Stance Prediction and Claim Verification: An Arabic Perspective

Jude Khouja
Latynt
jude@latynt.com

Abstract

This work explores the application of textual entailment in news claim verification and stance prediction using a new corpus in Arabic. The publicly available corpus comes in two perspectives: a version consisting of 4,547 true and false claims and a version consisting of 3,786 pairs (claim, evidence). We describe the methodology for creating the corpus and the annotation process. Using the introduced corpus, we also develop two machine learning baselines for two proposed tasks: claim verification and stance prediction. Our best model utilizes pretraining (BERT) and achieves 76.7 *F1* on the stance prediction task and 64.3 *F1* on the claim verification task. Our preliminary experiments shed some light on the limits of automatic claim verification that relies on claims text only. Results hint that while the linguistic features and world knowledge learned during pretraining are useful for stance prediction, such learned representations from pretraining are insufficient for verifying claims without access to context or evidence.

1 Introduction

Although fake news is not an emerging phenomenon and has been documented throughout history, the prevalence and wide spread of misinformation over the internet has captured significant proportion of public attention in recent years. This is in part linked to the low barrier for content generation through the advent of the internet and social media (Allcott and Gentzkow, 2017) and the fact that false news spread faster than true news (Vosoughi et al., 2018) rendering it increasingly dangerous to public discourse. The widespread exposure in the U.S. for example has been reported by researchers who found that the average American encountered between one and three stories from known publishers of fake news during the month before the 2016 election (Allcott and Gentzkow, 2017).

Since manual fact-checking by human experts does not scale well with the amount of information shared on the web, there is a growing body of work in recent years aimed at developing automatic tools to target fake news, misinformation and credibility of content on social media in general (Rubin et al., 2016; El Ballouli et al., 2017; Baly et al., 2018a,b; Wang et al., 2018; Saleh et al., 2019; Zhang et al., 2019). Several datasets were developed to further aid research on this topic[1] (Darwish et al., 2017; Wang, 2017; Baly et al., 2018b; Thorne et al., 2018). We refer readers to (Thorne and Vlachos, 2018; Pierri and Ceri, 2019) for a more comprehensive overview of recent research on fake news, propaganda and misinformation.

Despite the increased attention, most of the work has been focusing on the English language. Tools, resources and datasets available in Arabic are limited (Darwish et al., 2017; Baly et al., 2018b; Elsayed et al., 2019). As such, this work contributes to recent efforts targeting Arabic by introducing a new publicly available corpus in Arabic that is suitable to study claim verification and semantic entailment (Katz, 1972).

2 Related Work

In recent years, there has been rapid progress in developing systems and tools for automatic fact checking and claim verification. Various approaches were developed which relied on a diverse set of methods and information to verify claims. Most relevant to this work are approaches that used content such as textual information in the title and/or body of the claims to predict their veracity. Among this direction of research those that considered a machine learning approach (Potthast et al.,

[1] FNC: http://www.fakenewschallenge.org/

Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER), pages 8–17
July 9, 2020. ©2020 Association for Computational Linguistics

Given a news title, write two news titles that:

A- Paraphrase the original title:
Has same meaning but is worded differently by
rephrasing and changing Syntax, using verb
synonyms, using different words to describe the
same information such as locations, counts and dates.

B- Contradict the original title:
Looks similar to the original title but has
contradicting meaning (both cannot be true in the
same context) by reversing meaning without
negating main verb, using antonym of main verb
with rephrasing, changing key information using world
knowledge such as locations, counts and dates.

Table 1: Guidelines for rewriting news titles.

2017; Wang et al., 2018; Alzanin and Azmi, 2019) including deep learning techniques (Hanselowski et al., 2017; Baly et al., 2018b; Popat et al., 2018; Chawla et al., 2019; Helwe et al., 2019; Lv et al., 2019).

Datasets: There are limited but growing datasets related to claim verification (Al Zaatari et al., 2016; Darwish et al., 2017; Wang, 2017; Baly et al., 2018b; Thorne et al., 2018; Alkhair et al., 2019; Alzanin and Azmi, 2019; Elsayed et al., 2019). However, datasets focusing on Arabic remain scarce (Darwish et al., 2017; Baly et al., 2018b; Elsayed et al., 2019). Recently, work on the application of textual entailment for claim verfication has been explored and new datasets combining stance prediction and claim verfication were introduced (Baly et al., 2018b; Thorne et al., 2018).

This work is most in line with that direction. We developed a new corpus in Arabic that can be used jointly for claim verification and textual entailment recognition. However, our new corpus differs from the aforementioned datasets in that it is at the sentence level, hence, we are disentangling the tasks of claim verification and textual entailment from the task of evidence extraction (Information Retrieval) and focusing on the former. We also start from real news titles and generate true/false claims from them. Our aim is to mitigate one type of bias that results from starting with fake news collected in the wild: bias in the distribution of topics among the true/false claims. While some forms of biases about the world are useful in determining the veracity of a claim, some can be problematic. We can imagine a dataset that contains more positive[2] news in the ”fake” class than in the ”true” class. A system trained on such data could predict the class ”fake” with higher confidence for any claim that has a positive tone compared to one that has a negative or neutral tone. Such surface level biases in topics and linguistic styles could arguably result in models that do not generalize well.

3 The corpus

In this part, we describe our Arabic News Stance (ANS) corpus.[3] We derived two perspectives of the corpus suitable for claim verification and stance classification. Please refer to Appendix A to read our data statement about the corpus.

3.1 Data Collection

In contrast to Baly et al. (2018b) and more in line with Thorne et al. (2018), we start with true news titles (reference) and generate fake/true claims from them. The corpus generating process can be summarized in two stages: 1) generating true/false modifications of existing news titles through crowdsourcing; and 2) validating the generated claims by annotating them in a separate phase.

We derive our corpus by sampling a subset of news titles from the most recent version of the Arabic News Texts (ANT) corpus (Chouigui et al., 2017); A collection of Arabic news from multiple news media sources in the Middle East. The dataset was suitable for our task as it covers several topics of news (politics, sports, *etc.*) sourced from several credible mainstream news outlets (BBC, CNN, Al Arabiya, *etc.*). The following is an example of a news title from this dataset:

حقائق سقوط صخرة تزن ١٠٠ كغ من الحائط
الغربي بالقدس

“Facts about the falling of a boulder weighing 100 kg. of the west wall in Jerusalem.”

Generating true/false claims We used crowdsourcing to generate true/false claims. Starting from a news title, we recruited annotators to modify each news title into a new claim. For true claims, annotators were asked to paraphrase the original sentence by changing its syntax and wording while maintaining the integrity of the information. We allowed for the use of world knowledge to modify the information. For example, replacing cities with

[2] Positive here refers to sentiment

[3] Data available at: https://github.com/latynt/ans

Type	Translation	Arabic
Reference	**Wall Street records largest losses in 6 weeks**	وول ستريت تسجل أكبر خسائر في ٦ أسابيع
Paraphrase	Losses in Wall Street are the highest in 6 weeks	خسائر في وول ستريت هي الأعلى في ستة اسابيع
Contradiction	Profits in Wall Street in the last six weeks	مكاسب في وول ستريت في الاسابيع الستة الاخيرة
Reference	**Death of a journalist who reported on Russian Mercenaries in Syria in mysterious circumstances**	وفاة صحفي كتب عن المرتزقة الروس في سوريا في ظروف غامضة
Paraphrase	Death of a journalist in mysterious circumstances after he reported on Russian Mercenaries in Syria	وفاة صحفى فى ظروف غامضة بعد أن كتب عن المرتزقة الروس فى سوريا
Contradiction	Death of a journalist after battling with illness	وفاة صحفى بعد صراع مع المرض
Reference	**5.5 Billion withdrawn from emerging markets by investors in one week**	٥.٥ مليار دولار سحوبات المستثمرين من الأسواق الناشئة بأسبوع
Paraphrase	Nearly 6 Billion withdrawn in a week from emerging markets	قرابة ستة مليار دولار سحوبات أسبوع في الأسواق الناشئة
Contradiction	Almost a million in withdrawals from emerging markets	سحوبات حوالي المليون في الأسواق الناشئة

Table 2: Examples of modifications by annotators. Green highlights a change in line with reference. Red highlights a conflicting part of the sentence with the reference sentence.

countries and celebrities and politicians with their nationalities.

For false claims, to insure that the modification results in meaningful mutation of the semantic information, the instructions (Table 1) stated that the modified sentence should contradict the original title in such a way that both cannot simultaneously be true in the same context. Annotators were asked to avoid simple negation and were encouraged to use different strategies for modifying the sentences. Our analysis of a sample of the collected data showed that different annotators utilized different strategies at different rates. For example, some annotators predominantly altered years, counts and locations that appeared in the original titles while others modified the semantics of the modified sentences to have opposite meaning (detained vs. released, supported vs. opposed, etc.).

We relied on Amazon Mechanical Turk[4] and Upwork [5] to recruit annotators. We only considered Arabic native speakers for news title rewriting. All annotators had to pass a language qualification test similar to our task. Data was randomly assigned to annotators in batches of 500. To insure the quality of the generated data, we sampled data during the annotation from each batch and re-annotated any batch containing errors in more than 10% of the sample by resending the batch to the annotator after explaining the errors. See Table 2 for examples of generated claims using different modification strategies.

3.2 Data Validation And Analysis

To evaluate the quality of our data, we performed a second round of annotation on the generated news titles. We derived a new task in which annotators were presented with a pair of sentences and asked to supply a hypothesis about how they are semantically related. This task is related to the the semantic concepts of entailment and contradiction (Katz, 1972; Bowman et al., 2015) but with the aim of validating our generated ture/fake claims. We highlight a notable difference compared to other work on stance classification. In contrast to the commonly used four classes adopted in other datasets [6] *(agree, contradict, discuss, unrelated)*, we elect to merge labels *(discuss, unrelated)* into one *(other/not enough information)* resulting in three classes – *paraphrase, contradiction, other/not enough information* for each pair of news titles. Our motivation is that despite the general value of discriminating between irrelevant documents[7] (unrelated) and documents that are related to the claim but do not make a stance about the claim (discuss), both classes represent the same position in the context of stance prediction. We, therefore, treat them as one class. We found that this is also similar to the approach by Thorne et al. (2018).

To present annotators with a small set of the third class (other), we first considered randomly pairing news titles from our corpus. We hypothesized that randomly paired news titles will be discussing unrelated news and would naturally be assigned the label *other* by annotators. However and upon examining examples of this method, we noticed that telling the (other) class apart from the two classes was dis-proportionally trivial since the randomly paired sentences differed significantly

[4] https://www.mturk.com

[5] https://www.upwork.com/

[6] For example: Fake News Challenge (FNC)

[7] Documents in this case refer to sentences but could be any body of text. Hence, in this work we use both terms interchangeably.

Number of Annotators	#	%
3	2594	60.9%
4	1239	29.1%
5	426	10.0%
Annotator Labels Overlap	#	%
$< 75\%$	470	11.0%
75 - 99%	210	4.9%
100%	3579	84.0%
Majority/Author Labels Overlap	#	%
Majority Label = Author's Label	3766	99.4%
Majority Label $\neq$ Author's Label	23	0.6%
Fleiss k		
3 total annotators		0.83
4 total annotators		0.81
5 total annotators		0.83

Table 3: Statistics for the annotation results. The author's label is the label obtained from the worker who rewrote the news title. Majority label is the consensus of 75% or higher of the annotators.

(discussed different topics and contained no overlapping words) compared to pairs from the *paraphrase, contradict* classes. Predicting this class, therefore, can be reduced to checking for the absence of overlap in words from the paired titles. As an alternative selection criteria to random pairing, we used a similarity metric to select pairs that look more similar. We calculate the F1 score of overlapping ngrams in the paired titles weighted by the ngram size similar to Trinh and Le (2018). In our case however, we consider ngrams at the character level given the short length of the sentences. We included ngrams of size 2 to 6 and set the minimum score to 0.1.

A total of 4,259 pairs were labeled by 3 to 5 annotators. We considered the author's rewritten sentences as labels (for the *paraphrase* and *contradict* classes). Table 3 shows the annotation statistics. The Fleiss k scores (calculated separately for examples labeled by 3, 4 and 5 annotators) show overall a very high level of agreement (> 0.81) suggesting that the quality of the dataset is sufficiently high. For the final data, we included only pairs with inter annotator agreement of 75% or higher, hence, dismissing all data with 2 out of 3 majority vote or worse.

Figures 1 and 2 provide some details about the length of the written claims in the final dataset compared to the original reference sentences. We noticed that on average, claims are shorter than the original references with contradicting claims being shorter than paraphrasing claims. This could be due to workers aiming to minimize time spent per each example. Another likely explanation is the fact that contradicting a statements by replacing or removing key words is easier than paraphrasing a statement.

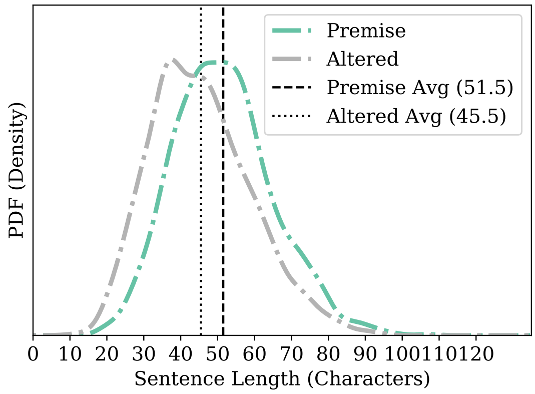

Figure 1: Length of sentences in dataset (rewritten vs. reference)

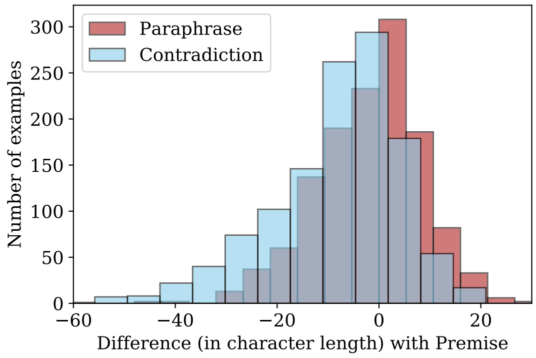

Figure 2: Comparison in rewritten sentences

4 Experiments

In this section, to demonstrate the utility of the corpus, we derive two tasks useful for evaluating news veracity and stance prediction and develop two baselines to evaluate on the proposed tasks. We describe the proposed tasks and details of the baselines in this section and the results in section 5.

4.1 Tasks

Claim Only Verification: In this setting, we explore the task of verifying claims based only on information in the claims themselves. In our corpus, we assess the veracity of a claim c_i from our corpus D based solely on the textual information of the claim. The task is, hence, a binary classification where an estimator needs to map an input to a label Y which can be either *fake* or *not fake*:

Class	#	%
Not Fake	3072	67.6%
Fake	1475	32.4%

Table 4: Class distribution for claim verification. (#: total number of examples. %: percentage of all data)

$$p(Y|c_i), \qquad c_i \in D$$

We consider all original news titles (reference sentences) in our corpus to belong to the *not fake* class. We rely on the fact that the reference sentences originated from reputable mainstream media in the Middle East. Our *fake* class examples consist of the sentences corrupted by annotators that passed the data validation process described in Section 3.1. Table 4 shows the distribution of classes for this task.

It is important to discern the limited scope in defining news veracity in this work: the incorrectness of the corrupted sentence is not a universal statement about the claim. We note the fact that several of the corrupted sentences can be factual/not fake in other contexts. As such, we consider them fake in regards to the related event/context - in this case our reference sentence (original news titles). Further analysis exposed two instances where the modified sentences matched other original news titles. Both examples were excluded from the corpus for this task. However, such cases hint at the limits of claim verification using claim text only. We further explore this in section 5 and share some insights.

Stance Prediction This task is a direct reflection of our annotation process. Given a reference sentence r_i and a claim c_i, predict the label Y (Agree, Contradict, Other/Not enough information) from the claim/reference pair (c_i, r_i).

$$p(Y|c_i, r_i), \qquad (c_i, r_i) \in D$$

Table 5 shows the distribution of classes in our corpus for the stance prediction task.

4.2 Methods

We evaluate two baselines on both tasks. For modeling, we considered two classes of models that have been largely adopted by the NLP community. The models are described in the next section.

Class	#	%
Disagree	2399	63.4%
Agree	1301	34.4%
Other	86	2.3%

Table 5: Class distribution for stance prediction. (#: total number of examples. %: percentage of all data)

Recurrent Perspective Matching: Our first baseline is a simple RNN model that uses Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) as the main building block to encode the input. LSTM models encode the input sequentially and can model temporal dependencies useful to semantic tasks. In our implementation for both tasks, we consider both character level and word level representations of the input sentence(s) separately. In each case, we represent every input word/character with a unique d-dimensional vector that is learned during training. These vectors are then passed through the LSTM layer in sequence and the output of the last step (at the end of the sentence) is used as the encoding of the sentence(s). For the claim verification task, the claim encoding $\overrightarrow{h}_t$ can be described by:

$$\overrightarrow{h}_t = \overrightarrow{LSTM}(\overrightarrow{h}_{t-1}, x_t) \qquad t = 1, ..., M_i$$

Where M_i is the length size of the sentence corresponding to example i and x_t is the character/word at position t.

In stance prediction, the input consists of a pair of sentences (reference r, claim c). Each is encoded using the same LSTM layer to obtain their encoding:

$$\begin{aligned} \overrightarrow{r}_t &= \overrightarrow{LSTM}(\overrightarrow{r}_{t-1}, x_t) \qquad t = 1, ..., M_i^r \\ \overrightarrow{c}_t &= \overrightarrow{LSTM}(\overrightarrow{c}_{t-1}, x_t) \qquad t = 1, ..., M_i^k \end{aligned}$$

To obtain the interaction representation $\overrightarrow{h}_t$, $\overrightarrow{\mathrm{r}_t}$ and $\overrightarrow{\mathrm{c}_t}$ are multiplied element-wise. We experimented with *cosine* similarity and concatenation and found the element-wise multiplication and concatenation to work slightly better than *cosine* similarity:

$$\overrightarrow{h}_t = (\overrightarrow{\mathrm{r}_t} \circ \overrightarrow{k}_t)$$

The resulting encoding in both tasks $\overrightarrow{h}_t$ is then passed through a linear layer with non-linearity

	(dev)				(test)			
Claim Verification	**Acc.**	**Prec.**	**Rec.**	F_1	**Acc.**	**Prec.**	**Rec.**	F_1
Majority Class	68.1	34.1	50.0	40.5	67.1	33.6	50.0	40.2
LSTM character level								
char, 10(emb), 100(hid), 0(dropout)	70.2	65.7	56.8	55.4	67.3	60.2	54.6	52.5
char, 10(emb), 100(hid), 30.0(dropout)	70.6	67.9	56.5	54.6	67.8	61.3	55.1	53.1
LSTM word level								
word, 50(emb), 50(hid), 0(dropout)	68.1	60.4	54.8	52.9	65.8	57.2	53.9	52.4
word, 50(emb), 100(hid), 0(dropout)	68.6	61.8	56.4	55.5	64.5	55.4	53.3	52.1
Stance Prediction								
Majority Class	62.4	20.8	33.3	25.6	63.8	21.3	33.3	26.0
LSTM character level								
char, 10(emb), 50(hid), 0(dropout)	62.2	20.7	33.3	25.6	64.4	21.5	33.3	26.1
char, 50(emb), 50(hid), 0(dropout)	62.4	20.8	33.3	25.6	64.1	21.4	33.3	26.0
char, 50(emb), 50(hid), 30.0(dropout)	62.5	43.0	33.7	26.6	64.4	46.4	34.0	27.5
LSTM word level								
word, 10(emb), 50(hid), 0(dropout)	62.1	38.7	39.2	38.8	62.0	37.8	38.1	37.8
word, 50(emb), 50(hid), 30.0(dropout)	63.0	39.9	40.7	40.3	59.8	37.4	38.2	37.8

Table 6: Results for the claim verification and stance prediction Tasks.

(*ReLu*) followed by a $softmax$ function to convert the output to probabilities for each class:

$$p(Y = c|h_i) = softmax(ReLu(W_c\overrightarrow{h_i} + b_c))$$

W_c and b_c are learnable parameters associated with each class c in the corresponding task.
Prediction in both tasks is done by selecting the label with the highest probability:

$$\arg\max_c p(Y = c|h_i)$$

Pretrained Transformer: Pretraining and transfer learning (Devlin et al., 2018a; Peters et al., 2018; Radford et al., 2019) has recently gained attention as a popular approach to acquiring universal linguistic features useful in many downstream NLP tasks and was shown to be successful in improving on the state of the art in many downstream NLP tasks with minimal fine-tuning. Lv et al. (2019) have successfully explored BERT for the task of fake news detection in English and proposed an extension that improves on fine-tuned BERT. In addition to the aforementioned supervised methods, we evaluate BERT (Devlin et al., 2018a) on both tasks in our corpus. We are not aware of any other work that explored pretraining for claim verification and stance prediction in Arabic.

BERT is based on the Transformer model first introduced by Vaswani et al. (2017). Transformer-based models have recently become common in many NLP tasks including question answering and entailment classification (Devlin et al., 2018b; Radford, 2018). For both tasks, we utilize a publicly available implementation that has been trained on a multilingual dataset including Arabic.[8]. We elect to adhere to the proposed approach recommended by Devlin et al. (2018a) for future reproducibility. Since our implementation is identical to the one provided by the authors, we will omit the detailed description of the model architecture and refer readers to (Vaswani et al., 2017)[9].

Task	**Prec.**	**Rec.**	F_1
Claim Verification			
Fake	51	55	53
Not Fake	77	75	76
Macro Avg.	**64.1**	**64.6**	**64.3**
Stance Detection			
Agree	65	63	64
Disagree	80	81	80
Other	86	86	86
Macro Avg.	**76.8**	**76.6**	**76.7**

Table 7: Results of using pretraining (BERT) on claim verification and stance prediction tasks.

5 Results

For the recurrent perspective models, we trained all models for 100 epochs using Adam optimizer (Kingma and Ba, 2014) with 0.001 learning rate. We conducted hyper-parameter tuning on the de-

[8] We use BERT-Base, Multilingual Cased: 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters

[9] See also:
http://nlp.seas.harvard.edu/2018/04/03/attention.html

Prediction	Label	Translation	Arabic
Fake	Fake	Historic agreement between Europe and Japan to support trump	اتفاق تاريخي بين أوروبا و اليابان لمساعدة ترامب
Fake	True	Historic agreement between Europe and Japan to confront trump	اتفاق تاريخي بين أوروبا و اليابان لمواجهة ترامب
True	Fake	First women's interest channel in Gaza soon to see the light	أول قناة تلفزيونية نسائية في غزة تظهر للنور قريبا
True	True	First women's interest channel in Gaza faces uncertain fate	أول قناة تلفزيونية نسائية في غزة تواجه مصيرا مجهولا
Fake	True	Ethiopia assures Egypt of its Nile share	أثيوبيا تؤكد حرصها على حصة مصر بالنيل
Fake	Fake	Ethiopia apathetic about Egypt's right of the Nile water	أثيوبيا غير معنية بحصة مصر من مياه النيل

Table 8: Examples of claim verification task predictions using fine-tuned BERT highlighting the model's invariant labels for similar sentences with different meanings.

velopment set. For the pretrained BERT model, we fine-tune on our data for 3 epochs using BERT BPE units.

Table 6 shows the top results of all experiments for both tasks. We report the accuracy and F_1 (Macro unweighted average). In the claim verification task, results show that in general, word based models perform comparably to character based models but we note that all results do not provide significant gains (53.1 vs. 40.2 F_1) compared to the baseline (majority class) which could be explained by the small training data size but might hint at an ill-defined task. We explore this further below. In the stance prediction task, experiments show word based models outperform character based models (37.8 vs 27.5 F_1). This could be due to the limited size of our corpus which is not sufficient for character based models to learn words and phrases from scratch and capture the semantic representation needed for stance prediction.

Results for the pretraining experiments (shown in Table 7) show significant improvement of the pretrained model over the models trained only on our corpus. This is similar to findings by Lv et al. (2019). However, the improvement is disproportionally larger in the stance prediction task (76.7 vs. 37.8 F_1) and the large gains do not carry over to the claim verification task (64.3 vs. 53.1 F_1). The imbalance in gains also confirms our intuition about the limitation of claim only verification which we discuss next.

Limits Of Claim Only Verification: We briefly mentioned in Section 4.1 the limited scope of claim verification in a setting were the decision about the veracity of a claim can be made using only the text of the claim. We hypothesize that the task might not be learnable through a direct mapping from the claim text to the veracity space. Given that the initial results of the fine-tuned BERT model supported this intuition, we elected to manually inspect a sample of the predictions and noticed that in many cases the model was predicting the same label for claims that look similar but are semantically different. We share a sample of these cases in Table 8. This suggests that while the linguistic features learned during pretraining were useful for textual entailment (stance prediction task), the veracity of a claim cannot be made using only implicit world knowledge learned during pretraining. A simple example highlighting this limitation is the reference news title الذهب يصعد مع تراجع الدولار *"Gold prices increase amidst a falling dollar."*' and its contradicted rewritten version "أسعار الذهب تهبط عالمياْ" *"Gold prices fall globally"*. Here, it is easy to argue that the contradiction can be true in another context and hence, a decision about the veracity of this claim should only be made in reference to a particular context/event. We believe that explicitly associating each claim with evidence or context is the more appropriate approach for claim verification.

These initial experiments suggest that discriminate models trained using claim only information might rely on biases in the topics, linguistic styles, tones and implicit world-knowledge learned from training data to make predictions. Results of the performance of such models could, therefore, be inflated if the training data is not uniformly distributed across languages, topics, writing styles, political ideologies etc. While we believe that our dataset collection process which yields classes that share the same distribution of topics and news sources mitigated these types of biases, we also note that the annotation process and human factor introduced other types of biases that could be present in the data.

6 Conclusion

In this work we presented a new publicly available corpus for textual entailment and its use in studying misinformation in the Arabic language. We shared some insights about the creation of the corpus and the baselines developed to evaluate the corpus. We further explored the use of pretraining (Devlin et al., 2018a) and developed a strong baseline for our tasks. Our experiments additionally shed light on the limits of ”claim-only” misinformation detection methods that rely solely on the stated claims without use of accompanying evidence. We hope to explore this further in future work. As we plan to also explore the use of generated data in studying the robustness of misinformation detection methods against adversarial data with varying linguistic styles, political ideologies and world-knowledge.

Acknowledgements

We are grateful for Ayah Zirikly, Bart Desmet and René F. Kizilcec for their valuable insights, critical commentary and helpful discussions during the course of this work. We also thank our anonymous reviewers for their thoughtful comments and suggestions.

References

Ayman Al Zaatari, Rim El Ballouli, Shady ELbassouni, Wassim El-Hajj, Hazem Hajj, Khaled Shaban, Nizar Habash, and Emad Yahya. 2016. Arabic corpora for credibility analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4396–4401.

Maysoon Alkhair, Karima Meftouh, Kamel Smaïli, and Nouha Othman. 2019. An arabic corpus of fake news: Collection, analysis and classification. In *Arabic Language Processing: From Theory to Practice*, pages 292–302. Springer International Publishing.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *J. Econ. Perspect.*, 31(2):211–236.

Samah M Alzanin and Aqil M Azmi. 2019. Rumor detection in arabic tweets using semi-supervised and unsupervised expectation–maximization. *Knowledge-Based Systems*, 185:104945.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. Predicting factuality of reporting and bias of news media sources.

Ramy Baly, Mitra Mohtarami, James Glass, Lluis Marquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus.

Emily Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6(0):587–604.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. Learning natural language inference from a large annotated corpus. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Piyush Chawla, Diego Esteves, Karthik Pujar, and Jens Lehmann. 2019. SimpleLSTM: A Deep-Learning approach to Simple-Claims classification. In *Progress in Artificial Intelligence*, pages 244–255. Springer International Publishing.

A Chouigui, O B Khiroun, and B Elayeb. 2017. ANT corpus: An arabic news text collection for textual classification. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 135–142.

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*, pages 145–148.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: Pre-training of deep bidirectional transformers for language understanding.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Rim El Ballouli, Wassim El-Hajj, Ahmad Ghandour, Shady Elbassuoni, Hazem Hajj, and Khaled Shaban. 2017. CAT: Credibility analysis of Arabic content on twitter. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 62–71, Valencia, Spain. Association for Computational Linguistics.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the CLEF-2019 CheckThat! lab: Automatic identification and verification of claims.

Andreas Hanselowski, P V S Avinesh, Benjamin Schiller, and Felix Caspelherr. 2017. Description of the system developed by team athene in the fnc-1. *Fake News Challenge*.

Chadi Helwe, Shady Elbassuoni, Ayman Al Zaatari, and Wassim El-Hajj. 2019. Assessing arabic weblog credibility via deep co-learning. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 130–136.

S Hochreiter and J Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Jerrold J Katz. 1972. *Semantic Theory*. New York: Harper & Row.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Zhengwei Lv, Duoxing Liu, Haifeng Sun, Xiao Liang, Tao Lei, Zhizhong Shi, Feng Zhu, and Lei Yang. 2019. AUTOHOME-ORCA at SemEval-2019 task 8: Application of BERT for fact-checking in community forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 870–876.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Francesco Pierri and Stefano Ceri. 2019. False news on social media: A Data-Driven survey.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using Evidence-Aware deep learning.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. 2019. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. (Nips).

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD ’18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. Tanbih: Get to know what you are reading.

A Data Statement

In line with recent efforts addressing ethical issues that can result from the use of data and technology and following the recommendations of Bender and Friedman (2018), we are sharing the following information that we believed is relevant to our dataset and the collection process. We encourage future use of the data to include a summary of this information.

A.1 Language Variety

To study and build tools in the areas of stance prediction and claim verification. Data was selected from news titles and rewritten by annotators for the purpose of generating statements and statement pairs. Part of the dataset was a random subset of the ANT corpus which was created through webcrawling news sources in the Middle East. As different tools and annotation were included in the creation of the data, we expect the distribution of topics, opinions and language to incorporate different types and levels of bias. To the best of our knowledge, the data is in Standard Arabic (’arb’) with few exceptions such as abbreviations. At least Latin script (’Latn’) is present.

A.2 Annotator Demographic

A total of 8 crowd-source workers mostly from the Middle East contributed to the annotations. Annotators were selected based on their fluency in the Arabic language. Demographic information was not available at the time annotation for all recruited individuals. Of the information available, we are aware of at least 1 woman, 2 men and 3 individuals who are Arabic native speakers.

A.3 Text Characteristics

The dataset includes a subset of the news titles from ANT news corpus (v2.1) which included 5 news sources (BBC, Al Arabiya, CNN, Sky News, France24) and 6 categories (culture, economy, international news, Middle East, sport, technology) collected from February 2018 to October 2018. Data also includes rewritten versions of the news titles by the annotators following the provided guidelines (see Table 1).