

AraStance: A Multi-Country and Multi-Domain Dataset of Arabic Stance Detection for Fact Checking

Tariq Alhindi,¹ Amal Alabdulkarim,² Ali Alshehri,³
Muhammad Abdul-Mageed,⁴ and Preslav Nakov⁵

¹Department of Computer Science, Columbia University

²Georgia Institute of Technology, ³The State University of New York at Buffalo

⁴The University of British Columbia, ⁵Qatar Computing Research Institute, HBKU
tariq@cs.columbia.edu, amal@gatech.edu, alimoham@buffalo.edu
muhammad.mageed@ubc.ca, pnakov@hbku.edu.qa

Abstract

With the continuing spread of misinformation and disinformation online, it is of increasing importance to develop combating mechanisms at scale in the form of automated systems that support multiple languages. One task of interest is claim veracity prediction, which can be addressed using stance detection with respect to relevant documents retrieved online. To this end, we present our new Arabic Stance Detection dataset (AraStance) of 4,063 claim–article pairs from a diverse set of sources comprising three fact-checking websites and one news website. AraStance covers false and true claims from multiple domains (e.g., politics, sports, health) and several Arab countries, and it is well-balanced between related and unrelated documents with respect to the claims. We benchmark AraStance, along with two other stance detection datasets, using a number of BERT-based models. Our best model achieves an accuracy of 85% and a macro F1 score of 78%, which leaves room for improvement and reflects the challenging nature of AraStance and the task of stance detection in general.

1 Introduction

The proliferation of social media has made it possible for individuals and groups to share information quickly. While this is useful in many situations such as emergencies, where disaster management efforts can make use of shared information to allocate resources, this evolution can also be dangerous, e.g., when the news shared is not precise or is even intentionally misleading. Polarization in different communities further aggravates the problem, causing individuals and groups to believe and to disseminate information without necessarily verifying its veracity (misinformation) or even making up stories that support their world views (disinformation). These circumstances motivate a need to develop tools for detecting fake news online, including for a region with opposing forces and ongoing conflicts such as the Arab world.

Our work here contributes to these efforts a new dataset and baseline results on it. In particular, we create a new dataset for stance detection of claims collected from a number of websites covering different domains such as politics, health, and economics. The websites cover several Arab countries, which enables wider applicability of our dataset. This compares favorably to previous work for Arabic stance detection such as the work of Baly et al. (2018), who focused on a single country. We use the websites as our source to collect true and false claims, and we carefully crawl web articles related to these claims. Using the claim–article pairs, we then manually assign *stance* labels to the articles. By stance we mean whether an article *agrees*, *disagrees*, *discusses* a claim or it is just *unrelated*. This allows us to exploit the resulting dataset to build models that automatically identify the stance with respect to a given claim, which is an important component of fact-checking and fake news detection systems. To develop these models, we resort to transfer learning by fine-tuning language models on our labeled dataset. We also benchmark our models on two existing datasets for Arabic stance detection. Finally, we make our dataset publicly available.¹

Our contributions can be summarized as follows:

1. We release a new multi-domain, multi-country dataset labeled for both stance and veracity.
2. We introduce a multi-query related document retrieval approach for claims from diverse topics in Arabic, resulting in a dataset with balanced label distributions across classes.
3. We compare our dataset to two other Arabic stance detection datasets using four BERT-based (Devlin et al., 2019) models.

¹The data can be found at <http://github.com/Tariq60/arastance>.

2 Related Work

Stance detection started as a standalone task, unrelated to fact-checking (Küçük and Can, 2020). One type of stance models the relation (e.g., *for*, *against*, *neutral*) of a text segment towards a *topic*, usually a controversial one such as abortion or gun control (Mohammad et al., 2016; Abbott et al., 2016). Another one models the relation (e.g., *agree*, *disagree*, *discuss*, *unrelated*) between two pieces of text (Hardalov et al., 2021b; Ferreira and Vlachos, 2016). The latter definition is used in automatic fact-checking, fake news detection, and rumour verification (Vlachos and Riedel, 2014).

There are several English datasets that model fact-checking as a stance detection task on text from multiple genres such as Wikipedia (Thorne et al., 2018), news articles (Pomerleau and Rao, 2017; Ferreira and Vlachos, 2016), and social media (Gorrell et al., 2019; Derczynski et al., 2017). Most related to our work here is the Fake News Challenge, or FNC, (Pomerleau and Rao, 2017), which is built by randomly matching claim–article pairs from the Emergent dataset (Ferreira and Vlachos, 2016), which itself pairs 300 claims to 2,500 articles. In FNC, this pairing is done at random, and it yielded a large number of unrelated claim–article pairs. There are several approaches attempting to predict the stance on the FNC dataset using LSTMs, memory networks, and transformers (Hanselowski et al., 2018; Conforti et al., 2018; Mohtarami et al., 2018; Zhang et al., 2019; Schiller et al., 2021; Schütz et al., 2021).

There are two datasets for Arabic stance detection with respect to claims. The first one collected their false claims from a single political source (Baly et al., 2018), while we cover three sources from multiple countries and topics. They retrieved relevant documents and annotated the claim–article pairs using the four labels listed earlier (i.e., *agree*, *disagree*, *discuss*, *unrelated*). They also annotated “rationales,” which are segments in the articles where the stance is most strongly expressed. The other Arabic dataset by Khouja (2020) uses headlines from news sources and generated true and false claims by modifying the headlines. They used a three-class labeling scheme of stance by merging the *discuss* and the *unrelated* classes in one class called *other*.

Our work is also related to detecting machine-generated and manipulated text (Jawahar et al., 2020; Nagoudi et al., 2020).

3 AraStance Construction

We constructed our AraStance dataset similarly to the way this was done for the English Fake News Challenge (FNC) dataset (Pomerleau and Rao, 2017) and for the Arabic dataset of Baly et al. (2018). Our dataset contains true and false claims, where each claim is paired with one or more documents. Each claim–article pair has a stance label: *agree*, *disagree*, *discuss*, or *unrelated*. Below, we describe the three steps of building AraStance: (i) claim collection and pre-processing, (ii) relevant document retrieval, and (iii) stance annotations.

3.1 Claim Collection and Preprocessing

We collected false claims from three fact-checking websites: ARAANEWS², DABEGAD³, and NORUMORS⁴, based in the UAE, Egypt, and Saudi Arabia, respectively. The claims were from 2012 to 2018 and covered multiple domains such as politics, sports, and health. As the three fact-checking websites only debunk false claims, we looked for another source for true claims: following Baly et al. (2018), we collected true claims from the Arabic website of REUTERS⁵, assuming that their content was trustworthy. We added topic and date restrictions when collecting the true claims in order to make sure they were similar to the false claims. Moreover, in order to ensure the true claims were from the same topics as the false ones, we used a subset of the false claims as seeds to retrieve true claims that were within three months of the seed false claims, and we ranked them by TF.IDF, similarity to the seeds. We kept a maximum of ten true claims per seed false claim. For all claims, we removed the ones that contained no-text and/or were multimedia-centric. Moreover, we manually modified the false claims by removing phrases like “*It is not true that*”, “*A debunked rumor about*”, or “*The reality of*”, which are often used by fact-checking websites. This sometimes required us to add a noun at the beginning of the claim based on the text of the target articles, or to make some grammatical edits. We show examples of two false claims before and after preprocessing in Table 1. Note that the headlines we retrieved from REUTERS were already phrased as claims, and thus we did not have to edit them in any way.

²<http://araanews.ae>

³<http://dabegad.com>

⁴<http://norumors.net>

⁵<http://ara.reuters.com>

Original Claim	Preprocessed Claim
<p>لا صحة لما يتم تداوله حول تسجيل مكالمات ورسائل المقيمين</p> <p><i>What is being circulated about recording residents' calls and messages is not true</i></p>	<p>الحكومة تسجل مكالمات ورسائل المقيمين</p> <p><i>The government is recording calls and messages of residents</i></p>
<p>اشعة الكوزمو القادمة من كوكب المريخ اشاعة كاذبة من ٢٠٠٨</p> <p><i>Cosmo rays coming from Mars is a false rumor from 2008</i></p>	<p>ناسا تحذر من اشعة الكوزمو الخطيرة القادمة من كوكب المريخ هذه الليلة من الساعة ٣٠.٣٠-٣.١٢</p> <p><i>NASA warns of dangerous cosmic rays coming from Mars tonight from 12.30-3.30</i></p>

Table 1: Examples of false claims before and after preprocessing.

3.2 Document Retrieval

For each claim, we retrieved relevant documents using multiple queries and the Google Search API. It was harder to find relevant documents for the false claims by passing their preprocessed version as queries because of their nature, locality, and diversity. For some false claims, there were extra clauses and modifiers that restricted the search results significantly as shown in the examples below:

1. طفلة تظهر بنصف جسم بشري،
النصف الاخر ثعبان
A female child with half a human body, and the other half is a snake
2. يتم تنظيف الرئتين عند المدخنين من خلال
شم بخار اللبن و الماء لمدة عشرة أيام
Lungs of smokers are cleaned by smelling the steam of milk and water for ten days

To remedy this, we boosted the quality of the retrieved documents by restricting the date range to two months before and after the date of the claim, prepending named entities and removing extra clauses using parse trees. In order to emphasize the presence of the main entity(s) in the claim, we extracted named entities using the Arabic NER corpus by Benajiba et al. (2007) and Stanford’s CoreNLP Arabic NER tagger (Manning et al., 2014). We further used Stanford’s CoreNLP Arabic parser to extract the first verb phrase (VP) and all its preceding tokens in the claim, as this has been shown to improve document retrieval results for claim verification, especially for lengthy claims (Chakrabarty et al., 2018). For the two examples shown above, we would keep the claims until the comma for the first example and the word *and* for the second one, and we would consider those as the queries.

For each false claim, we searched for relevant documents using the following five queries: (i) the manually preprocessed claim as is, (ii) the preprocessed claim with date restriction, (iii) the preprocessed claim with named entities and date restriction, (iv) the first VP and all preceding tokens with date restriction, and lastly (v) the first VP and all preceding tokens with named entities and date restriction. For the true claims, due to wider coverage that led to easier retrieval, we only ran two queries, using the claim with and without date restriction.

We combined the results from all queries, and we kept a maximum of ten documents per claim. If the retrieved documents exceeded this limit, we only kept documents from news sources⁶, or from sources used in previous work on Arabic stance detection (Baly et al., 2018; Khouja, 2020). If we still had more than ten documents after filtering by source, we ranked the documents by their TF.IDF similarity with the claim, and we kept the top ten documents. We limited the number of documents to ten per claim in order to avoid having claims with very high numbers of documents and others with only one or two documents. Ultimately, this helped us keep the dataset balanced in terms of both sources and topics.

3.3 Stance Annotation

We set up the annotation task as follows: given a claim–article pair, what is the stance of the document towards the claim? The stance was to be annotated using one of the following labels: *agree*, *disagree*, *discuss*, or *unrelated*, which were also used in previous work (Pomerleau and Rao, 2017; Baly et al., 2018).

⁶We used Google News as a reference of news sources for the three countries of the fact-checking websites: <https://news.google.com/?hl=ar&gl=X&ceid=X%3Aar>, where **X** is **AE**, **EG**, or **SA**, standing for UAE, Egypt and Saudi Arabia, respectively.

Claim	Document Title	A1	A2	A3
اللحوم والدواجن المستوردة من البرازيل فاسدة <i>Meat and poultry imported from Brazil are rotten</i>	فضيحة البرازيل: اللحوم الفاسدة تهدد الانتعاش الاقتصادي <i>Brazil scandal: Rotten meat threatens economic recovery</i>	D	A	D
تشارك مصر في عملية انقاذ أطفال الكهف في تايلاند <i>Egypt participates in the rescue operation of cave children in Thailand</i>	الإنقاذ يبدأون عملية لإخراج صبية من داخل كهف في تايلاند <i>Rescuers begin operation to remove the teenage boys from the cave in Thailand</i>	D	U	U

Table 2: Disagreement between the annotators on the *discuss* (*D*) label with the *agree* (*A*) (first example) and the *unrelated* (*U*) labels (second example).

We explained the labels to annotators as follows:

- *agree*: the document agrees with the main claim in the statement clearly and explicitly;
- *disagree*: the document disagrees with main claim in the statement clearly and explicitly;
- *discuss*: the document discusses the same event without taking a position towards its validity;
- *unrelated*: the document talks about a different event, regardless of how similar the two events might be.

Our annotators were three graduate students in computer science and linguistics, all native speakers of Arabic. We adopted guidelines similar to the ones introduced by Baly et al. (2018). First, we conducted a pilot annotation round on 315 claim–article pairs, where each pair was annotated by all annotators. The annotators agreed on the same label for 220 out of the 315 pairs (70% of the pairs), while for 89 pairs (28%) there were two annotators agreeing on the label, and for the remaining 6 pairs (2% of the pairs) there was a three-way disagreement. The main disagreements between the annotators were related to the *discuss* label, which was confused with either *agree* or *unrelated*.

We show two examples in Table 2 where the annotators labeled the example on the top of the table as *discuss* and *agree*. The two annotators that labeled this example as *discuss* justified their choice by arguing that the document only mentioned the claims without agreeing or disagreeing and mainly analyzed the impact of rotten meat on Brazil’s economy in great detail. The example in the bottom of the table was labeled by one annotator as *discuss* and by two annotators as *unrelated*. The annotators who labeled it as *unrelated* argued that there was no mention of Egypt’s involvement in the rescue efforts, while the annotator who labeled the pair as *discuss* maintained that the document discussed the same event of children trapped in the cave.

These disagreements were resolved through discussions between the annotators, which involved refining the guidelines to label a pair as *discuss* if it only talks about the exact same event of the claim without taking any clear position. The annotators were also asked not to take into consideration any other factors, e.g., the date of article, its publisher, or its veracity.

For the rest of the data, each claim–article pair was annotated by two annotators, where the differences were resolved by the third annotator. This is very similar to labeling all pairs by three annotators with majority voting, but with less labor requirements. We measured the inter-annotator agreement (IAA) using Fleiss kappa, which accounts for multiple annotators (Fleiss and Cohen, 1973), obtaining an IAA of 0.67, which corresponds to substantial agreement.

3.4 Statistics About the Final Dataset

Table 3 shows the number of claims and articles for each website with their veracity label (by-publisher) and final stance annotations. The distribution of the four stance classes in training, development, and test is shown in Table 4. After selecting the gold annotations, we discarded all claims that had all of their retrieved documents labeled as *unrelated*, aiming to reduce the imbalance with respect to the *unrelated* class, and we only focused on claims with related documents, which can be seen as a proxy for check-worthiness. We ended up with a total of 4,063 claim–articles pairs based on 910 claims: 606 false and 304 true. The dataset is imbalanced towards the false claims, but as our main task is stance detection rather than claim veracity, we aimed at having a balanced distribution for the four stance labels. As shown in Table 4, around half of the labels are from the *unrelated* class, but it is common for stance detection datasets to have higher proportion of this class (Pomerleau and Rao, 2017; Baly et al., 2018).

Source	Veracity	Claims	Articles	Stance			
				Agree	Disagree	Discuss	Unrelated
ARANNEWS	False	170	518	80	82	51	305
DABEGAED	False	278	1,413	225	249	143	796
NORUMERS	False	158	490	26	103	32	329
REUTERS	True	304	1,642	691	15	161	775
Total	–	910	4,063	1,022	449	387	2,205

Table 3: Statistics about the number of claims, articles and claim–article pairs and the distribution of their stances for each source.

Label	Train	Dev	Test
Agree	739	129	154
Disagree	309	76	64
Discuss	247	70	70
Unrelated	1,553	294	358
Total	2,848	569	646

Table 4: Statistics about the claim–article pairs with stances in the training, development and test sets.

There are various approaches that can mitigate the impact of the class imbalance caused by the *unrelated* class. These are related to (i) task setup, (ii) modeling, and (iii) evaluation.

First, the task can be approached differently by only doing stance detection on the three related classes (Conforti et al., 2018), or by merging the *discuss* and the *unrelated* classes into one class, e.g., called *neutral* or *other* (Khouja, 2020).

Second, it is possible to keep all classes, but to train a two-step model: first to predict related vs. unrelated, and then, if the example is judged to be related, to predict the stance for the three related classes only (Zhang et al., 2019).

Third, one could adopt an evaluation measure that rewards models that make correct predictions for the related classes more than for the *unrelated* class. Such a measure was adopted by the Fake News Challenge (Pomerleau and Rao, 2017). However, such measures have to be used very carefully, as they might be exploited. For example, it was shown that the FNC measure can be exploited by random prediction from the related classes and never from the *unrelated* class, which has a lower reward under the FNC evaluation measure (Hanselowski et al., 2018). We leave such considerations about the impact of class imbalance to future work.

4 Experimental Setup

4.1 External Datasets

We experimented with a number of BERT-based models, pre-trained on Arabic or on multilingual data, which we fine-tuned and applied to our dataset, as well as to the following two Arabic stance detection datasets for comparison purposes:

- **Baly et al. (2018) Dataset.** This dataset has 1,842 claim–article pairs for training (278 *agree*, 37 *disagree*, 266 *discuss*, and 1,261 *unrelated*), 587 for development (86 *agree*, 25 *disagree*, 73 *discuss*, and 403 *unrelated*), and 613 for testing (110 *agree*, 25 *disagree*, 70 *discuss*, and 408 *unrelated*).
- **Khouja (2020) Dataset.** This dataset has 2,652 claim–article pairs for training (903 *agree*, 1,686 *disagree*, and 63 *other*), 755 for development (268 *agree*, 471 *disagree*, and 16 *other*) and 379 for testing (130 *agree*, 242 *disagree*, and 7 *other*).

The dataset by Baly et al. (2018) has 203 true claims from REUTERS and 219 false claims from the Syrian fact-checking website VERIFY-SY,⁷ which focuses on debunking claims about the Syrian civil war. Thus, the dataset contains claims that focus primarily on war and politics. They retrieved the articles and performed manual annotation of claim–article pairs for stance, following a procedure that is very close to the one we used for AraStance. Moreover, their dataset has annotations of rationales, which give the reason for selecting an *agree* or a *disagree* label. The dataset has a total of about 3,000 claim–article pairs, 2,000 of which are from the *unrelated* class. The dataset comes with a split into five folds of roughly equal sizes. We use folds 1-3 for training, fold 4 for development, and fold 5 for testing.

⁷<http://www.verify-sy.com/>

The dataset by [Khouja \(2020\)](#) is based on sampling a subset of news titles from the Arabic News Text (ANT) corpus ([Chouigui et al., 2017](#)), and then making true and false alterations of these titles using crowd-sourcing. The stance detection task is then defined between pairs of original news titles and their respective true/false alterations. This essentially maps to detecting paraphrases for true alterations (stance labeled as *agree*) and contradictions for false ones (stance labeled as *disagree*). They further have a third stance label, *other*, which is introduced by pairing the alterations with other news titles that have high TF.IDF similarity with the news title originally paired with the alteration. Overall, [Khouja \(2020\)](#)’s dataset is based on *synthetic statements* that are paired with *news titles*. This is quite different from AraStance and the dataset of [Baly et al. \(2018\)](#), which have *naturally occurring claims* that are paired with *full news articles*. Moreover, as both AraStance and [Baly et al. \(2018\)](#)’s datasets have naturally occurring data from the web, they both exhibit certain level of noise and irregularities, e.g., some very long documents, words/characters in other languages such as English, etc. Such a noise is minimal in [Khouja \(2020\)](#)’s dataset, which is a third differentiating factor compared to the other two datasets. Nevertheless, we include [Khouja \(2020\)](#)’s dataset in our experiments in order to empirically test the impact of these differences.

4.2 Models

We fine-tuned the following four models for each of the three Arabic datasets:

1. Multilingual BERT (mBERT), base size, which is trained on the Wikipedias of 100 different languages, including Arabic ([Devlin et al., 2019](#)).
2. ArabicBERT, base size, which is trained on 8.2 billion tokens from the OSCAR corpus⁸ as well as on the Arabic Wikipedia ([Safaya et al., 2020](#)).
3. ARBERT, which is trained on 6.2 billion tokens of mostly Modern Standard Arabic text ([Abdul-Mageed et al., 2020](#)).
4. MARBERT, which is trained on one billion Arabic tweets, which in turn use both Modern Standard Arabic and Dialectal Arabic ([Abdul-Mageed et al., 2020](#)).

⁸<http://oscar-corpus.com>

The four models are comparable in size, all having a *base* architecture, but with varying vocabulary sizes. More information about the different models can be found in the original publications about them. We fine-tuned each of them for a maximum of 25 epochs with an early stopping patience value of 5, a maximum sequence length of 512, a batch size of 16, and a learning rate of $2e-5$.

5 Results

The evaluation results are shown in Tables 5 and 6 for the development and for the test sets, respectively. We use accuracy and macro-F1 to account for the different class distributions; we also report per-class F1 scores. Note that [Khouja \(2020\)](#) uses three labels rather than four, merging *discuss* and *unrelated* into *other*. Their label distribution has a majority of *disagree*, followed by *agree*, and very few instances of *other*, which is different from our dataset and from [Baly et al. \(2018\)](#)’s.

We can see that ARBERT yields the best overall and per-class performance on dev for the [Khouja \(2020\)](#) dataset and AraStance. It also generalizes very well to the test sets, where it even achieved a higher macro-F1 score for the [Khouja \(2020\)](#) dataset. The performance of the other three models (mBERT, ArabicBERT, and MARBERT) drops slightly on the test set compared to dev for both AraStance and the [Khouja \(2020\)](#) dataset. This might be due to ARBERT being pre-trained on more suitable data, which includes Books, Gigaword and Common Crawl data primarily from MSA, but also a small amount of Egyptian Arabic. Since half of our data comes from an Egyptian website (DABEGAD), this could be helpful. Indeed, while ArabicBERT is pretrained on slightly more data than ARBERT, it was almost exclusively pretrained on MSA, without dialectal data, and AraStance it performs worse.

About the other models: The datasets on which ArabicBERT was trained have duplicates, which could explain the model being outperformed. For MARBERT, it is pretrained on tweets that have both MSA and dialectal Arabic. MARBERT’s data come from social media, which is different from the news articles or titles from which all the experimental downstream three datasets are derived. Also, it seems that ARBERT and MARBERT are better than the other two models at predicting the stance between a pair of sentences, as it is the case with the [Khouja \(2020\)](#) dataset.

Model	Baly et al. (2018) Dataset						Khouja (2020) Dataset					AraStance					
	A	D	Ds	U	Acc	F1	A	D	O	Acc	F1	A	D	Ds	U	Acc	F1
mBERT	.63	0	.11	.84	.73	.40	.74	.84	.76	.81	.78	.81	.68	.58	.92	.82	.75
ArabicBERT	.58	.14	.24	.82	.69	.45	.74	.86	.84	.82	.81	.85	.75	.56	.92	.84	.77
ARBERT	.56	.14	.30	.83	.70	.46	.81	.89	.87	.86	.86	.85	.82	.60	.93	.86	.80
MARBERT	.44	.14	.23	.78	.62	.40	.80	.88	.79	.85	.82	.85	.80	.53	.89	.84	.77

Table 5: Results on the development set for the three Arabic Stance Detection datasets. Shown are the F1-scores for each class (A: Agree, D: Disagree, Ds: Discuss, U: Unrelated, O: Other), as well as the overall Accuracy (Acc), and the Macro-Average F1 score (Macro-F1).

Model	Baly et al. (2018) Dataset						Khouja (2020) Dataset					AraStance					
	A	D	Ds	U	Acc	F1	A	D	O	Acc	F1	A	D	Ds	U	Acc	F1
mBERT	.64	0	.12	.85	.73	.40	.67	.81	.86	.76	.78	.83	.77	.51	.93	.85	.76
ArabicBERT	.66	.35	.27	.80	.67	.52	.72	.85	.71	.81	.76	.84	.74	.52	.94	.85	.76
ARBERT	.65	.29	.27	.81	.68	.51	.80	.89	1.0	.86	.90	.85	.78	.55	.92	.85	.78
MARBERT	.51	0	.25	.77	.60	.38	.78	.88	.92	.84	.86	.86	.72	.41	.90	.84	.72

Table 6: Results on the test set for the three Arabic Stance Detection datasets. Shown are the F1-scores for each class (A: Agree, D: Disagree, Ds: Discuss, U: Unrelated, O: Other), as well as the overall Accuracy (Acc), and the Macro-Average F1 score (Macro-F1).

This could be due to the diversity of their pre-training data, which improves the model’s ability to capture inter-sentence relations such as paraphrases and contradictions. Another factor that could explain ARBERT’s better performance compared to MARBERT is that the latter is trained with a masking objective only, while ARBERT is trained with both a masking objective *and* a next sentence prediction objective. The use of the latter objective by ARBERT could explain its ability to capture information in our claim–stance pairs, although these pairs are different from other types of pairs such as in the question and answer task, where the pair occurs in an extended piece of text.

On the other hand, there is no consistently best model for the Baly et al. (2018) dataset. This could be due to a number of reasons. First, that dataset has a severe class imbalance, as we have explained in Section 4. Second, the dataset (especially the false claims) is derived from one particular domain, i.e., the Syrian war, which might not be well represented in the pretraining data. Therefore, additional modeling considerations such as adaptive pretraining on a relevant unlabelled corpus before fine-tuning on the target labeled data could help.

Surprisingly, ArabicBERT and ARBERT perform much better on the test set than on the development set of the Baly et al. (2018) dataset for the *disagree* class, which has the lowest frequency: from 0.14 F1 to 0.29–0.35 F1.

Since the number of *disagree* instances is very low (25 documents for 10–12 unique claims), it is possible that the claims in the test set happen to be more similar to the ones in the training data than it is for development. This is plausible because we did our train-dev-test split based on the five-folds prepared by the authors as explained in Section 4. It is worth noting that the multilingual model (mBERT) has the highest overall accuracy and F1 score for the *unrelated* class of the Baly et al. (2018) dataset. Multilingual text representations such as mBERT might over-predict from the majority class, and thus would perform poorly on the two low-frequency classes; indeed, mBERT has an F1-score of 0 for *disagree*, and no more than 0.12 for *discuss* on development and testing.

Finally, we observe very high performance for all models for the *unrelated* class of AraStance. This could be an indication of strong signals that differentiate the *related* and the *unrelated* classes, whereas the *discuss* class is the most challenging one in AraStance, due to its strong resemblance to *agree* in some examples such as the one shown in Table 2. This indicates that all models offer an area for improvement, where a single classifier can excel for both frequent and infrequent classes for the stance detection within and across datasets. We leave further experimentation, including with models developed for FNC and the Baly et al. (2018) dataset, for future work.

6 Conclusion and Future Work

We presented AraStance, a new multi-topic Arabic stance detection dataset with claims extracted from multiple fact-checking sources across three countries and one news source. We discussed the process of data collection and approaches to overcome challenges in related document retrieval for claims with low online presence, e.g., due to topic or country specificity. We further experimented with four BERT-based models and two additional Arabic stance detection datasets.

In future work, we want to further investigate the differences between the three Arabic stance detection datasets and to make attempts to mitigate the impact of class imbalance, e.g., by training with weighted loss, by upsampling or downsampling the classes, etc. We further want to examine the *discuss* class across datasets and to compare the choice of annotation scheme—three-way vs. four-way—on this task. Moreover, we plan to enrich AraStance by collecting more true claims from other websites, thus creating a dataset that would be more evenly distributed across the claim veracity labels. Furthermore, we would like to investigate approaches for improving stance detection by extracting the parts of the documents that contain the main stance rather than truncating the documents after the first 512 tokens. Finally, we plan to experiment with cross-domain (Hardalov et al., 2021a) and cross-language approaches (Mohtarami et al., 2019).

Acknowledgements

Tariq Alhindi is supported by the KACST Graduate Studies Scholarship. Muhammad Abdul-Mageed gratefully acknowledges support from the Social Sciences and Humanities Research Council of Canada through an Insight Grant on contextual misinformation detection for Arabic social media, the Natural Sciences and Engineering Research Council of Canada, Canadian Foundation for Innovation, Compute Canada, and UBC ARC-Sockeye.⁹ Preslav Nakov is supported by the Tanbih mega-project, which is developed at the Qatar Computing Research Institute, HBKU, and aims to limit the impact of “fake news,” propaganda, and media bias by making users aware of what they are reading. We would also like to thank Firas Sabbah for his help in collecting the false and true claims, and the anonymous reviewers for their helpful feedback.

⁹<https://doi.org/10.14288/SOCKEYE>.

References

- Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC '16, pages 4445–4452, Portorož, Slovenia.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. *arXiv preprint arXiv:2101.01785*.
- Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 21–27, New Orleans, Louisiana, USA.
- Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. ANERsys: An Arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, CI-CLING '07, pages 143–153.
- Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. 2018. Robust document retrieval and individual evidence modeling for fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and VERification*, FEVER '18, pages 127–131, Brussels, Belgium.
- Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. ANT corpus: an Arabic news text collection for textual classification. In *Proceedings of the 14th International IEEE/ACS Conference on Computer Systems and Applications*, AICCSA '17, pages 135–142, Hammamet, Tunisia.
- Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards automatic fake news detection: cross-level stance detection in news articles. In *Proceedings of the First Workshop on Fact Extraction and VERification*, FEVER '18, pages 40–49, Brussels, Belgium.
- Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, pages 69–76, Vancouver, Canada.
- 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*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA.

- William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT '16*, pages 1163–1168, San Diego, California, USA.
- Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.
- Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval '19*, pages 845–854, Minneapolis, Minnesota, USA.
- Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING '18*, pages 1859–1874, Santa Fe, New Mexico, USA.
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021a. Cross-domain label-adaptive stance detection. *arXiv preprint arXiv:2104.07467*.
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021b. A survey on stance detection for mis- and disinformation identification. *arXiv preprint arXiv:2103.00242*.
- Ganesh Jawahar, Muhammad Abdul-Mageed, and VS Laks Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING '20*, pages 2296–2309, Barcelona, Spain (Online).
- Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification, FEVER '20*, pages 8–17, Online.
- Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL '14*, pages 55–60, Baltimore, Maryland, USA.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16*, pages 31–41, San Diego, California, USA.
- Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT '18*, pages 767–776, New Orleans, Louisiana, USA.
- Mitra Mohtarami, James Glass, and Preslav Nakov. 2019. Contrastive language adaptation for cross-lingual stance detection. 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 '19*, pages 4442–4452, Hong Kong, China.
- El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, Tariq Alhindi, and Hasan Cavusoglu. 2020. Machine generation and detection of Arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop, ANLP '20*, pages 69–84, Barcelona, Spain (Online).
- Dean Pomerleau and Delip Rao. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake News Challenge*.
- Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval '20*, pages 2054–2059, Barcelona (online).
- Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*.
- Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. 2021. Automatic fake news detection with pre-trained transformer models. In *Proceedings of the ICPR International Workshops and Challenges*, pages 627–641.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification, FEVER '18*, pages 1–9, Brussels, Belgium.
- Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, Maryland, USA.
- Qiang Zhang, Shangsong Liang, Aldo Lipani, Zhaochun Ren, and Emine Yilmaz. 2019. From stances' imbalance to their hierarchical representation and detection. In *Proceedings of the World Wide Web Conference, WWW '19*, pages 2323–2332, Lyon, France.