Long live Egypt under the rule of the president Sisi. | Pro |
| Maghreb | Illegal Immigration to Europe | الحرقة مهيش حل وعمرها ما كانت حل | The immigration is not a solution, and never has it ever been a solution. | Against |
| | Tunisian General Labor Union | الاتحاد العام التونسي للشغل هو أكبر منظمة نقابية تونسية وتضم 500 000 عضو في 2017 | The UGTT is the largest Tunisian labor organization with 500,000 members in 2017. | Neutral |
| Gulf Region | Fifa World Cup 2022 | انطلاق بطولة كأس العالم 2022 بمواجهة قطر والإكوادور بتاريخ 20 نوفمبر | The start of the World Cup 2022 with the confrontation between Qatar and Ecuador on the 20th of November | Neutral |
| | Normalization With Israel | والله عن نفسي راضي بالتطبيع بس اعرف غيري مو راضي الإختلاف وارد | I am personally satisfied with the normalization, but I know others are not, disagreements are possible. | Pro |
| Levant Region | Presence of Refugees in Jordan | الاردن بيتكم وانتو الان جزء منا واحنا كثير مبسوطين فيكو تحياتي لكل السوريين في كل العالم | Jordan is your home, and you are now part of us, and we are very happy with you. Greetings to all Syrians around the world. | Pro |
| | Feminism and Women's Rights | النساء دمرو المجتمع بالمطالبه بالحقوق | Women destroyed the society by demanding rights. | Against |

**Table 1:** Examples from each region of our corpus along with their translations and stances

| Region's dialect | Topic | Keywords | Translation |
|---|---|---|---|
| Egyptian | New Administrative Capital in Egypt | العاصمة الإدارية الجديدة ; العاصمة الجديدة ; العاصمة الادارية ; | The New Administrative Capital ; The New capital ; The Administrative Capital |
| Maghreb | Illegal Immigration to Europe | الحرقة ; الهجرة الغير نظامية ; الهجرة السرية : أوروبا; إيطاليا | Immigration; Irregular Immigration; Clandestine Immigration: Europe; Italy |
| Gulf Region | Fifa World Cup 2022 | كأس العالم في قطر; تنظيم كأس العالم قطر; فيفا ٢٠٢٢ ; فيفا قطر; مونديال قطر | World Cup in Qatar; The Organization of Qatar's World Cup; FIFA 2022; FIFA Qatar; Qatar's Mundial |
| Levant Region | Feminism and Women's Rights | النسويات؛ الحركة النسوية؛ الاحتجاجات النسوية؛ الذكورية | Feminists; The Feminist Movement; Feminist Protests; Masculinity |

**Table 2:** Examples of seed keywords from each region's dialect and topic

be in MSA, and the remaining half should be in the respective regional dialect.

We developed dashboards for each region that update data as entered to enforce the above rules. The dashboards keep track of the total number of sentences per topic, the number of sentences per stance class, and the portion of sentences in each dialect. The dashboards can be filtered by topic, dialect, and stance class.

Table 3 shows the final number of sentences and their distribution per topic, dialect, and stance. This table reflects the results after the annotation and discussion meetings. Consequently, some sentences were excluded, resulting in a lower final sentence count than our initial intent. For instance, we have somewhat less than 504 sentences for each of the two topics of Egypt. The final total number of sentences in our corpus is 4657.

| Region's Dialect | Topic | Sentence Count | Dialect Distribution | | Stance | | |
|---|---|---|---|---|---|---|---|
| | | | MSA | Dialect | pro | neutral | against |
| Egyptian | New Administrative Capital in Egypt | 494 | 50.6% | 49.4% | 156 | 157 | 181 |
| | Egypt's 2013 Political Transition | 480 | 56.7% | 43.3% | 165 | 146 | 169 |
| Maghreb | Illegal Immigration to Europe | 590 | 55.8% | 44.2% | 185 | 174 | 231 |
| | Tunisian General Labor Union | 713 | 65.2% | 34.8% | 213 | 201 | 299 |
| Gulf Region | Fifa World Cup 2022 | 528 | 54.0% | 46.0% | 243 | 134 | 151 |
| | Normalization With Israel | 577 | 55.5% | 44.5% | 171 | 208 | 198 |
| Levant Region | Presence of Refugees in Jordan | 628 | 44.2% | 55.8% | 237 | 163 | 228 |
| | Feminism and Women's Rights | 647 | 46.1% | 53.9% | 203 | 183 | 261 |

**Table 3:** Corpus statistics and sentence distribution

### 3.3. Annotation Guidelines

Establishing comprehensive guidelines is fundamental to any annotation project to ensure consistency, accuracy, and reliability in the annotated data, as explained in (Zaghouani et al., 2014). Similar principles are applied in their later works, focusing on Arabic diacritized corpora (Zaghouani et al., 2016a) and machine translation post-editing (Zaghouani et al., 2016b), underscoring the importance of clear, well-defined guidelines in producing high-quality annotated datasets. These methodologies not only facilitate the annotation process but also enhance the utility and credibility of the resulting corpora, as seen in their contributions to Arabic language resources (Bouamor et al., 2018; Habash et al., 2018; Zaghouani and Charfi, 2018).

We developed detailed annotation guidelines for our corpus to ensure consistent and reliable annotations. The guidelines provide clear instructions and examples to the annotators, helping them accurately identify the stance expressed in each sentence toward the given topic. These guidelines covered the following key aspects:

- **Definition of Stance**: A clear explanation of stance was given, differentiating it from sentiment or opinion. The guidelines defined stance as the expression of being in favor of, against, or neutral towards a specific target.
- **Stance Labels**: Detailed descriptions of the three stance labels used in the annotation: "Pro" (expressing support or agreement), "Against" (expressing opposition or disagreement), and "Neutral" (expressing neither support nor opposition).
- **Annotation Process**: Step-by-step instructions on the annotation process, including the number of annotators per sentence, conflict resolution procedures, and quality control measures.
- **Examples and Edge Cases**: Numerous examples illustrating each stance label, edge cases, and ambiguous instances to help annotators understand and consistently apply the guidelines.

### 3.4. Corpus Annotation

The corpus was annotated by a team of skilled native Arabic speakers from the four regions we selected. These employees were hired as part-time employees with an undergraduate academic degree in either Computer Science or Information Systems. They also have a background in NLP, which they gained through involvement in at least one prior project.

Each sentence was annotated blindly by two annotators. For each topic/sentence pair, both annotators were given detailed annotation guidelines and instructed to determine the stance of the sentence towards the given topic. After these two rounds of annotation, a script was run to detect conflicts, i.e., cases where the two annotators disagreed. Conflicts were then resolved by including a third annotator. If two annotations out of three match, then the two matching annotations would be considered as the final annotation. If all three annotations were different or if the sentence was vague for either the second or the third annotator, whether because she considered it as unrelated to the topic or its meaning was unclear, the sentence was labeled as "discuss", and a discussion meeting took place, which was attended by the three annotators involved. During the discussion meeting, the three annotators discussed the conflicted sentences, and a majority vote was used to determine the final annotation. If the three annotators agreed that a particular sentence was unclear or unrelated to the topic, the sentence would be eliminated from the corpus, leading to the final numbers shown in Table 3. In all annotation rounds, the annotators were familiar with the dialect of their assigned sentences.

To assess the quality of the annotation, we calculated the inter-annotator agreement. Since there are two main annotators, we used Cohen's Kappa, as shown in Table 4, for the different regions. All kappa scores are above 0.80, indicating sub-

stantial agreement among the annotators, which speaks to the quality of the annotation.

| Dialect | Cohen's Kappa |
|---|---|
| Egyptian | 0.8157 |
| Maghreb | 0.8921 |
| Gulf Region | 0.8041 |
| Levant Region | 0.8265 |
| Overall | 0.8376 |

**Table 4:** Corpus' inter-annotator agreement metrics using Cohen's Kappa

## 4. Experiments

This section presents the methodology and results of our machine-learning experiments to test and evaluate our dataset for stance detection in Arabic.

### 4.1. Setup

We split our dataset into training and testing sets with a 75:25 ratio, comprising 3,492 samples for training and 1,165 samples for testing.

Each sample extracted from the dataset comprises a topic name, a corresponding sentence, and a classification label indicating the sentence's stance toward the topic. We removed URLs, emails, stop words, punctuation, and non-Arabic characters to preprocess the text data. The proposed model is designed to determine the stance of a given sentence towards a specified topic, whether it is in favor, against, or neutral. Since machine learning models cannot understand text inputs directly, we defined several feature extraction methods, such as unigram-vectorizer, bigram-vectorizer, trigram-vectorizer, and tfidf-vectorizer, to extract features from the text data. These methods were applied to the training and testing sets, generating sparse matrices that can be used as inputs to the machine learning classifiers. We trained traditional and neural network classifiers to identify the most effective approach for our task. Initially, we tested traditional ML classifiers such as Support Vector Machines (SVM), Logistic Regression (LR), Random Forests, and Decision Trees. Then, we tested a neural network model, the Multi-Layer Perceptron.

All experiments were conducted using Google Colab[2], which served as the cloud-based computing infrastructure. It provided the computational resources required for training and evaluating the models. We used the *pandas* package to manipulate and analyze the data and the *nltk* package during the preprocessing phase for tasks such as tokenization and stop word removal. We relied on the package *sklearn* for training and evaluating all the machine learning classifiers for training and evaluating all the machine learning classifiers.

[2] https://colab.research.google.com/

### 4.2. Results

To evaluate each model's performance with our dataset, we used accuracy and F1-score as evaluation metrics. We also generated a classification report, which provides a detailed analysis of the model's performance for each class (pro, against, and neutral). This report helped us understand which models were better suited for our task. In addition, we analyzed the performance of each extracted feature to identify the most suitable feature for each model.

During the preprocessing step, we removed stop words, which negatively impacted the performance. This can be explained by the importance of some stop words, such as negation words, in determining the stance of a sentence. Based on that observation, we decided to skip the stop-word removal.

Table 5 compares the performance of the different classifiers using various vectorization techniques for feature extraction, including Unigram, Bigram, Trigram, and TF-IDF. The F1-score represents a weighted average of the three labels, calculated using precision and recall, which considers true positives, false positives, and false negatives. Accuracy, however, measures the proportion of correctly classified instances among all instances in the dataset. This analysis provides insights into the effectiveness of the classifiers across different feature extraction methods. Additionally, we separately vectorized the topics and sentences and then combined their vectors, which led to the results shown in Table 6.

| Classifier | Unigram | | Bigram | | Trigram | | TF-IDF | |
|---|---|---|---|---|---|---|---|---|
| | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc |
| SVM | .60 | .60 | .41 | .45 | .35 | .42 | .66 | .66 |
| Logistic Regression | .65 | .65 | .56 | .56 | .48 | .50 | .65 | .65 |
| Random Forest | .60 | .60 | .51 | .53 | .42 | .47 | .59 | .59 |
| Decision Tree | .55 | .55 | .53 | .54 | .46 | .48 | .50 | .50 |
| Multi-Layer Perceptron | .65 | .65 | .59 | .59 | .49 | .51 | .65 | .65 |

**Table 5:** Performance of ML models for stance detection using combined vectorization

| Classifier | Unigram | | Bigram | | Trigram | | TF-IDF | |
|---|---|---|---|---|---|---|---|---|
| | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc |
| SVM | .61 | .62 | .41 | .44 | .37 | .41 | .63 | .64 |
| Logistic Regression | .64 | .64 | .56 | .56 | .44 | .45 | .66 | .66 |
| Random Forest | .64 | .64 | .51 | .53 | .41 | .42 | .62 | .62 |
| Decision Tree | .56 | .56 | .53 | .54 | .42 | .43 | .52 | .52 |
| Multi-Layer Perceptron | .66 | .66 | .58 | .59 | .41 | .42 | .66 | .66 |

**Table 6:** Performance with separate vectorization and combined vectors

We observe slight variations in the performance metrics for each classifier and vectorization method by comparing the two approaches by comparing the two approaches. These differences highlight the impact of treating the topics and sentences as separate features during the vectorization process, leading to slight variations in the classification performance. For the SVM classifier, using separate vectorization and combined vectors, the F1 scores range from 0.37 to 0.63, while the accuracy scores range from 0.41 to 0.64. Logistic Regression also shows similar variations, with F1 scores ranging from 0.44 to 0.66 and accuracy scores ranging from 0.45 to 0.66. For Random Forest, the F1 scores range between 0.41 and 0.64, and accuracy scores range between 0.42 and 0.62. Decision Trees achieves F1 scores from 0.42 to 0.56 and accuracy scores from 0.43 to 0.52. Lastly, Multi-Layer Perceptron achieves F1 scores between 0.41 and 0.66 and accuracy scores between 0.42 and 0.66.

Although the experimental results display several combinations with very similar results, we opted for the one that performed better with both features, even if the differences were slight. The MLP achieved the best performance using unigram and TF-IDF features for separate and combined vectorization. This model consisted of multiple hidden layers, resulting in 2,178,100 parameters representing the weights and biases learned during the training. The model's RAM (Random Access Memory) usage was 3.7 GB (out of the total available 12.7 GB). Additionally, the GPU RAM usage was 0.8 GB (out of the total available 15.0 GB).

### 4.3. Discussion

We explored different feature combinations to enhance the classifiers' performance. However, the results of our experiments demonstrated that these feature combinations did not yield significant improvements. Therefore, our focus shifted towards selecting each classifier's most effective individual feature extraction method to achieve the best performance.

The results presented in Table 7 show variations in the performance of different classifiers and feature extraction methods. The SVM classifier performs better using the TF-IDF feature extraction method, achieving the highest F1 score of 0.63 and an accuracy score of 0.64. Similarly, the Logistic Regression classifier performs better with the TF-IDF feature extraction method, achieving an F1 score of 0.66 and an accuracy score of 0.66. These results indicate that the TF-IDF feature extraction method captures essential information for both SVM and Logistic Regression.

In contrast, Random Forest performs better when using the unigram feature extraction method, achieving an F1 score of 0.64 and an accuracy score of 0.64. The Decision Trees classifier performed lower than the other classifiers with unigram and TF-IDF features. Its F1 scores range from 0.42 to 0.56, and its accuracy scores range from 0.43 to 0.52, showing some limitations.

On the other hand, the MLP classifier achieved an F1-score of 0.66 and an accuracy of 0.66 with the Unigram and TF-IDF features, which is the highest among all the evaluated classifiers. MLP models, like other deep learning models, often require larger datasets to train their multiple layers of neurons effectively. However, it is essential to consider the data's quantity, quality, and representation. While larger datasets can generally provide more diverse examples and improve model performance, it is still possible to achieve good results with MLP models using smaller datasets if the data is high quality and representative of the problem at hand.

Among the state-of-the-art datasets for Arabic stance detection, we found that the MAWQIF dataset (Alturayeif et al., 2022) is quite similar to ours as it considers the stance of sentences towards specific targets. Therefore, we used the MAWQIF dataset for comparison purposes and tested it using the MLP model with TF-IDF features, demonstrating its best performance. Table 7 compares evaluation metrics between our dataset and the MAWQIF dataset for each stance class. Three key metrics, namely Precision, Recall, and F1-score, are provided for each stance class, namely, "Favor," "Against," and "Neutral". Additionally, we included the overall accuracy and F1-score to provide a global perspective of model performance with both datasets.

It is noteworthy that the MARASTA and MAWQIF datasets were separately trained and tested using the same model. We observed the following performance metrics: the model achieved an accuracy of 0.73 and an F1-score of 0.71 with the MAWQIF dataset. In contrast, this model achieved an accuracy with our dataset MARASTA and an F1-score of 0.66. However, a deeper analysis reveals significant differences in performance within individual stance classes for the MAWQIF dataset. Specifically, within the 'Neutral' class, the precision was 0.42, the recall was 0.19, and the F1-score was 0.26. For the 'Against' class, the precision was 0.64, the recall was 0.63, and the F1-score was 0.64. Meanwhile, for the 'Pro' class, the precision was 0.78, the recall was 0.86, and the F1-score was 0.82. On the other hand, for MARASTA, within the 'Neutral' class, the precision

| Class | Metric | Dataset | |
|---|---|---|---|
| | | MARASTA | MAWQIF |
| Pro | Precision | 0.69 | 0.78 |
| | Recall | 0.63 | 0.86 |
| | F1-score | 0.66 | 0.82 |
| Against | Precision | 0.67 | 0.64 |
| | Recall | 0.63 | 0.63 |
| | F1-score | 0.65 | 0.64 |
| Neutral | Precision | 0.61 | 0.42 |
| | Recall | 0.73 | 0.19 |
| | F1-score | 0.66 | 0.26 |
| Overall Accuracy | | 0.66 | 0.73 |
| Macro F1-score | | 0.66 | 0.71 |

**Table 7:** Performance comparison of our dataset MARASTA and MAWQIF dataset for detecting each stance class

was 0.61, the recall was 0.73, and the F1-score was 0.66. In the 'Against' class, the precision was 0.67, the recall was 0.63, and the F1-score was 0.65. For the 'Pro' class, the precision was 0.69, the recall was 0.63, and the F1-score was 0.66. Unlike MAWQIF, which shows excellent performance for the 'Pro' class but poor performance for the 'Neutral' class, our dataset demonstrates more balanced performance across all stance classes. The significant disparity in performance across different stance labels can be directly attributed to the imbalance in label distribution in MAWQIF, underscoring the importance of addressing class label imbalances. We attribute the better model performance on our corpus to the balanced distribution of sentences among the three stance classes, which ensures that the model is exposed to a diverse range of examples for each stance, allowing it to learn and generalize better.

## 5. Conclusion

We presented a novel cross-domain multi-dialectal Arabic stance corpus, which covers four regions of the Arab World: Maghreb, Egypt, Levantine, and the Gulf. This corpus includes over 4,500 sentences grouped into eight distinct topics across all regions. The collected sentences were annotated by going through at least two rounds of annotation, with a third round added in case of a conflict. Furthermore, we reported on machine learning experiments that we conducted on our stance corpus for the stance detection task using classical classifiers and neural network models with different feature combinations to build a model that can automatically predict the stance of any sentence in Arabic. The Multi-Layer Perceptron (MLP) classifier gave the best when using Unigram or TF-IDF features. Combining features to increase performance was less effective than selecting the best individual feature extraction techniques for each classifier. Additionally, we demonstrated that the balanced distribution of sentences across classes in our dataset contributes to the consistent performance of the model across different stance labels. This highlights the importance of balancing distribution for improved performance. Our future work will focus on building tools for stance detection in Arabic and applying these in real-world scenarios.

## 6. Limitations

One limitation of our corpus is its relatively small size, with slightly more than 4,500 annotated sentences. Additionally, the number of sentences of around 500 per topic could be considered small. This might limit the number of words that are covered by each dialect. However, our experiments show that the small size of the corpus is compensated by its sound quality. Furthermore, for some regions covered by our corpus, only one dialect was included as representative of the region, such as the Maghreb region, where we only included sentences from the Tunisian dialect. The issue of specific words to certain countries with similar dialects might arise. Besides, while our corpus covers four main Arabic dialects, some Arabic dialects are still not covered, such as Iraqi and Yemeni.

## 7. Ethical Considerations

In developing the MARASTA corpus and conducting the experiments for stance detection, we have upheld stringent ethical standards to ensure the respect and protection of individual privacy and data. Our data collection processes were meticulously designed to use publicly available information, avoiding personally identifiable information to maintain anonymity and confidentiality.

During the annotation process, we employed annotators who are native speakers and possess a background in NLP. This ensured a high level of understanding and sensitivity towards the cultural and linguistic nuances of the Arabic language and its various dialects. Annotators were trained to handle data ethically, maintaining impartiality and objectivity in their annotations.

We acknowledge that stance detection can be sensitive and potentially controversial, particularly in politically and socially charged contexts. Therefore, we emphasize that our research aims to advance understanding of NLP and not to foster or endorse any particular viewpoint or stance. We strive to contribute to the broader academic community's knowledge and develop tools to assist in understanding language, stance, and sentiment in a multicultural and multilingual world.

## Acknowledgements

This publication was made possible by NPRP13S-0206-200281 from the Qatar National Research Fund. The contents herein reflect the work and are solely the authors' responsibility.

## References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. Daict: A dialectal arabic irony corpus extracted from twitter. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6265–6271.

Arfan Ahmed, Nashva Ali, Mahmood Alzubaidi, Wajdi Zaghouani, Alaa A Abd-alrazaq, and Mowafa Househ. 2022. Freely available arabic corpora: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2:100049.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Inf. Process. Manag.*, 58(4):102597.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. *arXiv preprint arXiv:1804.08012*.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse Processes*, 11(1):1–34.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Rong Cao, Xiangyang Luo, Yaoyi Xi, and Yaqiong Qiao. 2022. Stance detection for online public opinion awareness: An overview. *International Journal of Intelligent Systems*, 37(12):11944–11965.

Anis Charfi, Wajdi Zaghouani, Syed Hassan Mehdi, and Esraa Mohamed. 2019. A fine-grained annotated multi-dialectal arabic corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 198–204.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. 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*, ASONAM '17, page 145–148, New York, NY, USA. Association for Computing Machinery.

Tim Draws, Karthikeyan Natesan Ramamurthy, Ioana Baldini, Amit Dhurandhar, Inkit Padhi, Benjamin Timmermans, and Nava Tintarev. 2023. Explainable cross-topic stance detection for search results. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 221–235.

Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, et al. 2018. Unified guidelines and resources for arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mohammad Mehdi Jaziriyan, Ahmad Akbari, and Hamed Karbasi. 2021. ExaASC: A general target-based stance detection corpus in arabic language. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*. IEEE.

Dilek Küçük and Fazli Can. 2021. Stance detection: A survey. *ACM Comput. Surv.*, 53(1):12:1–12:37.

Yingjie Li and Cornelia Caragea. 2019. 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,

Hong Kong, China. Association for Computational Linguistics.

Nolan McCarty. 2019. *Polarization: What everyone needs to know®*. Oxford University Press.

Dana Mckay, Stephann Makri, Marisela Gutierrez-Lopez, Andrew MacFarlane, Sondess Missaoui, Colin Porlezza, and Glenda Cooper. 2020. We are the change that we seek: Information interactions during a change of viewpoint. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, page 173–182, New York, NY, USA. Association for Computing Machinery.

Francisco Rangel, Paolo Rosso, Wajdi Zaghouani, and Anis Charfi. 2020. Fine-grained analysis of language varieties and demographics. *Natural Language Engineering*, 26(6):641–661.

Paolo Rosso, Francisco Rangel, Irazu Hernández Farías, Leticia Cagnina, Wajdi Zaghouani, and Anis Charfi. 2018. A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass*, 12(4):e12275.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. A multi-task model for sentiment aided stance detection of climate change tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 854–865.

Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 290–297.

Wajdi Zaghouani. 2014. Critical survey of the freely available arabic corpora. In *International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Reykjavik, Iceland, 26-31 May 2014*.

Wajdi Zaghouani, Houda Bouamor, Abdelati Hawwari, Mona Diab, Ossama Obeid, Mahmoud Ghoneim, Sawsan Alqahtani, and Kemal Oflazer. 2016a. Guidelines and framework for a large scale arabic diacritized corpus. In *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3637–3643. European Language Resources Association (ELRA).

Wajdi Zaghouani and Anis Charfi. 2018. Guidelines and annotation framework for arabic author profiling. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wajdi Zaghouani, Nizar Habash, Ossama Obeid, Behrang Mohit, Houda Bouamor, and Kemal Oflazer. 2016b. Building an arabic machine translation post-edited corpus: Guidelines and annotation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1869–1876.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale arabic error annotation: Guidelines and framework.