Article:

Althabiti, S. orcid.org/0000-0002-4646-0577, Alsalka, M.A. and Atwell, E. (2022) Detecting Arabic Fake News on Social Media using Sarcasm and Hate Speech in Comments. International Journal on Islamic Applications in Computer Science And Technology - IJASAT, 10 (4). pp. 28-36. ISSN 2289-4012

Detecting Arabic Fake News on Social Media using Sarcasm and Hate Speech in Comments

Saud Althabiti[1,2,a], Mohammad Ammar Alsalka[1,b], Eric Atwell[1,c]

[1]School of Computing, University of Leeds, Leeds, United Kingdom
[2]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[a]scssal@leeds.ac.uk, [b]m.a.alsalka@leeds.ac.uk, [c]e.s.atwell@leeds.ac.uk

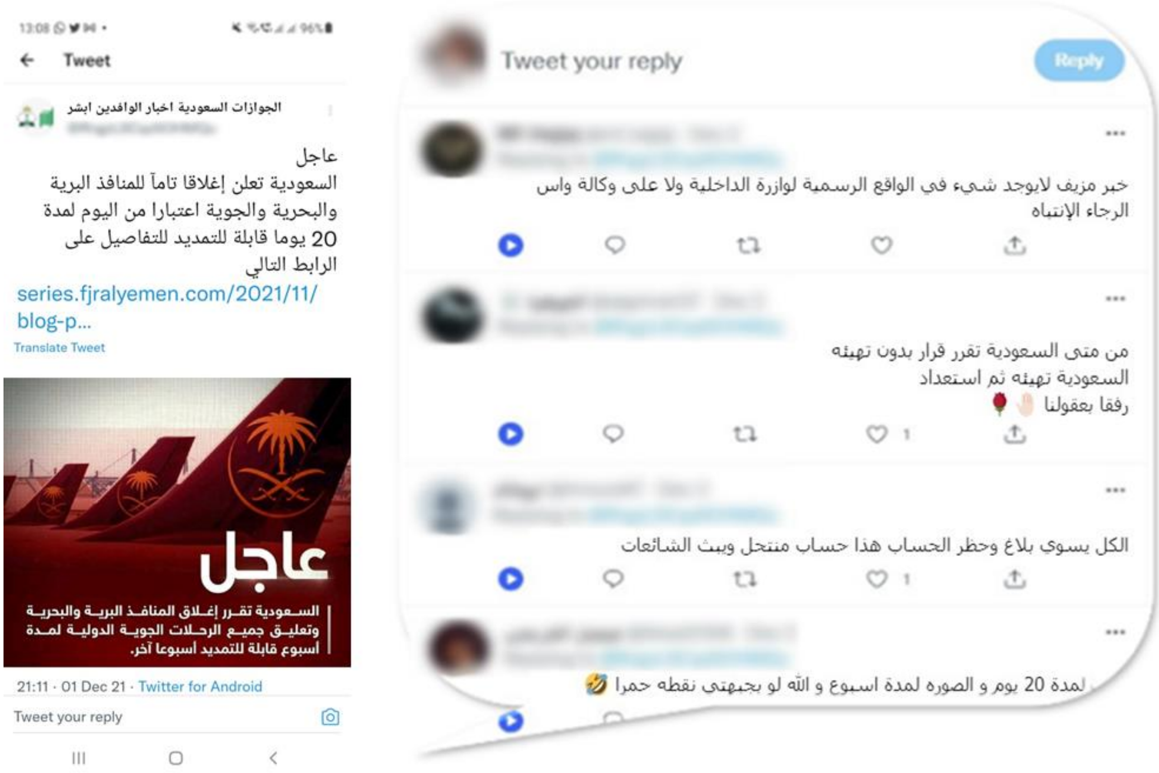
ABSTRACT

The rapidly increasing popularity of social networking sites and the widespread acceptance of anonymous users have encouraged an environment where unidentified accounts can act maliciously and propagate fake news. The motivation behind that could either be to begin hype or to gain individuals' attention and negatively impact society. Several studies attempt to establish models to detect fake news based on news content, source, or propagation path. However, fewer studies have investigated more profound signs, such as people's responses to the information posted on social media. We hypothesize that the existence of sarcasm or hate language in the comments and responses to a news post may be used as an indicator of the authenticity of the post itself. Therefore, this paper proposes a new technique incorporating hate language and sarcasm detected in users' comments as significant features for identifying fake news. We used three Arabic datasets to conduct this study and experimented with various state-of-the-art models. As a result, we conclude that considering these features in news responses can help detect fake news since we found that the existence of sarcasm or hate speech in comments of false tweets is approximately double that in true ones.

***Keywords*: fake news detection, comments, hate speech, sarcasm, machine learning, Arabic NLP, social media**

1. Introduction

Online social media and microblogs have evolved into popular means for obtaining news. With this significant advancement, the general public tend to gain information through these platforms as they facilitate reaching and spreading the desired story (Shu et al., 2017). Moreover, social media encourages users to perform additional actions, such as commenting, sharing, resharing, and liking, which propagates and expands the posts' discussions. Unlike traditional news, any person can use microblogs by creating a personal account and posting or sharing messages on any matter, even if inaccurate, which could raise detrimental societal consequences. This paper presents a vital research question: Can hate speech and sarcasm in response to fake news be used as an indicator that the news is fake, and increase fake news detection model performance? A possible method we used is investigating possibly related tasks, such as sarcasm and hate speech expressed in tweets' comments. The idea began when deeply interpreting and analyzing an actual fake news example. Figure 1 exhibits an Arabic post on Twitter that carries misinformation, saying, "Urgent: Saudi Arabia announces a complete closure of land, sea, and airports from today for 20 days, which can be extended. For more details, check the following link". The style of the written tweet as well as the source seems to be a trustworthy and credible post. However, several responses to this post agreed that it was fake, which explains the efficacy of employing the thread's replies in addition to the post content.



الجوازات السعودية اخبار الوافدين ابشر

عاجل
السعودية تعلن إغلاقا تاماً للمنافذ البرية
والبحرية والجوية اعتبارا من اليوم لمدة
20 يوما قابلة للتمديد للتفاصيل على
الرابط التالي
series.fjralyemen.com/2021/11/
blog-p...

السعودية تقرر إغلاق المنافذ البرية والبحرية
وتعليق جميع الرحلات الجوية الدولية لمدة
أسبوع قابلة للتمديد أسبوعا آخر.

21:11 · 01 Dec 21 · Twitter for Android

خبر مزيف لايوجد شيء في الواقع الرسمية لوزارة الداخلية ولا على وكالة واس
الرجاء الإنتباه

من متى السعودية تقرر قرار بدون تهيئه
السعودية تهيئه ثم استعداد
رفقا بعقولنا ✋🌹

الكل يسوي بلاغ وحظر الحساب هذا حساب منتحل وبيث الشائعات

لمدة 20 يوم و الصوره لمدة اسبوع و الله لو بجبهتي نقطه حمرا 🤣

Figure 1: Example of an Arabic Fake News

Despite the agreement on disbelieving the news post, each user expressed their opinions differently. For example, some showed their feelings about the subject, some commented sarcastically, and some even included hateful words when responding to the news. Therefore, developing a different model that adopts commentators' intentions as new features has the potential to improve fake news detection performance.

The subsequent section includes a literature review of fake news detection issues and presents four main approaches. Section three describes the three datasets we used to conduct this study. The proposed methodology, models used and experiments are detailed in the fourth section. Then, we present and discuss our findings in the results and discussion section. Finally, we conclude this paper and suggest future work.

2. Related work

Various automatic classification approaches were used in different studies to tackle the difficult challenge of fake news detection. A review (Zhou & Zafarani, 2020) examined and assessed techniques from four different perspectives for identifying fake news: the inaccurate information it contains (knowledge-based), the writing style (style-based), the propagation patterns (propagation-based), and the source's credibility (source-based). The knowledge-based method evaluates news' authenticity by comparing the extracted knowledge with confirmed or true facts. The second method attempts to catch misinformation by capturing the writing style of the published misleading information since those manipulators usually have malicious intent to spread distorted posts for the purpose of influencing vast communities of users (Shu et al., 2017). Another vital factor for detecting misinformation is investigating how posts are diffused throughout a network. Several studies analyzed how users pass the news to each other, analogous to the study of Isnad or chain of narrators in verifying the credibility of Hadith (Tarmom et al., 2020; 2021). These prevalent posts' paths form propagation structures that include information such as the cascade's depth and breadth. These incorporated features can then be used as input in classification models to classify news as fake or real. In addition, a source-based method is heavily used, which relies on the primary source (author or publisher) or the latest source (the user who shared the information).

Because fake news is prepared to portray strong attitudes towards a certain subject (Alonso et al., 2021), various studies have utilized sentiment analysis to help tackle the challenge of detecting fake news. Some approaches consider sentiment analysis as the essential basis for detecting fake news. At the same time, other studies used sentiment analysis as one feature to be combined with other features acquired from the posts. Studies (AlRubaian et al., 2015; Bhutani et al., 2019; Diakopoulos et al., 2010) analyzed sentiments of the news text posted to social networking sites such as Twitter since plenty of fake news shows exaggeratedly negative or positive feelings. While other studies have proven that most real news spreads naturally without including emotional sentences to mislead readers (Alonso et al., 2021). The original news could be real in some cases, but manipulators could convey it differently. Therefore, some researchers, such as (Ross & Thirunarayan, 2016) investigated how similar news polarity is to the expressed description following the tweets so that if there is a similarity, this may indicate the credibility of the news and vice versa. Finally, fewer studies have exploited users' responses to the given information to evaluate the total sum of the readers' feelings towards the news (Zhang et al., 2021).

However, what about more subtle indications that users' responses contain, such as the presence of ironic or offensive phrases in each comment? To the best of our knowledge, no previous study has addressed these kinds of features to identify fake news. Since responses to news can support determining credibility, we could take advantage of users' comments to extract new features to detect fake news along with the text of the news.

3. Datasets

To assess the method proposed in this paper and conduct experiments using state-of-the-art (SOTA) models, we used several Arabic social media datasets.

3.1 ArSarcasm

One of the main tasks is sarcasm detection, so our first step was acquiring a labelled dataset that would help identify sarcastic tweets. A study by (Farha & Magdy, 2020) presented ArSarcasm, an Arabic sarcasm detection dataset. The dataset has more than ten thousand tweets, and it is not only labelled for detecting sarcasm purposes but also annotated for sentiment and dialects. It contains 1682 sarcastic tweets. Hence, in this case study, we only utilized the text of tweets and the sarcasm label column.

3.2 ARACOVID19-MFH

The other task we considered is hate-speech detection. ARACOVID19-MFH (Hadj Ameur & Aliane, 2021) is also a multi-label dataset that involves ten features and labels manually annotated and contains approximately 11000 Arabic tweets. Due to regulations imposed by Twitter, the texts of these tweets are not publicly available. Alternatively, they provided tweet IDs to encourage future researchers to retrieve the text of these posts and implement further investigations. In addition, the dataset has ten different labels, including "contain hate?" and "contain fake information?" In this experiment, we employed the text of tweets and the 'contain hate' label column.

3.3 ArCOV19-Rumors

The third dataset we considered for this experiment was one with data labelled for fake news detection, with comments. This helps us study users' engagements and extract new features. Several Arabic datasets related to the COVID-19 pandemic are available for researchers, including reliable and unreliable information, mostly collected from Twitter. For example, one of the recent studies introduced a dataset called ArCOV19-Rumor (Haouari et al., 2020). This dataset, which holds about 3.6k manually annotated tweets as either true or false, aims to detect Arabic rumours on Twitter. It was derived from about 138 claims verified from popular fact-checking websites. The dataset comprises the following categories: health, sports, social, politics, religions, and entertainment, and they utilise tweets' content, account profiles, propagation networks, and users' comments on the tweet-level verification. In this experiment, we employed the text of tweets and comments within each tweet thread.

4. Methodology

In this study, we conducted several experiments, including dataset collection, model selection, feature extraction and automatic labelling, and detecting fake news using the extracted features. These experiments are explained in the following subsections.

4.1 Collection

Firstly, we used Twitter API to collect published annotated datasets from previous studies for the three different tasks explained in section 3. We will refer to the three datasets ArSarcasm, ARACOVID19-MFH, and ArCOV19-Rumors as DS-T1, DS-T2, and DS-FND (tweets and replies) respectively, as shown in the model architecture in Figure 2. In the DS-FND dataset, the tweet file contains 1069 out of 3,584 tweets that has responses. While the replies file has 38514 tweet ids. However, this number is further reduced due to suspension of some accounts, deleted comments, and other reasons. In addition, we removed threads that contained self-replies that make long news threads. The result was 18,521 replies that we used in our study.

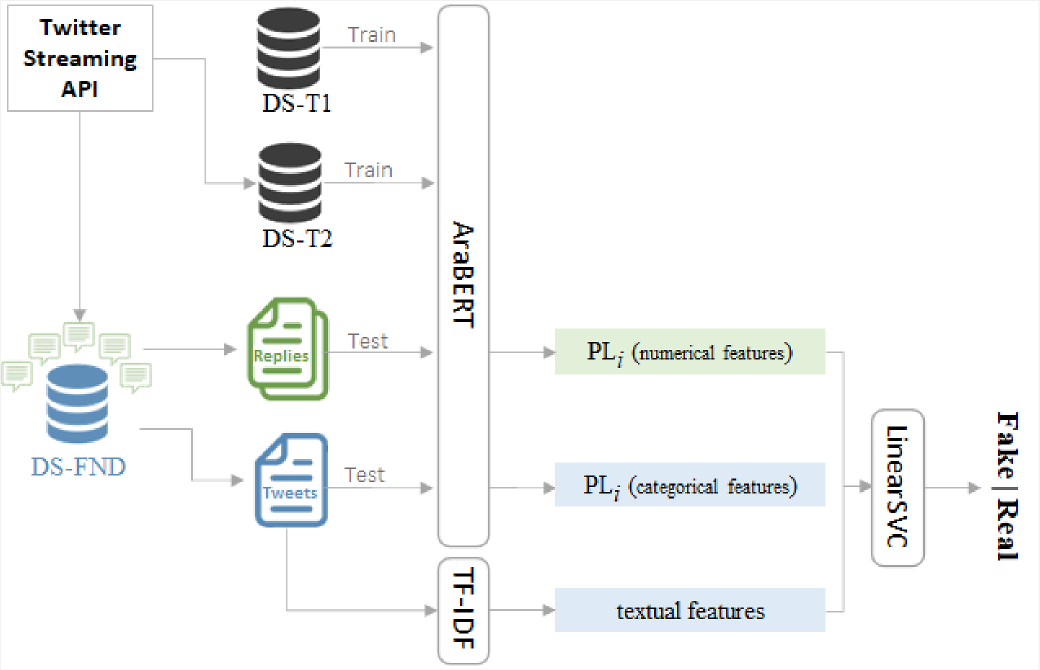


Figure 2: Model Architecture

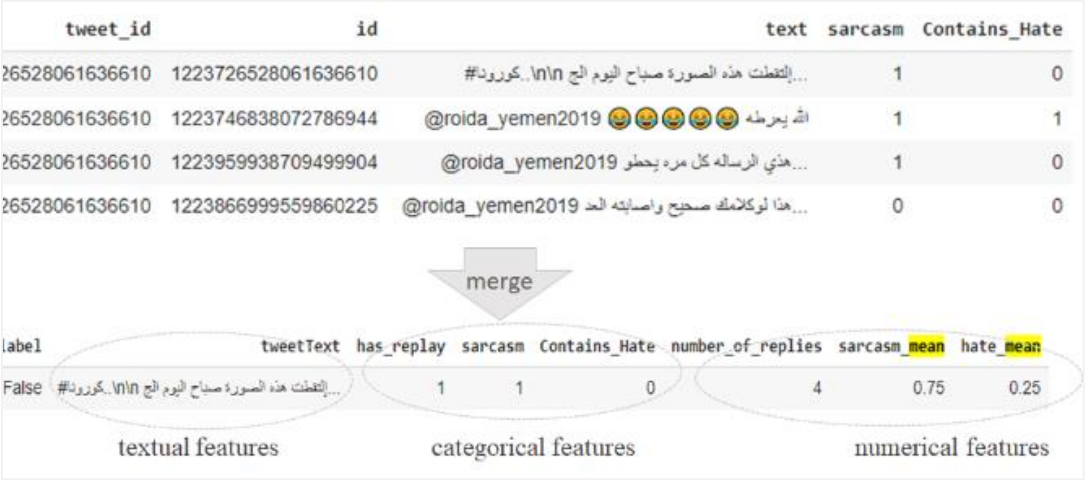
4.2 Model selection

After collection, we split each task's dataset (DS-T$_i$) into training and testing. Next, we trained various state-of-the-art models using the datasets and made some comparisons. Firstly, we started with supervised algorithms such as Support Vector Classifiers (SVC). Although many machine learning algorithms are used in classification problems, SVC usually has better accuracy than other machine learning classification algorithms (Althabiti et al., 2021; Shaar et al., 2021). In this model, we used TF-IDF (Term Frequency - Inverse Document Frequency) as the metric, to build a TFIDF_SVC model. Moreover, we wanted to use transformer-based models for natural language processing (NLP), such as BERT (Devlin et al., 2018) and Roberta (Liu et al., 2019). Many studies have proven these bidirectional models to be powerful methods in NLP to comprehend context-heavy texts (Althabiti et al., 2022). Therefore, we also used AraBERT (Antoun et al., 2020), an Arabic pre-trained language model based on BERT architecture, because we used Arabic datasets in our experiments. Then, we evaluated the results using (precision, recall, and F1 score) to determine the best-performing model (M$_i$) for each task (T$_i$).

4.3 Extracting new features

We trained the selected model (M$_i$) for each task by adopting the entire dataset, DS-T$_i$. Then tested on both tweets and replies from DS-FND. The predicted labels PL_{ij} from the tweet and reply file are categorical; for example (does a tweet i or a comment j contain sarcasm?). Accordingly, the average of sarcasm or hate speech $\bar{x}_i$ in each tweet's comments ij has been added as a new numerical feature. We used equation 1 to calculate these features. The DS-FND file was then augmented with automatic labels derived from these newly extracted features, as shown in Figure 3. We also published these automatically labelled dataset files, which can be found on GitHub[1].

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} PL_{ij} \qquad (1)$$



tweet_id	id	text	sarcasm	Contains_Hate
26528061636610	1223726528061636610	...إلتقطت هذه الصورة صباح اليوم الح \n\n..كورونا#	1	0
26528061636610	1223746838072786944	@roida_yemen2019 😂😂😂😂😂 الله يرحطه	1	1
26528061636610	1223959938709499904	...هذي الرساله كل مره يحطو @roida_yemen2019	1	0
26528061636610	1223866999559860225	...هذا لوكلامك صحيح واصابته الحد @roida_yemen2019	0	0

label	tweetText	has_replay	sarcasm	Contains_Hate	number_of_replies	sarcasm_mean	hate_mean
False	...إلتقطت هذه الصورة صباح اليوم الح \n\n..كورونا#	1	1	0	4	0.75	0.25

Figure 3: Merging extracted features from comments into the tweets file

[1] https://github.com/althabiti/Sarcasm-and-Hate-in-comments-for-fake-News-Detection

4.4 Detecting fake news

The final step was to consider the following features:

- Textual features: the text of a tweet.
- Two categorical features extracted from the tweets file (namely, is it a sarcastic or an offensive tweet?)
- Two numerical features extracted from the replies file (namely, the percentage of sarcastic comments and the percentage of offensive comments).

We first applied the model selection step to the textual features and then used it for our final experiment, which considered all extracted features along with the text.

4.5 Implementation tools

Along with Twitter API, we used the "Hydrator[2]", an Electron-based desktop application for crawling Tweets' IDs to harvest datasets. Experimentations with transformer-based models were conducted on Google's Colab GPU environment. This platform was chosen as it is free to use and does not require installing, setting up, and configuring many Python packages and libraries. We also used local machines to implement other experiments, analyses, and visualization.

5. Results and discussion

We conducted five main experiments in this study. The first two aimed at selecting a model for both sarcasm detection and hate speech detection. Our results, demonstrates that the newly developed AraBERTv0.2-Twitter outperforms other state-of-the-art models in both experiments, as shown in

Table 1. In order to rely on this model, we also manually evaluated it. We tried several samples by writing new comments and making the model predict if our input includes sarcasm or hate speech, so we used it to label both the tweet file and the replies file automatically. As a result, the model predicted 48 and 15 sarcastic and offensive tweets, respectively. Also, the model predicted 1042 out of 18521 sarcastic replies and 592 responses that included hate speech.

Table 1: Comparing SOTA models for model selection

Model	ARACOVID19-MFH			ArSarcasm		
	P	R	F1	P	R	F1
TFIDF_SVC	0.99	0.97	0.98	0.82	0.85	0.82
arabertv02-twitter	**0.99**	**0.98**	**0.98**	**0.93**	**0.93**	**0.93**
Roberta	0.80	0.90	0.85	0.84	0.84	0.84
BERT	0.96	0.96	0.96	0.71	0.84	0.77

The bar chart in Figure 4 gives the number of tweets containing one or more of three features in eight combinations. The three features of the combinations on the x-axis are the tweet label (true or false), the existence of sarcastic users' engagements (0: does not existent, 1: exists) and hateful language in users' engagements (0: does not existent, 1: exists). After merging the newly extracted categorical features and the numerical features explained in subsection 4.4- with the tweets file, we found that most of the tweets' replies do not include sarcasm or hate language, as shown in the first two bars in Figure 4. Looking at the other six situations when sarcasm or hate

[2] https://github.com/DocNow/hydrator

language is present, the number of false tweets is about double that of true tweets. This shows that the existence of sarcasm or hate language in comments indicates fake news in the original post and supports our initial hypotheses in this paper.

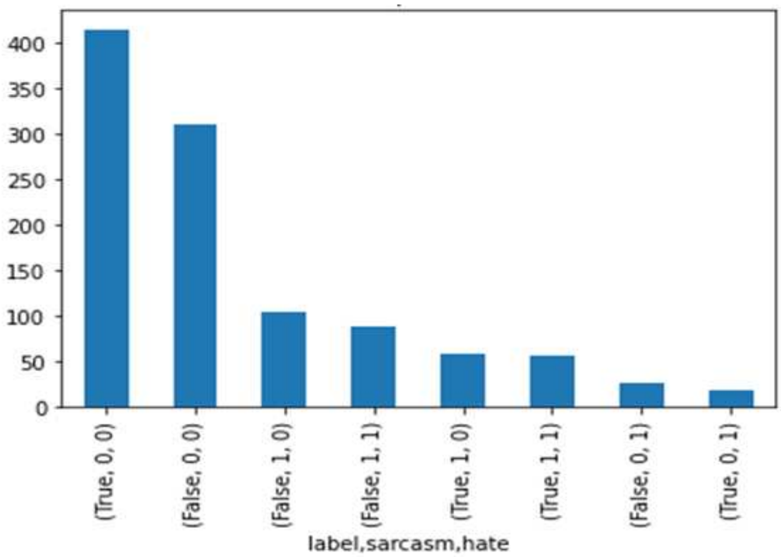


Figure 4: Number of sarcastic and hateful comments in both true and false news.

After that, we repeated the model selection step on ArCOV19-Rumors dataset to select the best-performing model for fake news detection from the tweets' text. The AraBERT model in this case study only got 83% accuracy compared with the SVC, which outperformed other models with a 90.4%. Therefore, we employed this model for the final experiment. In the final step, we considered the newly extracted features along with textual features. Our findings reveal that the model could predict tweet authenticity with an accuracy of 0.895 considering all features, compared to 0.904 using only textual features. Surprisingly, the analysis showed that sarcasm and hate speech in comments might slightly decrease the performance of predicting fake news in tweets in this case.

There are some possibilities that might justify this reduction. Firstly, because of the significant absence of sarcastic and hateful replies within the entire dataset, as illustrated in Figure 4, the model may not consider these extracted features a good indication for detecting fake news. Additionally, there is still a possibility that the selected model could mis predict due to the inherently ambiguous nature of sarcasm, and the same applies to hate speech. Therefore, tackling this challenge is a crucial stage in detecting fake news from comments.

6. Conclusions and future work

Since the advent of social media, many people have preferred to read news from such sites. Unfortunately, due to its general acceptance, this has increased the dilemma of fake news. Although diverse researchers have investigated different methods to tackle this issue, fewer studies have explored other indications, such as examining commenters' engagement in a discussion. Hence, we strive to incorporate new features, including offensive or sarcastic reactions towards a particular post, to detect fake news. To conduct this study, we trained state-of-the-art models using two published datasets for sarcasm and hate detection. Then, we tested the trained models on a third dataset labelled as fake or real, including tweets and user comments. Consequently, our analysis illustrates the possibility of using these extracted features as good indicators for identifying false news. Nevertheless, our findings using the training models on new extracted features along with the text slightly reduced the performance compared to using only the textual features. In future work, we endeavour to investigate further indications from comments, such as sentiment analysis. We published the automatically-labelled dataset to enable future researchers to investigate further. In addition, the development of explainable models is becoming more significant since most of the widely used SOTA models behave similarly to black boxes that provide adequate results but can be challenging to justify.

Acknowledgements

We would like to acknowledge and thank the Ministry of Education in Saudi Arabia, King Abdulaziz University, and the University of Leeds for their continuous support.

References

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. In *Electronics (Switzerland).* https://doi.org/10.3390/electronics10111348

AlRubaian, M., Al-Qurishi, M., Al-Rakhami, M., Rahman, S. M. M., & Alamri, A. (2015). A multistage credibility analysis model for microblogs. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1434–1440.

Althabiti, S., Alsalka, M. A., & Atwell, E. (2022, August). SCUoL at CheckThat! 2022: fake news detection using transformer-based models. In *CEUR Workshop Proceedings* (Vol. 3180, pp. 428-433). CEUR Workshop Proceedings.

Althabiti, S., Alsalka, M., & Atwell, E. (2021). SCUoL at CheckThat! 2021: An AraBERT model for check-worthiness of Arabic tweets. *CEUR Workshop Proceedings*.

Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *ArXiv Preprint ArXiv:2003.00104*.

Bhutani, B., Rastogi, N., Sehgal, P., & Purwar, A. (2019). Fake News Detection Using Sentiment Analysis. *2019 12th International Conference on Contemporary Computing, IC3 2019*. https://doi.org/10.1109/IC3.2019.8844880

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Diakopoulos, N., Naaman, M., & Kivran-Swaine, F. (2010). Diamonds in the rough: Social media visual analytics for journalistic inquiry. *2010 IEEE Symposium on Visual Analytics Science and Technology*, 115–122.

Farha, I. A., & Magdy, W. (2020). From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 32–39.

Hadj Ameur, M. S., & Aliane, H. (2021). AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset. *Procedia CIRP*. https://doi.org/10.1016/j.procs.2021.05.086

Haouari, F., Hasanain, M., Suwaileh, R., & Elsayed, T. (2020). ArCOV19-rumors: Arabic COVID-19 twitter dataset for misinformation detection. *ArXiv Preprint ArXiv:2010.08768*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv Preprint ArXiv:1907.11692*.

Ross, J., & Thirunarayan, K. (2016). Features for ranking tweets based on credibility and newsworthiness. *2016 International Conference on Collaboration Technologies and Systems (CTS)*, 18–25.

Shaar, S., Hasanain, M., Hamdan, B., Ali, Z. S., Haouari, F., Nikolov, M. K. A., Kartal, F. A. Y. S., da San Martino, G., Barrón-Cedeño, A., & Míguez, R. (2021). Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates. *Working Notes of CLEF*.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Tarmom T, Atwell E, Alsalka M. (2021). Deep Learning vs Compression-Based vs Traditional Machine Learning Classifiers to Detect Hadith Authenticity. *8th International Conference on Information Management and Big Data (SIMBig 2021)*

Tarmom T, Atwell E, Alsalka MA. (2020). Non-authentic Hadith Corpus: Design and Methodology. *International Journal on Islamic Applications in Computer Science And Technology.* 8(3), pp. 13-19

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. *Proceedings of the Web Conference 2021*, 3465–3476.

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*. https://doi.org/10.1145/3395046

Abstract in Arabic

الشعبية المتزايدة بشكل سريع في مواقع الشبكات الاجتماعية والقبول واسع النطاق شجعت للمستخدمين المجهولين بيئة بحيث يمكن للحسابات الغير معروفة أن تتصرف بشكل سلبي وتنشر أخبارًا مزيفة. قد يكون الدافع وراء ذلك إما البدء في الضجيج أو جذب إنتباه الأفراد والتأثير سلبًا على المجتمعات. تحاول العديد من الدراسات إنشاء نماذج وخوارزميات للكشف عن الأخبار المزيفة بناءً على محتوى الأخبار أو المصدر أو مسار الانتشار. ومع ذلك، يوجود فقط القليل من الدراسات التي بحثت في علامات أكثر عمقًا، مثل ردود فعل الناس على المعلومات المنشورة التي تنقل أخبارعلى وسائل التواصل الاجتماعي. في هذا الدراسة نحن نفترض أن وجود السخرية أو لغة الكراهية في التعليقات والردود على منشور إخباري يمكن استخدامه كمؤشر على صحة المنشور نفسه .لذلك ، تقترح هذه الورقة تقنية جديدة تتضمن لغة الكراهية والسخرية المكتشفة في تعليقات المستخدمين كعلامات مهمة لتحديد الأخبار المزيفة .استخدمنا ثلاث مجموعات بيانات عربية لإجراء هذه الدراسة وجربنا عدة من نماذج الذكاء الاصطناعي الحديثة .ونتيجة لذلك ، نستنتج أن النظر في هذه العلامات في الردود الإخبارية يمكن أن يساعد في الكشف عن الأخبار المزيفة لأننا وجدنا أن وجود السخرية أو الكلام الذي يحض على الكراهية في التعليقات على التغريدات الكاذبة يقارب ضعف نظيره في التعليقات الصحيحة.