ISHFMG_TUN at StanceEval: Ensemble Method for Arabic Stance Evaluation System

Ammar Mars
University of Tunis, ISG
Smart Lab Laboratory
Tunis,Tunisia
ammar.mars@gmail.com

Mustapha Jaballah
University of Tunis, ENSIT
SIME Laboratory
Tunis, Tunisia
musjmusj4@gmail.com

Dhaou Ghoul
HIGHSYS,
HF Lab,
Paris, France
dhaou.ghoul@gmail.com

Abstract

It is essential to understand the attitude of individuals towards specific topics in Arabic language for tasks like sentiment analysis, opinion mining, and social media monitoring. However, the diversity of the linguistic characteristics of the Arabic language presents several challenges to accurately evaluate the stance. In this study, we suggest ensemble approach to tackle these challenges. Our method combines different classifiers using the voting method. Through multiple experiments, we prove the effectiveness of our method achieving significant F1-score value equal to 0.7027. Our findings contribute to promoting Natural Language Processing (NLP) and offer treasured enlightenment for applications like sentiment analysis, opinion mining, and social media monitoring.

Keywords: Stance detection, ensemble method, machine learning, Arabic language, classification.

1 Introduction

Nowadays, almost all people use social media for different purposes such as communication, business, and advertisement. Social media platforms have revolutionized the way people interact, share information, and conduct business, creating a global community that is more interconnected than ever before.

This digital transformation has led to an immense amount of user-generated content being produced daily, encompassing various languages and dialects. A large number of these users are Arabic people. In fact, Arabic is among the most commonly worldwide spoken languages, with over 300 million native speakers across more than 20 countries.

This widespread use underscores the significance of Arabic in the digital sphere and highlights the need for effective Natural Language Processing (NLP) tools tailored to this language.

Arabic represents an important target for NLP due to its linguistic richness and diverse user base. However, Arabic presents different characteristics making it challenging for NLP tasks. For instance, Arabic script is written without spaces between letters, which poses difficulties in tokenization and word segmentation.

Additionally, there are several different orthographic conventions for representing short vowels, long vowels, and other vowel sounds, which add to the complexity of processing Arabic text (Saiegh-Haddad and Henkin-Roitfarb, 2014). These linguistic features necessitate the development of specialized algorithms and models that can accurately interpret and analyze Arabic text.

Many researches have been conducted on this topic, addressing various aspects of Arabic NLP, including text classification, named entity recognition, and machine translation. Among these research efforts, stance detection has emerged as a crucial task. Stance detection involves determining the position or attitude expressed in a text towards a particular target, which is essential for applications like opinion mining, sentiment analysis, and social media monitoring.

This paper describes our participation in such a task: Arabic stance evaluation (Alturayeif et al., 2024). Our work focuses on developing and refining techniques to accurately assess stances expressed in Arabic texts, leveraging the latest advancements in NLP and machine learning.

The implications of effective stance detection in Arabic are far-reaching. It can enhance the capabilities of automated systems to gauge public opinion, monitor trends, and identify emerging issues in real-time, providing valuable insights for businesses, policymakers, and researchers.

This work can be exploited in tasks like opinion mining, sentiment analysis, and social media monitoring, enabling more nuanced and culturally aware analyses of Arabic content.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 832–836

This paper is organized into four sections: Section 2 describes the dataset, detailing the sources, preprocessing steps, and characteristics of the data used in our experiments. Section 3 presents the methods employed, including the models and techniques applied to tackle the stance detection task. Section 4 is reserved for experimental results and discussions, where we analyze the performance of our approach and compare it with existing methods. Finally, Section 5 provides perspectives on future work, outlining potential improvements and directions for further research in Arabic NLP.

2 Data

In the experimentation of this work, we used MAWQIF Dataset (Alturayeif et al., 2022), which is the first Arabic dataset for target-specific stance detection, composed of 4,121 tweets annotated about three targets : "COVID-19 vaccine" (1,373 tweets), "digital transformation" (1,348 tweets) and "women empowerment" (1,400 tweets) as a shown in the table 1.

MAWQIF is a multi-label dataset that provides annotations for stance,sentiment, and sarcasm as a shown in the table 2. This rich dataset provides a benchmark for the three tasks and allows opportunities for studying the interaction between different opinion dimensions and evaluating a multi-task model.

In preparing our data and through extensive experiments, we opted for applying series of cleaning up steps to each tweet. First, we removed punctuation marks,numbers, non Arabic words, and duplicated characters. Second, we eliminated emoticons. Third, we dropped stop words and Arabic diacritics. Finally we normalized the dataset. (Ayedh et al., 2016).

Furthermore, to address class imbalance in our training data, we implemented the Synthetic Minority Oversampling TEchnique (SMOTE) (Bowyer et al., 2011). SMOTE operates by identifying neighboring instances in the feature space, drawing a line between them, and generating new samples along that line.

3 System

Our model is constructed upon a conventional machine learning approach. For model construction and training, we utilize the FeatureUnion module in SCIKIT-LEARN (Pedregosa et al., 2018), which facilitates the seamless combination of various n-gram representations at both the word and character levels, as illustrated in figure 1.

During model training, we concatenate three vectors containing the following features weighted using Term Frequency-Inverse Document Frequency (TF-IDF): word n-grams (1 to 5-grams), character n-grams (1 to 4-grams), and character n-grams (1 to 5-grams) with word boundaries explicitly delineated by spaces.

We employ series of classifiers organized through a voting technique utilizing the Voting Classifiers Stochastic Gradient Descent(SGD)(Ketkar, 2017), Linear Support Vector Classification(LinearSVC) (Pedregosa et al., 2011), Multinomial Naive Bayes (MNB)(Murphy et al., 2006), Ridge Classifier (Grüning and Kropf, 2006), Random Forest (RF) (Liu et al., 2012) .

This approach constructs an ensemble voting classifier employing hard voting, which aggregates predicted class labels based on majority rule. The ensemble comprises the following classifiers:

- SGD with alpha = 0.00001 and penalty = 'l2'.
- LinearSVC with penalty = 'l2' and Tolerance for stopping set to 0.001.
- MNB with alpha = 0.01.
- Ridge Classifier with alpha = 1.
- RF with defaults parameters.

Note that our model draws inspiration from a system designed for Arabic dialect classification (Ghoul and Lejeune, 2020).

4 Results

The Arabic stance evaluation system using the ensemble approach showed promising outcomes. The F1-scores for Digital Transformation, Covid Vaccine, and Women Empowerment are shown in the table 3.

These scores demonstrate the accuracy of the system. In fact, it performs well in categorizing stances taking on the challenges of the Arabic language.

The F1-score for Women Empowerment was significantly higher than for the other themes, indicating that the system performed a particularly good job of assessing positions on this particular topic.

This could be due to a number of things, including the obvious sentiments that are common in

Target	Train			Dev			Test			Total
	#Favor	#Against	#None	#Favor	#Against	#None	#Favor	#Against	#None	
Women empowerment	596	303	44	164	68	15	134	65	11	1400
Digital Transformation	717	117	98	162	25	26	156	25	22	1348
Covid-19 Vaccine	402	402	122	107	106	28	90	90	26	1373
Total	1715	822	264	433	199	69	380	180	59	4121

Table 1: Data split statistics

Target	Tweet	Stance	Sentiment	Sarcasm
Women empowerment	انا مع تمكين المرأة اصلا (I'm all for empowering women.)	Favor	Positive	No
Digital Transformation	ممكن تفهمونا شنو التحول الالكتروني لو هو مجرد خبر (Could you explain what digital transformation is, even if it's just briefly?)	None	Neutral	No
Covid-19 Vaccine	كله من تطعيم كورونا(All from Corona vaccination)	Against	Negative	No

Table 2: Examples of annotated tweets

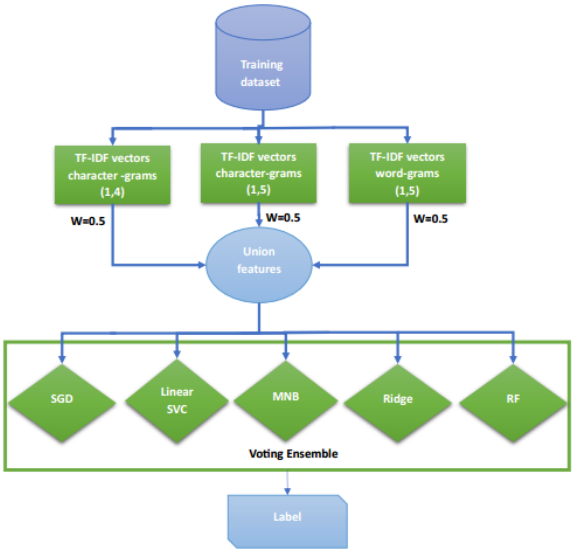

Figure 1: Model of the Ensemble Classifier

conversations about women empowerment in most tweets.

Nevertheless, the F1-score for Digital Transformation was marginally lower. This suggests that it would be difficult to assess stances on this issue.

To overcome this issue more investigation is necessary. To sum up, these outcomes show that the used approach is valuable for stance evaluation with performance varying depending on the topic.

5 Discussion

The application of the ensemble method in the Arabic stance evaluation process has proved the importance of using advanced methods to overcome the challenges that arise in Arabic NLP.

Ensemble methods, which combine multiple models to ameliorate prediction accuracy, have shown promise in dealing with the complexity of the Arabic language such as its rich morphology, diverse dialects, orthographic variations, and complex tokenization task.

Arabic is renowned for its rich morphology, characterized by a complex system of word formation and inflection. Arabic verbs and nouns undergo extensive inflectional changes to indicate tense, aspect, mood, person, number, and gender. For example, the verb root لقح (to Vaccinate) can be conjugated as لقح (he vaccinated), يلقح (He vaccinates),لقحوا (they vaccinate).

In addition, Arabic is known for its diverse dialects, which vary significantly across regions where Arabic is spo-

Target	Trained on	Tested on	F1-score
Women Empowerment	Train Set	Dev Set	0.7927
	Train+Dev	Test Set	0.7393
Covid Vaccine	Train Set	Dev Set	0.7548
	Train+Dev	Test Set	0.70193
Digital Transformation	Train Set	Dev Set	0.6981
	Train+Dev	Test Set	0.6670
Overall F1-score	Train Set	Dev Set	0.7485
	Train+Dev	Test Set	0.7027

Table 3: F1-scores for different targets and overall F1 score on development and test data

ken. For example, اول يوم بس كلمة and ما ناخذ تطعيم كورونا عشان ما نتتهي و تفوتنا الحياه. In these tweets the two words (بس and عشان) are used in Gulf dialect ,but in other dialects (فقط and لأن) are used. We can mention that the Orthographic variations in Arabic primarily involve differences in the way Arabic script is written across different regions and contexts.

Other noticed complexity is the tokenization task. It can be challenging due to the complex nature of the language, especially with regards to morphology and to the different forms that words can take based on their context. For example the tweet يرصدون الاثار السلبيه (They monitor the negative effects) , the tokenization challenge is to discover the root (رصد) of the conjugated verb يرصدون and discover the definite article ال attached to the nouns (اثار and سلبيه).

These methods use the strengths of certain models and reduce their weaknesses resulting in more reliable performance in stance detection.

Differences in F1-scores between subjects arise interesting questions about the impact of subject specificity on the accuracy of the stance detection task. For instance, the ambiguity that comes from idiomatic expressions can impact the effectiveness of NLP models.

The significantly higher F1-scores for the women empowerment target raise the possibility that some topics have unique linguistic characteristics that facilitate correct classification.

Moreover, the relatively low F1-scores of digital transformation highlight the importance of continuous research programs to solve problems related unique to specific areas.

In figure 2, the confusion matrix of the test set indicates that predicting the “against” label was more challenging, with 118 true positives (TP), 66 false positives(FP), and 62 false negatives (FN). This difficulty might be attributed to the use of indirect language and cultural context.

Similarly, the confusion matrix highlights the difficulty in accurately predicting the “None” label, with only 10 TP, 24 FP, and 50 FN. These prediction challenges stem from the imbalanced nature of the data.

To overcome these shortcomings, other techniques could be investigated:

- Ameliorating preprocessing
- Investigating other techniques of feature extraction such as Marabert
- Augmenting the dataset
- Balancing the dataset

6 Conclusion

In summary, this study utilized an ensemble approach to examine Arabic stance evaluation using the MAWQIF Dataset. Our system demonstrated a notable degree of success in detecting stance from Arabic tweets, showcasing the potential of ensemble methods in this domain. Despite these promising results, there remains room for improvement through advanced modeling techniques, such as transformer-based architectures and fine-tuning pre-trained language models for Arabic.

The availability of annotated datasets like MAWQIF is crucial for advancing Arabic NLP research, providing a solid foundation for innovative applications. Future research should focus on expanding dataset diversity and exploring transfer learning to adapt models trained on high-resource languages to Arabic.

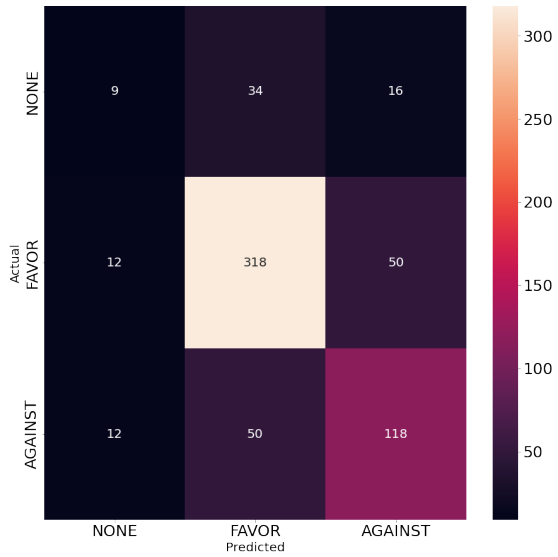

Figure 2: Confusion matrix on the test set for the Voting Ensemble classifier

Our findings contribute to the growing body of work in Arabic NLP, highlighting the potential of ensemble approaches. By continuing to innovate and build on this research, we aim to develop more accurate, efficient, and contextually aware systems for Arabic text analysis, fostering better communication and understanding in the Arabic-speaking world.

References

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. Stanceeval 2024: The first arabic stance detection shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

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.

Abdullah Ayedh, Guanzheng TAN, Khaled Alwesabi, and Hamdi Rajeh. 2016. The effect of preprocessing on arabic document categorization. *Algorithms*, 9(2).

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

Dhaou Ghoul and Gaël Lejeune. 2020. Voting Classifier vs Deep learning method in Arabic Dialect Identification. In *: Proceedings of the Fifth Arabic Natural Language Processing Workshop, COLING 2020*, Barcelone, Spain.

Martin Grüning and Siegfried Kropf. 2006. A ridge classification method for high-dimensional observations. In *From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation eV University of Magdeburg, March 9–11, 2005*, pages 684–691. Springer.

Nikhil Ketkar. 2017. *Stochastic Gradient Descent*, pages 113–132. Apress, Berkeley, CA.

Yanli Liu, Yourong Wang, and Jian Zhang. 2012. New machine learning algorithm: Random forest. In *Information Computing and Applications*, pages 246–252, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kevin P Murphy et al. 2006. Naive bayes classifiers. *University of British Columbia*, 18(60):1–8.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in python.

Elinor Saiegh-Haddad and Roni Henkin-Roitfarb. 2014. *The Structure of Arabic Language and Orthography*, pages 3–28. Springer Netherlands, Dordrecht.