Machine Generation and Detection of Arabic Manipulated and Fake News

El Moatez Billah Nagoudi[1], **AbdelRahim Elmadany**[1], **Muhammad Abdul-Mageed**[1], **Tariq Alhindi**[2], **Hasan Cavusoglu** [3]
[1] Natural Language Processing Lab,
[1,3] The University of British Columbia
[2] Department of Computer Science, Columbia University
[1] {moatez.nagoudi,a.elmadany,muhammad.mageed}@ubc.ca,
[2] tariq@cs.columbia.edu, [3] cavusoglu@sauder.ubc.ca

Abstract

Fake news and deceptive machine-generated text are serious problems threatening modern societies, including in the Arab world. This motivates work on detecting false and manipulated stories online. However, a bottleneck for this research is lack of sufficient data to train detection models. We present a novel method for automatically generating Arabic manipulated (and potentially fake) news stories. Our method is simple and only depends on availability of true stories, which are abundant online, and a part of speech tagger (POS). To facilitate future work, we dispense with both of these requirements altogether by providing AraNews, a novel and large POS-tagged news dataset that can be used off-the-shelf. Using stories generated based on AraNews, we carry out a human annotation study that casts light on the effects of machine manipulation on text veracity. The study also measures human ability to detect Arabic machine manipulated text generated by our method. Finally, we develop the first models for detecting manipulated Arabic news and achieve state-of-the-art results on Arabic fake news detection (macro $F_1 = 70.06$). Our models and data are publicly available.

1 Introduction

The last few years witnessed a striking rise in creation and dissemination of fake news (Egelhofer and Lecheler, 2019; Allcott et al., 2019). Such fake stories are propagated not only by individuals, but also by groups or even nation states (Allcott et al., 2019). For example, Allcott and Gentzkow (2017) discuss the role fake news have played in the 2016 U.S. presidential election, arguing that Donald Trump's voters have been more influenced to believe fake stories. More recently, concerns have also been raised about possible abuse of machine-generated text such as by GPT3 (Brown et al., 2020) for deceiving readers.

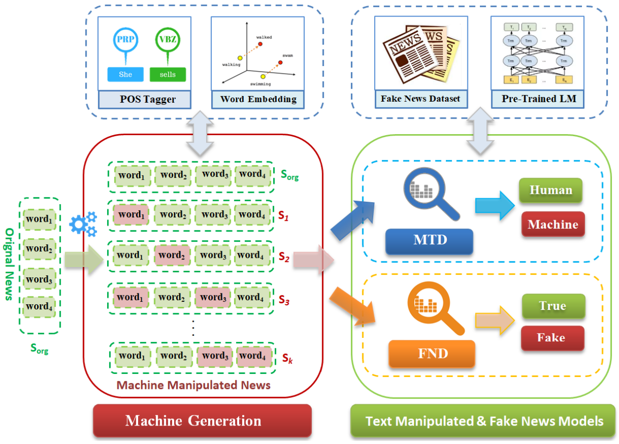

Figure 1: Our proposed methods. **Left:** Machine generation of manipulated text. **Top Right:** manipulated text detection model (MTD). **Bottom Right:** fake news detection model (FND). **word$_i$** : original word. **word$_j$** : substituted word.

In the Arab context, Arab countries have had their share of misinformation. This is especially the case due to the sweeping waves of uprisings and popular protests (Torres et al., 2018; Helwe et al., 2019). Although there has been considerable research investigating the

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 69–84
Barcelona, Spain (Online), December 12, 2020

legitimacy, or lack thereof, of news in many languages (Conroy et al., 2015; Kim et al., 2018; Bondielli and Marcelloni, 2019), work on the Arabic language is still lagging behind.

In this paper, we first report an approach to automatically generate manipulated (and possibly fake) stories in Arabic. Our approach is simple: Given a dataset of legitimate news, a part-of-speech (POS) tagger, and a word embedding model, we are able to automatically generate significant amounts of news stories. Since these generated stories are machine manipulated such that original words (e.g., named entities, factual information such as numbers and time stamps) are substituted, some of these stories can be used as training data for fakes news detection models.

To illustrate our method, we provide the following scenario: Given a human-authored sentence, we output a manipulated version of the original. The veracity of the manipulated version can either: (1) **Stay Intact.** For instance when changing an adjective with its synonym, e.g., أفضل ("top") with أحسن ("best") in أفضل هاتف ذكي هو الأيفون ("The best smartphone is the iPhone") or (2) **Change.** For example, when substituting a named entity with another that does not necessarily communicate the meaning of the original as closely. For example, changing the named entity أرامكو ("Aramco") with أمازون ("Amazon") in أرامكو تحقق هذه السنة أعلى أرباح ("Aramco achieved the highest profit this year").

As such, we emphasize that changing a certain POS does not automatically flip the sentence veracity. For example changing مصر ("Egypt") with المحروسة ("Almahrousa") does not alter the sentence veracity. We manually validate the claim that our method of text manipulation can generate fake stories via a human annotation study (Section 5). We then use our generated data to create models that can detect manipulated stories from our method and empirically show the impact of exploiting our generated stories on the fake news detection task on a manually-crafted external dataset (Section 6). We make our models and data publicly available.[1]

We make the following contributions: (1) We introduce AraNews, a new large-scale POS-tagged news dataset covering a wide range of topics from diverse sources. (2) We propose a simple, yet effective, method for automatic manipulation of Arabic news texts. Applying this methods on AraNews, we create and release the first dataset of manipulated Arabic news dataset to accelerate future research. (3) We perform a human annotation study to measure the ability of native speakers of Arabic to detect (a) machine manipulated and (b) fake news stories without resorting to external resources such as fact checking websites. The annotation study aims at gauging the extent to which a human can fall prey to deceptive news in a semi-real situation (i.e., where an average reader do not check third party sources when reading through a news story). (4) We develop effective models for detecting manipulated news stories, and then test the utility of our generated data for improving fake news detection on an external dataset.

The rest of the paper is organized as follows: Section 2 provides an overview of related work. In Section 3, we describe the two *true*[2] news datasets used in this work. Section 4 is about our methods for generating manipulated text (and potentially fake news stories). Section 5 describes our human annotation study. In Section 6, we present our detection models. We conclude in Section 7.

2 Related Work

Knowledge-Based Fact Checking. Recent work on developing automatic methods for fake news detection has mainly followed two lines of research as categorized in the literature (Thorne and Vlachos, 2018; Potthast et al., 2018). First, work that compares a claim against an evidence from (trusted) collections of factual information whether the evidence is a sentence (i.e. fact-checking modeled as textual entailment) or a full document (i.e. stance detection between a claim-document pair). This includes work that created synthetic claims verified against Wikipedia (Thorne et al., 2018), and naturally occurring claims verified against news articles (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017), discussion forums (Joty et al., 2018), or debate websites (Chen et al., 2019). These datasets are labeled using 2 tags (*true*, *false*) (Alhindi et al., 2018) 3 tags (*supported*, *refuted*, *not-enough-information*) (Thorne et al., 2018), or 4 tags (*agree*, *disagree*, *discuss*, *unrelated*) (Pomerleau and Rao, 2017). They vary in size from 300 claims (Fer-

[1] Models and data are at: `https://github.com/UBC-NLP/wanlp2020_arabic_fake_news_detection`.

[2] We use the terms "true" and "legitimate" interchangeably to refer to stories that are not "fake".

reira and Vlachos, 2016) to 185,000 claims (Thorne et al., 2018). Approaches on developing models to predict claim veracity using these datasets include hierarchical attention networks (Ma et al., 2019), pointer networks (Hidey et al., 2020), graph-based reasoning (Zhou et al., 2019; Zhong et al., 2019), and (similar to our methods) fine-tuning of pre-trained transformers (Hidey et al., 2020; Zhong et al., 2019).

Style-Based Detection. The second line of research focuses on analyzing the linguistic features of a claim to determine its veracity without considering external factual information. This approach is based on investigating linguistic characteristics of fake content in comparison to true content. In news and various fact-checked political claims, Rashkin et al. (2017) found that first and second person pronouns, superlatives, modal adverbs, and hedging are more prevalent in fake content, while concrete and comparative figures, and assertive words are more widespread in truthful content. Other work found the properties of deceptive language to differ between domains (Pérez-Rosas et al., 2018). Misleading content itself has been classified into sub-categories such as (a) the 3 types of fake (serious fabrication, hoaxes, and satire) (Rubin et al., 2015), (b) propaganda and its different techniques (Da San Martino et al., 2019), and (c) misinformation and disinformation (Ireton and Posetti, 2018). The differences between these different categories depend on many factors such as genre and domain, targeted audience, and deceptive intent (Rubin et al., 2015; Rashkin et al., 2017). In addition to categories, truth was classified to more than two *levels*. For example, `Politifact.com` introduced 6 levels: pants-on-fire, false, mostly-false, half-true, mostly-true and true. These different levels have been exploited in previous work, with a goal to automate this more challenging six-way classification task (Rashkin et al., 2017; Wang, 2017; Alhindi et al., 2018).

Automatic Generation of Data. The development of automatic fake news detection models was possible as the afore-mentioned datasets became available. More related to our work, previous work has focused on developing methods to automatically generate more robust, and large-scale, fake news datasets. Thorne et al. (2019) showed that current fact-checking systems are vulnerable to adversarial attacks by doing simple alteration to the training data. To increase robustness of such systems, previous work has extended available fake news datasets both manually and automatically using lexical substitution (Alzantot et al., 2018), rule-based alterations (Ribeiro et al., 2018), phrasal addition and temporal reasoning (Hidey et al., 2020), or using transformer models such as GPT-2 (Radford et al., 2019) and Grover (Zellers et al., 2019) for claim and news article generation (Niewinski et al., 2019; Zellers et al., 2019). As a way to increase our understanding and trust in fact-checking systems, Atanasova et al. (2020) developed a transformer-based model for generating fact-checking textual explanations along with the prediction of claim veracity.

Arabic Work. All of the datasets described above, however, are in English with limited availability of similar ones in other languages such as Arabic. Available Arabic datasets cover tasks such as determining claim check-worthiness of tweets (Barrón-Cedeño et al., 2020), news and claims from fact-checking websites (Elsayed et al., 2019), and translated political claims from English (Nakov et al., 2018). In addition, there are datasets for stance and factuality prediction of claims from news or social media with or without the evidence retrieval task (Baly et al., 2018; Khouja, 2020; Elsayed et al., 2019; Alkhair et al., 2019; Darwish et al., 2017). These corpora are created by either using credibility of publishers as proxy for veracity (*true*/*false*) then manually annotating the stance between a claim-document pair (*agree*, *disagree*, *discuss*, *unrelated*) (Baly et al., 2018) or by manual alteration of true claims to generate fake ones about the same topic (Khouja, 2020)–all requiring a manual, slow, and labor-intensive process. We alleviate this by introducing our simple and scalable approach for automatic generation of Arabic manipulated text, including potential fake stories, using the abundant legitimate online news data as seeds for the generation model. We also introduce a large-scale dataset in true and manipulated form for detection work. We now introduce our datasets.

3 Datasets

3.1 ATB: Arabic TreeBank

We exploit a number of Arabic Treebank datasets from the Linguistic Data Consortium (LDC). Namely, we use 4 LDC resources comprising Arabic news stories in Modern Standard Arabic (MSA). These are: Arabic Treebank (ATB) Part 1 v4.1 (LDC2010T13), Part 2 v3.1 (LDC2011T09), Part 3 v3.2 (LDC2010T08) and Broadcast News v1.0 (LDC2012T07), the latter being a collection of Arabic news stories built as part of of the DARPA TIDES project.[3] These 4 parts contain over $2,000$ news stories produced by a handful of Arabic news services with a total of 1.5M tokens. Moreover, we use the Arabic Treebank Weblog (LDC2016T02), which contains 13K Arabic news and a total of 308K tokens. We refer to all the 5 LDC resources collectively as **ATB**. For each token in ATB, there is a Latin-based transliteration, a unique identifier (lemma ID), a breakdown of the constituent morphemes (prefixes, stem, and suffixes), POS tag(s), and the corresponding English gloss(es).

3.2 AraNews: A New Large-Scale Arabic News Dataset

In order to study misinformation in Arabic news, we develop, **AraNews**, a large-scale, multi-topic, and multi-country Arabic news dataset. To create the dataset, we start by manually collecting a list of 50 newspapers belonging to 15 Arab countries, the United States of America (USA), and the United Kingdom (UK). Then, we scrape the news articles from this list of newspapers. Ultimately, we collected a total of $5,187,957$ news articles. The map in Figure 2 shows the geographic distribution of AraNews.

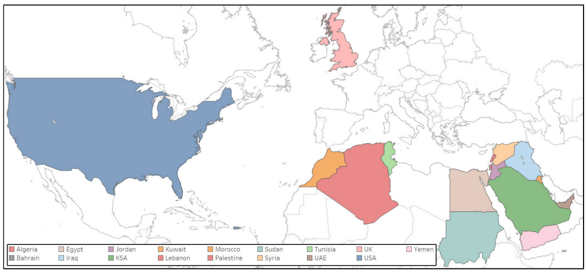

Figure 2: Geographical distribution of AraNews.

We assign each article in AraNews a thematic category as follows: We first consider the category assigned on each newspaper website to the article. We identify a total of 118 unique categories, which we manually map to only 17 categories using the dictionary illustrated in Table A2 in Appendix A.2. The 17 categories are in the set {*Politics, History, Society, Media, Entertainments, Weather, Sports, Social Media, Heath, Culture and Art , Economy, Religion, Education, Technology, Fashion, Local News, International News*}. For each article in the AraNews collection, we document several types of information. These include: (1) name of the newspaper in Arabic and English, (2) newspaper origin country, (3) newspaper link, (4) title, (5) content, (6) summary (if available), (7) author (if available), (8) URL, (9) date, and (10) topic. More details about AraNews are in Table A1 in Appendix A.1. AraNews is available for research.[4]

4 Methods

To generate a large scale manipulated news dataset, we exploit ATB (see Section 3.1) and 1M news articles extracted from AraNews (Section 3.2). In the following, we describe our data splits and methodology for automatically generating manipulated text from these two 'legitimate' sources.[5]

4.1 Data Splits

We split both ATB and AraNews at the article level into TRAIN, DEV, and TEST. Table 1 provides the related statistics at both the article and sentence levels across the different data sources for all three splits.

4.2 POS Tagging

The first step in our approach is to perform POS tagging of the news articles. ATB is already POS tagged. Thus, we use MADAMIRA (Pasha et al., 2014), a morphological analysis and disambiguation tool for Arabic, to POS-tag AraNews.[6]

[3] https://www.ldc.upenn.edu/collaborations/past-projects.

[4] https://github.com/UBC-NLP/wanlp2020_arabic_fake_news_detection.

[5] We do not check the veracity of stories in these two sources, but we have no reason to think they may have fake stories. As such, we make the assumption they consist of "true" stories.

[6] MADAMIRA was trained on the training sets of Penn Arabic Treebank corpus (parts 1, 2 and 3) (Maamouri et al., 2004) and the Egyptian Arabic Treebanks (Maamouri et al., 2014).

Data	**TRAIN** (80%)			**DEV** (10%)			**TEST** (10%)		
	Artic.	Sent.	Tokens	Artic.	Sent.	Tokens	Artic.	Sent.	Tokens
ATB Weblog	1.9K	8.6K	154.8K	235	1K	17.9K	235	1.2K	19.5K
ATB Part 1	587	4.7K	117.3K	73	580	14.9K	74	536	13.3K
ATB Part 2	400	3.4K	117.6K	50	387	13.9K	51	382	12.7K
ATB Part 3	479	10.5K	268.5K	60	1.5K	36.5K	60	1.4K	34.7K
ATB BN	96	21.5K	334.3K	12	3.1K	47.7K	12	2.4K	40K
AraNews	800K	3.3M	1.1B	100K	55.1K	209.9M	100K	61.6K	197.2M

Table 1: Statistics of ATB and AraNews (only 1M articles) datasets across the data splits.

4.3 News Word Embedding Model

The second component needed in our model is a word vector model. We train a fastText model (Joulin et al., 2016) on a concatenation of MSA data sources (Wikipedia Arabic,[7] Arabic Gigaword Corpus (Parker et al., 2009), and ATBP1V3 [8]). We perform light pre-processing involving removing punctuation marks, non-letters, URLs, emojis, and emoticons. We also convert elongated words back to their original form by reducing consecutive repetitions of the same character as suggested in (Lachraf et al., 2019). For example : استفسااااار (*inquiries*) and الجزائــــــــر (*Algeria*) are converted to استفسار and الجزائر. We then train our model using the Python Gensim library (Řehřek and Sojka, 2011). We set the vector size to 300, minimum word frequency at 100, and a window size of 5 words. We call this model **AraNewsEmb**. We then use this model to retrieve the most similar tokens of a given token in the original text using cosine similarity. Next, we use one of the set of relevant tokens to replace the original token, focusing only on tokens corresponding to the following POS tags: proper nouns (N_PROP), cardinal numbers (N_NUM), common adjective (ADJ), comparative adjective (ADJ_COMP), ordinal numbers (ADJ_NUM), and negative particles (NEG_PART). In theory, substitution of these words should have no syntactically harmful effect on the sentence. However, changes can happen if the gold or predicted POS tag is wrong.

4.4 Automatic Text Manipulation

To generate a machine manipulated story, we substitute the selected words (ones matching the listed POS tags) by a chosen one from the k most similar (k-closest) words in our AraNewsEmb model as described in (Nagoudi and Schwab, 2017). We remove negation from the sentence, using the negative particle (NEG_PART) POS as a guide, and substitute the cardinal number related to (N_NUM) with a random number. For tokens related to the rest of POS tags, we needed to identify a reasonable *character-level* similarity threshold between the original token and the retrieved most-similar token to ensure the two belong to different lemmas. [9]

We performed a manual analysis based on $5,000$ random substitution examples from AraNewsEmb and identify a similarity $ratio$ of 50%. This threshold gave us new words in 100% of the cases. For instance, if we want to substitute the word لبنان (Lebanon), we exclude three words: ولبنان, لبنانيا, بلبنان, before considering the 4^{th}-closest word which is سوريا (Syria). Other examples for the substitution process are illustrated in Table 3. We also provide in Table 2 the average number of k-closest words excluded in each POS class. The results of this step are two new machine manipulated datasets. We refer to these datasets as **ATB⁺** and **AraNews⁺**. More details about these two datasets are in Table A3 in Appendix A.3. We now provide an example illustrating how our text manipulation method works.

POS Label	Count	Avg	Median
ADJ	$99,538$	3.79	3.00
ADJ_COMP	$4,513$	2.81	2.00
ADJ_NUM	$5,752$	3.06	3.00
N_NUM	$60,615$	0.55	0.00
N_PROP	$75,771$	2.93	2.00

Table 2: Descriptive statistics of $k-$closest words excluded in each POS class. We simply remove the negation token corresponding to **NEG_PART** from the sentence, and so the embedding model is not used in this case.

[7] https://archive.org/details/arwiki-20190201.

[8] https://catalog.ldc.upenn.edu/LDC2010T08

[9] We use the following formula to compute the character-level similarity ratio between two tokens: $ratio = 2 * M/T$, where M is matching characters and T is total of characters.

Word	Translation	POS	*k*-closet (ratio similarity%)	Token rank
قصير	Short	ADJ	طويل (25%)	0
أكثر	More	ADJ_COMP	واكثر (89%), أقل (28%)	1
باكستان	Pakistan	N_PROP	لباكستان (93%), اوزباكستان (82%), بنغلاديش (26%)	2
الثالث	The third	ADJ_NUM	والثالث (92%), الثاني (67%), الاول (72%) , الرابع (49%)	4
وسبعة	And seven	N_NUM	وسبع (89%), وسبعه (80%), وسبعون (73%) , وثلاثون (17%)	4

Table 3: Illustration of substitution process based on the word embeddings model. **Token rank:** refers to rank of chosen word in the returned word embedding list (from AraNewsEmb) after applying our char-based cosine similarity threshold. **Light red:** excluded word. **Light green:** selected word. **Under lined** words represent the false negative of the selection process (i.e., words based on a different lemma and hence could work but were ignored by the algorithm).

4.5 Illustrative Example

We present a typical example illustrating the automatic text manipulation process by our method. Consider the following sports news sentence: محرز ينتقل الى برشلونة مقابل 120 مليون دولار ("Mahrez moves to Barcelona for $ 120 million"). The method proceeds in the following steps:

Step 1: Identify POS tags. The sentence can be POS-tagged as shown in Table 4.

Words		POS Tags
محرز	→	N_PROP
ينتقل	→	VERB
الى	→	PREP
برشلونة	→	N_PROP
مقابل	→	NOUN
120	→	NUM
مليون	→	NOUN
دولار	→	NOUN

Table 4: POS tags of our example

Step 2: POS and Token Selection. In this step, tokens corresponding to one or more POS tags must be chosen for substitution. For our illustrative example, we will select and substitute only the *proper noun* and *digit* tokens. The sentence has two proper nouns, برشلونة and محرز and one digit (120).

Step 3: Sentence Manipulation. If we select only the noun proper: برشلونة (Barcelona), we can retrieve the 5-closest words from AraNewsEmb. In this case, we obtain:مدريد (Madrid), ميلان (Milan), باريس (Paris), فالنسيا (Valencia), and مانشستر (Manchester). Indeed, we can generate 5 fake sentences from the original sentence. However, if we select two proper nouns برشلونة, محرز and the digit token 120, we can generate 75 (3 ∗ 5 ∗ 5) manipulated sentences from the single human sentence. Both scenarios are presented in Table 5.

Subs. with 5-closest of برشلونة	**Subs. with 5-closest of** برشلونة , محرز **and 120**
محرز ينتقل الى مدريد مقابل 120 مليون دولار	صلاح ينتقل الى ليدز مقابل 350 مليون دولار
محرز ينتقل الى باريس مقابل 120 مليون دولار	ميسي ينتقل الى مدريد مقابل 450 مليون دولار
محرز ينتقل الى فالنسيا مقابل 120 مليون دولار	رونالدو ينتقل الى باريس مقابل 155 مليون دولار
محرز ينتقل الى ميلان مقابل 120 مليون دولار	ماني ينتقل الى فالنسيا مقابل 280 مليون دولار
محرز ينتقل الى مانشستر مقابل 120 مليون دولار	اغويرو ينتقل الى مرسيليا مقابل 70 مليون دولار

Table 5: Illustrative output example from our text manipulation method. Given a sentence and a target POS tag, we substitute the word corresponding to the POS tag with the word closest to it (based on cosine similarity) in the AraNewsEmb model. **Left:** Substitution of word برشلونة (*Barcelona*) with its 5-closets words. **Right:** Substitution of برشلونة, محرز and *120* (*Barcelona, Mehrez* [name of a soccer player], and 120) each with 5-closest words.

5 Human Annotation Study

5.1 Annotation Data

We perform a human annotation study in order to identify (1) the ability of humans to detect machine manipulated text using our method, and (2) the extent to which text identified as machine manipulated can be *fake*. For this purpose, we randomly select 300 samples from the ATB development set (see Table 1), among which 145 sentences are from the original ATB sentences and the rest (i.e., 155 samples) are machine manipulated.

5.2 Annotation Procedures

For annotation, we follow two stages: The first stage is for **manipulated text detection**. We shuffle the samples and ask the annotators to label each sentence as either original/produced by humans (*human*) or generated by machine (*machine*). The second stage is for detecting **veracity of manipulated text**. This stage is applied only on the 155 machine manipulated sentences generated from ATB. Note that here we provide annotators with a sentence *pair* including the machine generated sentence itself and its human counterpart (original sentence in ATB). Annotators are then asked to compare the manipulated sentence to its original and assign the label *fake* if the manipulated sentence differs in meaningful ways (e.g., provides contradictory information) from the original, but a *true* label otherwise. That is, a *true* tag is assigned if difference between the sentence pair is only grammatical such as cases where the machine sentence is a paraphrase. Each sample is annotated by two experts, both of whom is native speakers of Arabic with a Ph.D. degree. Inter-annotator agreement in term of Kappa (κ) scores is 79.46% for *human* vs. *machine* and 81.07% for *fake* vs. *true*. As shown in Table 6, the substitution of tokens with the POS tags ADJ_NUM, NEG_PART, N_NUM, and N_PROP changes between 76.32% and 83.33% of sentence veracity. Meanwhile, changing tokens whose POS tags are ADJ or ADJ_COMP changes the veracity of the sentence less than 50% of the time. The reason is that the selected $k-$closets tokens in the second scenario is more or less of a paraphrase. Table 7 provides examples where annotators disagree on either or both tasks, and Table 8 illustrates cases where annotators agree.

			Annotators Agreement (%)		
		#Sent.	**Hum/Mach**	**True/Fake**	**%Fake**
Hum		145	97.93	N/A	N/A
Mach	**ADJ**	27	96.30	74.07	48.15
	ADJ_COMP	24	100	91.67	58.33
	ADJ_NUM	26	76.92	73.08	78.85
	NEG_PART	32	87.50	90.63	76.56
	N_NUM	19	100	73.68	76.32
	N_PROP	27	92.59	74.07	83.33
	Overall	155	94.67	80	70.32

Table 6: Percentages of inter-annotator agreement on a random sample of 300 sentences (original and manipulated).

Annotator 2		**Annotator 1**		**POS**	**Gold**	**Sentence**
T/F	**M/H**	**T/F**	**M/H**			
True	Mach	Fake	Mach	N_PROP	Hum	ذخائر القديسه تريز **المراهق** يسوع في القبيات واحتفالات دينيه تكريما لها حتي الخميس
					Mach	ذخائر القديسه تريز **الطفل** يسوع في القبيات واحتفالات دينيه تكريما لها حتي الخميس
True	Mach	Fake	Hum	N_PROP	Hum	هذه المؤامره تستهدف احباط العمليه السياسيه وارجاع **العراق** الي اه حكم اه العصابه البعثيه
					Mach	هذه المؤامره تستهدف احباط العمليه السياسيه وارجاع **الاردن** الي اه حكم اه العصابه البعثيه
Fake	Hum	True	Mach	ADJ	Hum	وتابع ان التوجه الاعلامي الجديد **مرفوض** والحكومه تطلب منا المستحيل
					Mach	وتابع ان التوجه الاعلامي الجديد **مضلل** والحكومه تطلب منا المستحيل
Fake	Hum	Fake	Mach	NEG_PART	Hum	واوضح ان مقعدين **لم** يتم البت بهما بعد في الاسكندريه وان ذلك عائد الي قرار قضائي
					Mach	واوضح ان مقعدين يتم البت بهما بعد في الاسكندريه وان ذلك عائد الي قرار قضائي
Fake	Hum	True	Mach	ADJ_NUM	Hum	ابيكم تعطوني سعر هالقطعه هذي ووين تركب حقت اي بندق **ثانية** من نفس الصناعه
					Mach	ابيكم تعطوني سعر هالقطعه هذي ووين تركب حقت اي بندق **رابع** من نفس الصناعه

Table 7: Examples of disagreement between annotators on either one or the two tasks.

Sentence	Gold	POS	Task Labels	
			H/M	T/F
وصدر بيان عن اتحاد علماء المسلمين في **العراق** جاء فيه	Hum	N_PROP	Hum	Fake
وصدر بيان عن اتحاد علماء المسلمين في **الاردن** جاء فيه	Mach			
حياك **الله** اخي الغالي	Hum	N_PROP	Hum	True
حياك **الرحمن** اخي الغالي	Mach			
هذه الصور اعادت الي اذهان مشاهد مقتل نحو واحد **وعشرين** ألف شخص	Hum	N_NUM	Hum	Fake
هذه الصور اعادت الي اذهان مشاهد مقتل نحو واحد **وثلاثين** ألف شخص	Mach			
احتل الهولندي كارستن المركز الاول في المرحله الثامنة من دوره فرنسا ال **٨١** للدراجات	Hum	N_NUM	Hum	Fake
احتل الهولندي كارستن المركز الاول في المرحله الثامنة من دوره فرنسا ال **٨٩** للدراجات	Mach			
يلف الغموض **العديد** من المشكلات في اداء الجيش الاسرائيلي	Hum	ADJ	Hum	True
يلف الغموض **الكثير** من المشكلات في اداء الجيش الاسرائيلي	Mach			
احترم نفسك **احسن** لك والا ساشن حمله لمقاطعه مدونتك ابداها بتوقيعات الزملاء السعوديين	Hum	ADJ	Hum	True
احترم نفسك **افضل** لك والا ساشن حمله لمقاطعه مدونتك ابداها بتوقيعات الزملاء السعوديين	Mach			
الصابونجي : الهجره المسيحيه **لا** تتصل بموضوع ديني	Hum	NEG_Part	Hum	Fake
الصابونجي : الهجره المسيحيه تتصل بموضوع ديني	Mach			
واضاف كورماك ان واشنطن تفكر في الخضوع في المستقبل في حال **لم** تنجح الضغوط السياسيه	Hum	NEG_Part	Hum	Fake
واضاف كورماك ان واشنطن تفكر في الخضوع في المستقبل في حال تنجح الضغوط السياسيه	Mach			

Table 8: Example labels from one annotator on a sample of our data.

6 Manipulated Text and Fake News Detection

6.1 Manipulated Text Detection (MTD)

Approach. We use the ATB⁺ and AraNews⁺ datasets for training deep learning models for detecting manipulated text. From each of these datasets, we select 61K human and 61K machine manipulated sentences (total $\sim$ 122K) and split them into 80% training (TRAIN), 10% development (DEV), and 10% test (TEST) as shown in Table 9.

# Split	Human	Machine Manipulated						
	# Sent.	ADJ	ADJ_COMP	ADJ_NUM	N_NUM	N_PROP	NEG_PART	Total
TRAIN	48,727	9,600	4,513	5,752	9,600	9,600	9,600	48,665
DEV	6,573	1,300	638	844	1,300	1,300	1,300	6,682
TEST	5,895	1,200	592	665	1,200	1,200	1,200	6,057

Table 9: The TRAIN, DEV, and TEST splits form ATB⁺ (with a similar split from AraNews⁺) for developing our manipulated news detection models. The same amount of data from the different POS categories is extracted from each of the two datasets.

Models. For the purpose of training our manipulated text detectors, we exploit 4 large pre-trained masked language models (MLM): mBERT (Devlin et al., 2018), AraBERT (Antoun et al., 2020), XLM-R$_{\text{Base}}$, and XLM-R$_{\text{Large}}$ (Conneau et al., 2020). [10]

Training Data & Hyper-Parameters. We fine-tuned all these models on the TRAIN split of (1) ATB^{+}, and (2) AraNews^{+}, independently. For each model, we run for 25 epochs with a batch size of 32, maximum sequence length of 128 tokens, and a learning rate of $1e^{-5}$.

[10] Each of the mBERT, AraBERT, and XLM-R$_{\text{Base}}$ models has 12 layers each with 12 attention heads, and 768 hidden units. The XLM-R$_{\text{Large}}$ model has 24 layers each with 16 attention heads, and $1,024$ hidden units.

Evaluation Data. We evaluate each of the two models on its respective DEV and TEST splits (i.e., from either ATB$^+$ or AraNews$^+$). Although the data in the two classes are reasonably balanced, we use *both* accuracy and macro F_1 for evaluation. Table 10 shows the results on the two datasets.

Results & Discussion. As Table 10 shows, the best performance on ATB$^+$ is at 83.20 F_1 (acquired with XLM-R$_{\text{Base}}$). For AraNews$^+$, the best model is at 89.25 F_1 (acquired with AraBERT). These results show that it is harder to detect manipulated text exploiting ATB$^+$ than that exploiting AraNews$^+$. This could be due to two reasons: (a) ATB$^+$ contains news stories that are diachronically different from the data the language models are trained on, which is less true for the case of AraNews (since the latter dataset is crawled in late 2019 and early while most ATB data were acquired prior to 2004), (b) ATB$^+$ is POS tagged manually, which makes generations based on it less error-prone.

Data	Models	Dev		Test	
		Acc.	F1	Acc.	F1
ATB$^+$	mBERT	77.16	77.08	77.42	77.36
	XLM-R$_{\text{Base}}$	81.72	81.72	**83.22**	**83.20**
	XLM-R$_{\text{Large}}$	82.41	82.38	81.38	81.36
	AraBERT	**83.19**	**83.17**	82.63	82.62
AraNews$^+$	mBERT	79.39	79.38	83.51	83.52
	XLM-R$_{\text{Base}}$	82.77	82.56	86.09	86.08
	XLM-R$_{\text{Large}}$	82.12	82.10	86.35	86.35
	AraBERT	**87.21**	**87.21**	**89.23**	**89.25**

Table 10: Performance results of our the MTD models on the dev and test split of ATB$^+$ and AraNews$^+$.

6.2 Fake News Detection (FND)

Approach. Evaluating on an external human-crafted fake news dataset, we also develop a host of models for detecting fake news. The dataset is developed by Khouja (2020) by sampling a subset of news titles from the Arabic News Texts corpus (Chouigui et al., 2017), a collection of Arabic news from multiple news media sources in the Middle East. Crowd-sourcing is used to generate true and false claims starting from a news title. Khouja (2020) asks annotators to modify each news title into a new claim by: (1) paraphrasing the original title via changing wording and syntax while maintaining the same meaning, thus producing a legit (or *true*) sentence, and (2) modifying the meaning of the original title such that a sentence that contradicts that title is acquired (constituting a false, or *fake*, claim). We refer to this dataset from (Khouja, 2020) as **`Khouja`**. It comprises $3,072$ *true* sentences and $1,475$ *fake* sentences. We now describe our various fake news detection models.

Models. As explained, our primary goal is to test how data generated by our methods will fare on the problem of fake news detection, as evaluated on a human-created fake news dataset (i.e., Khouja). For this reason, we only test models reported in this section on the DEV and TEST splits of Khouja. We have the following modeling settings:

1. **Fine-Tuning on Khouja (Baseline).** Here, we fine-tune all MLMs (i.e., models from Section 6.1) on the train split of Khouja.
2. **Zero-Shot Detection.** Based on our human annotation study (Section 5), we hypothesize that our machine-manipulated sentences will be closer to the *fake* class than the *true* class in the fake news context. To test this hypothesis, we fine-tune our MLMs *only* on our generated data (and hence naming this setting *zero-shot*, i.e., since we do not train on Khouja TRAIN at all). We have the following configurations pertaining the parts of our data we fine-tune on: (a) ATB$^+$ TRAIN, (b) AraNews$^+$ TRAIN, and (c) double the size the TRAIN of AraNews$^+$.
3. **Data augmentation.** We augment the Khouja TRAIN split with the 3 training configurations from our data listed in the zero-shot setting above (i.e., a, b, and c), each time fine-tuning on Khouja and one of these 3 splits.

Evaluation Data & Hyper-Parameters. For the current experiments, as explained earlier, we use the original split of Khouja (2020) (i.e., 80% TRAIN, 10% DEV, and 10% for TEST). We evaluate all the

FND models on the DEV and TEST splits of Khouja and use the same hyper-parameters as in Section 6.1.

Results & Discussion. As Table 11 shows, best performance when training on *Khouja TRAIN (gold, our baseline)* is 67.21 F_1 (acquired with XLM-R$_{\text{Large}}$). This is already 2.91% points higher than the best system reported by Khouja (2020) (64.30 F_1, not shown in Table 11).

For our *zero-shot experiments*, our best model is at 52.71 F_1 when training on AraNews$^+$ base setting (i.e., setting **a** in Table 11, with TRAIN data = 48,655 sentences). This result shows that use of data generated by our method is effective on the fake news detection task, even without access to gold training data. In particular, the 52.71 F_1 we acquire is higher than the baseline majority class in Khouja (2020) (40.20 F_1) and close to their 53.10 F_1 character-level LSTM model trained on gold data.

Our *data augmentation experiments* show that using double-sized generated data from AraNews (Train= 97,310 sentences, our setting **c**) is most effective and results in 70.06 F_1. ***This is the best model we report in this paper. It is*** $\sim$ 2.85 F_1 ***higher than our own baseline, and*** 5.76 F_1 ***better than Khouja (2020)'s best model. Overall, our results clearly demonstrate the positive impact of our manipulated data on the fake news detection task, thereby lending value to our novel machine generation method.***

Setting	TRAIN Split	Model	DEV		TEST	
			Acc.	**F1**	**Acc.**	**F1**
Baseline	KH	mBERT	**73.40**	64.74	70.39	61.93
		XLM-R$_{\text{Base}}$	72.74	64.27	72.15	64.92
		XLM-R$_{\text{Large}}$	71.52	**65.60**	72.15	**67.21**
		AraBERT	73.07	67.10	72.59	67.05
Zero-Shot	(a)	mBERT	61.92	48.14	60.96	49.12
		XLM-R$_{\text{Base}}$	61.81	47.42	60.53	47.37
		XLM-R$_{\text{Large}}$	**62.36**	49.52	62.28	50.28
		AraBERT	62.03	47.72	61.62	49.27
	(b)	mBERT	53.09	49.12	53.73	50.70
		XLM-R$_{\text{Base}}$	58.28	47.66	57.89	48.59
		XLM-R$_{\text{Large}}$	58.06	46.99	61.18	**52.71**
		AraBERT	54.42	**49.94**	53.29	50.12
	(c)	mBERT	55.41	48.87	54.61	49.18
		XLM-R$_{\text{Base}}$	55.85	48.21	56.58	48.77
		XLM-R$_{\text{Large}}$	56.62	48.75	57.89	50.33
		AraBERT	54.86	48.65	57.24	51.49
Data Augmentation	KH+(a)	mBERT	71.96	65.51	68.20	60.72
		XLM-R$_{\text{Base}}$	70.86	62.39	69.96	62.71
		XLM-R$_{\text{Large}}$	65.89	61.40	66.67	62.86
		AraBERT	72.63	67.15	70.83	65.38
	KH+(b)	mBERT	70.20	64.68	69.74	64.58
		XLM-R$_{\text{Base}}$	72.52	67.05	72.37	67.40
		XLM-R$_{\text{Large}}$	**73.29**	65.71	72.37	65.79
		AraBERT	72.96	62.94	73.90	66.44
	KH+(c)	mBERT	69.54	64.79	68.42	64.11
		XLM-R$_{\text{Base}}$	69.65	64.65	72.15	66.94
		XLM-R$_{\text{Large}}$	71.85	**67.15**	**74.12**	**70.06**
		AraBERT	70.20	65.38	73.03	69.90

Table 11: Performance results of our the MTD models on the DEV and TEST splits of Khouja. **KH**: refer to Khouja TRAIN split. **(a)** ATB$^+$, **(b)** AraNews$^+$, and **(c)** 2x AraNews$^+$.

7 Conclusion

We presented a novel, simple method for automatic generation of Arabic manipulated text for the news domain. To enable off-the-shelf use with our method, we also collected and released a new POS-tagged Arabic news dataset. Exploiting our dataset, we developed and released the first Arabic model for detecting manipulated news text. We performed a human annotation study shedding light on the impact of our text manipulation approach on news veracity. Finally, we leveraged our generated data for augmenting gold fake news data from an external source and report a new SOTA on the task of fake news detection.

In the future, we plan to explore applying our method to languages other than Arabic. This should be straightforward, since the method itself is language-agnostic and only needs a POS tagger and a dataset from a given language. We also plan to investigate more sophisticated text manipulation methods, exploiting data from different domains. We will also study the impact of these methods on detection of machine generated text as well as fake news detection.

Acknowledgements

MAM gratefully acknowledges support from the Natural Sciences and Engineering Research Council of Canada, the Social Sciences Research Council of Canada, Compute Canada (www.computecanada.ca), and UBC ARC–Sockeye (https://doi.org/10.14288/SOCKEYE).

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium, November. Association for Computational Linguistics.

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

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online, July. Association for Computational Linguistics.

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, Volume 2 (Short Papers)*, pages 21–27.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557.

Amina Chouigui, Oussama Ben Khiroun, and Bilel 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. IEEE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5640–5650.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jana Laura Egelhofer and Sophie Lecheler. 2019. Fake news as a two-dimensional phenomenon: a framework and research agenda. *Annals of the International Communication Association*, 43(2):97–116.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeno, 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. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 301–321. Springer.

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*, pages 1163–1168.

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.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online, July. Association for Computational Linguistics.

Cherilyn Ireton and Julie Posetti. 2018. *Journalism, fake news & disinformation: handbook for journalism education and training*. UNESCO Publishing.

Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2018. Joint multitask learning for community question answering using task-specific embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4196–4207.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online, July. Association for Computational Linguistics.

Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 324–332. ACM.

Raki Lachraf, El Moatez Billah Nagoudi, Youcef Ayachi, Ahmed Abdelali, and Didier Schwab. 2019. ArbEngVec : Arabic-English cross-lingual word embedding model. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 40–48, Florence, Italy, August. Association for Computational Linguistics.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. 27:466–467.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2348–2354, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

El Moatez Billah Nagoudi and Didier Schwab. 2017. Semantic similarity of Arabic sentences with word embeddings. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 18–24, Valencia, Spain, April. Association for Computational Linguistics.

Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–387. Springer.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. Gem: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2009. Arabic gigaword.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

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*.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *ACL (1)*.

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.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Radim Řehřek and Petr Sojka. 2011. Gensim—statistical semantics in python. *statistical semantics; gensim; Python; LDA; SVD*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Victoria L Rubin, Yimin Chen, and Nadia K Conroy. 2015. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2937–2946.

Russell Torres, Natalie Gerhart, and Arash Negahban. 2018. Epistemology in the era of fake news: An exploration of information verification behaviors among social networking site users. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 49(3):78–97.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9054–9065.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking. *arXiv preprint arXiv:1909.03745*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy, July. Association for Computational Linguistics.

Appendices

A AraNews Data

A.1 AraNews: Country, Domain, and Statistics

Country	# Newspaper	Newspaper Name		#News/Newspaper	#News/Country
Morocco	7	الشارع ٢٠	Rue20	36, 556	178, 911
		خبر المغرب	Khabarmaroc	2, 196	
		يا بلادي	Yabiladi	28, 760	
		البيضاوي	Albidaoui	14, 019	
		الأسد	Assdae	18, 600	
		الصباح	Assabah	68, 564	
		الأخبار	Alakhbarpressma	1, 021	
Algeria	6	الشروق	Echoroukonline	187, 936	520, 162
		الخبر	Elkhabar	121, 441	
		الشعب	Ech chaab	147, 960	
		المساء	el-massa	59, 917	
		الجديد اليومي	Eljadidelyawmi	2, 556	
		الامة	Alomah	352	
Tunisia	5	الجريدة	Aljaridah	44, 354	451, 278
		الصريح	Assarih	99, 468	
		المغرب	Lemaghreb	76, 550	
		حقائق اونلاين	Hakaekonline	128, 553	
		الشروق	Alchourouk	102, 353	
Egypt	5	اليوم	Elyom	22, 993	3, 021, 352
		الأهالي	Alahalygate	25, 235	
		طريق الاخبار	Akhbarway	80, 561	
		صوت الامة	Soutalomma	133, 128	
		اليوم ٧	Youm7	2, 759, 435	
Saudi	5	أنحاء	An7a	70, 985	304, 899
		الرياض	Alriyadh	212, 666	
		أم القرى	Uqngovsa	20, 994	
		الحدث	Alhadath	220	
		الجزيرة	Aljazeera	34	
Syria	3	صدى الشام	Sadaalshaamnet	12, 994	47, 058
		الوطن	Alwatansy	104, 68	
		الأيام السورية	Ayyamsyrianet	23, 578	
Sudan	3	السوداني نيوز	Alsudaninews	11, 153	113, 121
		السودان اليوم	Alsudanalyoum	10, 1924	
		ألوان السودانية	Alwandaily	44	
Yemen	3	الشارع نيوز	Alsharaeanews	1, 261	83, 802
		الصمود	Alsomoud	94, 86	
		الثورة	Althawrah	73, 055	
USA	2	بيروت تايمز	Beiruttimes	9, 629	99, 080
		صدى الوطن	Sadaalwatan	11, 091	
		وطن سرب	Watanserb	78, 360	
UK	2	ميدل ايست اونلاين	Middleeastonline	295, 190	295, 566
		بي بي سي	BBC	376	
UAE	2	الأيام	Alayam	5471	63897
		البيان	Elbyan	58426	
Bahrian	1	البحرين	Bahrian	7, 612	7, 612
Iraq	1	الزمان	Azzaman	120, 311	120, 311
Kuwait	1	صحيفة الوسط	Alwasat	31, 354	31, 354
Jordan	1	الدستور	Addustour	689, 444	689, 444
Lebanon	1	أخبار الأرز	Cedarnews	42, 388	42, 388
Palestine	1	عرب ٤٨	Arab48	35, 286	35, 286

Table A1: Descriptive statistics of our ArNews dataset.

A.2 AraNews: Domain Normalization

Sub-Categories		Category	
ثقافة قرأنية. الاسلامي, اسلاميات, الدين والحياة	→	الدين	Religion
التربية و التعليم, تربية, تربية وتعليم, الصباح التربوي	→	تعليم	Education
ثقافية, الثقافة و الفن, فن وثقافة, منوعات و فنون, ثقافة وفنون, الثقافة, ثقافي	→	ثقافة	Culture
علوم وتكنولوجيا, تكنولوجيا, علوم تكنولوجية, علوم وتك, علوم, اخبار التكنولوجيا	→	تكنولوجيا	Technology
مال و اعمال, أخبار الاقتصاد, اقتصاد وسياحة, الاخبار الاقتصادية, اسواق, اقتصاد و بورصة	→	اقتصاد	Economy
سياسة, نقابات, برلمان, الاحزاب, مجلس النواب, قرارات وزارية, مراسيم ملكية, مجلس الوزراء	→	سياسة	Politics
رياضة, رياضة محلية, رياضة وطنية, رياضة دولية, أخبار الرياضة, رياضة عالمية, مواقف رياضية	→	رياضة	Sport
صحة, أخبار الصحة والطب, صحة وطب, فايروس كورونا, الصحة, طبّ و صحّة, العلم والصحة	→	صحة	Health

Table A2: Story sub-categories and main categories to which we map in AraNews.

A.3 ATB$^+$ and AraNews$^+$ Data Splits

Data	# Split	Human	Machine Manipulated					
		# Sent.	ADJ	ADJ_COMP	ADJ_NUM	N_NUM	N_PROP	NEG_PART
ATB$^+$	TRAIN	$48.7K$	$99.5K$	$4.5K$	$5.8K$	$60.6K$	$75.8K$	$43.6K$
	DEV	$6.6K$	$13.4K$	638	844	$7.1K$	$10.6K$	$5.6K$
	TEST	$5.9K$	$11.9K$	592	665	$8.1K$	$9.5K$	$5K$
AraNews$^+$	TRAIN	$3.27M$	$6.2M$	$251.4K$	$298.5K$	$1.4M$	$2.3M$	$387.6K$
	DEV	$5.51K$	$7.8M$	$290.6K$	$293.7K$	$1.4M$	$3.6M$	$704.7K$
	TEST	$6.16K$	$64M$	$343.6K$	$303.4K$	$1.3M$	$5.8M$	$496.9K$

Table A3: Data splits and distribution of POS tags in our machine manipulated datasets : ATB$^+$ and AraNews$^+$

.