EMET: EMBEDDINGS FROM MULTILINGUAL-ENCODER TRANSFORMER FOR FAKE NEWS DETECTION

Stephane Schwarz[⋆] *Antônio Theóphilo*[⋆†] *Anderson Rocha*[⋆]

stephane.schwarz@outlook.com, {antonio.theophilo, anderson.rocha}@ic.unicamp.br
[⋆] Institute of Computing – University of Campinas, Campinas/SP, Brazil
[†] CTI Renato Archer, Campinas/SP, Brazil

ABSTRACT

In the last few years, social media networks have changed human life experience and behavior as it has broken down communication barriers, allowing ordinary people to actively produce multimedia content on a massive scale. On this wise, the information dissemination in social media platforms becomes increasingly common. However, misinformation is propagated with the same facility and velocity as real news, though it can result in irreversible damage to an individual or society at large. Solving this problem is not a trivial task, considering the reduced size of the text messages usually posted on these communication vehicles. This paper proposes an end-to-end framework called EMET to classify the reliability of small messages posted on social media platforms. Our method leverages text-embeddings from multilingual-encoder transformers that take into consideration the semantic knowledge from preceding trustworthy news and the use of the reader's reactions to detect misleading content. Our findings demonstrated the value of user interaction and prior information to check social media post's credibility.

Index Terms— Fake News Detection, Information Forensics, Deep Learning, Natural Language Processing, Social Media Data.

1. INTRODUCTION

The social media landscape is changing the way people interact with the world. The sudden growth of these platforms reflects positively in several scenarios, such as business strategy, politics, and communication. Their popularization empowers people engagement so that they can share their opinions with thousands of other people around the globe almost instantaneously. Given that these tools allow users to spread any information into a broad public environment, it becomes a great instrument to amplify and propagate news. According to the Pew Research Center [1], in 2019, seven out of ten Americans engage in news content offered by social platforms.

Although social media may be helping to expand horizons, several negative impacts stem from this advent. As reported by the Oxford Internet Institute [2], in the last three years, governments and political parties performed suspicious actions in social media networks. These actions had the purpose of manipulating and shaping public opinion, attacking political opponents, and dividing societies online by polarizing news and text messages in 70 distinct countries.

After the 2016 United States presidential election, the debates around the term Fake News exploded. As a natural outcome, several researchers started to work over this theme to better understand this phenomenon and how to mitigate falsehood spread.

It is in this scenario that this work presents a deep-learning-based end-to-end framework called EMET for the identification of fake posts on social platforms. Our approach contributes with a novel approach in which the readers' reactions and trustworthy news articles related to the target post are also taken into account in the classification process. Results pose our method as a new state-of-the-art for the problem and show that online comments from the crowd are a valuable asset when investigating news pieces.

We organized the remaining of this paper into five sections. Sec. 2 briefly discusses the related works in the area of fake news detection and deep learning applied to textual tasks. Sec. 3 details our proposed method to solve the problem, while Sec. 4 unveils the performed experiments. Ultimately, Sec. 5 presents some conclusions and final remarks.

2. RELATED WORK

Despite having different definitions, according to Shu et al. [3], fake news is “news articles that are intentionally and verifiably false”. Two important aspects to highlight in this definition are: (1) information needs to be verified as false; and (2) it must have been produced with malicious intent. A simpler but similar definition can be found at [4] in which fake news is defined as “a made-up story with an intention to deceive”.

Castillo et al. [5] presented a seminal work in which features relevant to tasks like fake news detection over microblogs are classified into four categories: message-, topic-, user- and propagation-based. Despite being Twitter®-specific, the authors claim to be generic enough for modern online platforms.

According to Lillie and Middelboe [6], stance detection is a technique used to infer if some news post is true or false. The authors also discuss the difficulty in comparing different fake news detection methods due to the differences in the data and metrics used.

The Fake News Challenge [7] is an initiative to foster machine learning and natural language processing (NLP) techniques to automate the steps required in the identification if a story is real or not. The authors believe that stance detection is an important first step in this path, and defined this task as the first one to be dealt with by the Challenge. Recently, Theóphilo et al. [8] also approached the problem of fighting fake news through smaller tasks but, instead of stance detection, they focused on the authorship of small messages.

Derczynski et al. [9] described the PHEME Project that proposes to detect and classify fake news into the four categories: rumours,

This work was funded by the São Paulo Research Foundation (FAPESP), DéjàVu grant #2017/12646-3 and grant #2018/10204-6; by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), DeepEyes grant; by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant #304497/2018-5, by Google Latin America Research Awards (LARA 2019) and by Movile/Wavy.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: MKSSS CUMMINS COLLEGE OF ENGINEERING FOR WOMEN. Downloaded on November 21,2020 at 07:31:08 UTC from IEEE Xplore. Restrictions apply.

disinformation, misinformation, and speculation.

Wang et al. [10] presented an approach for fake news detection based on learning shared features present in many events instead of focusing on event-specific features in order to have good performance on posts of unseen events. For filtering out these specific features, the authors devised the clever strategy of applying Generative Adversarial Networks (GANs) techniques to learn features able to deceive an event discriminator. They claimed to have good results over two datasets based on Twitter® and Weibo® (a Chinese microblog platform). They presented two models: (1) a Convolutional Neural Network (CNN) to extract textual features from each post, encoded in a 32-dimension pre-trained word-embedding to predict whether this post is fake or not (EANN Text); and (2) three components to extract, discriminate and, detect fake content based on textual and visual information (EANN Multi).

Khattar et al. [11] proposed an end-to-end network aiming to learn a shared multimodal representation of data (textual + visual) to perform fake news detection. The authors also proposed two models, one for text and another for multimodal data. The first (MVAE Text) feeds the text into a Bi-LSTM network to represent each word sentence in a 32-dimension vector. The second (MVAE Multi) uses a unified representation from visual and textual modalities, encoding both information into a latent vector to learn a shared representation to predict news credibility. They also performed their experiments in the same datasets of Wang et al. [10].

Working over the same datasets of Wang et al. [10] and Khattar et al. [11], Qi et al. [12] proposed a model called Multi-domain Visual Neural Network (MVNN) to identify fake news based solely on images, using only visual features from frequency and pixel domains in order to classify the posts. They designed a CNN network to capture patterns of fake-news images in the frequency domain, and a CNN-RNN model to extract visual features from different semantic levels in the pixel domain. An attention mechanism was utilized to combine the frequency and pixel domain feature representations.

Zubiaga et al. [13] defined an annotation methodology for tweets that can also be extended for other kinds of social interaction platforms. According to the authors, there are two types of tweets (source and response) and they can be categorized into three dimensions: support/response, certainty and evidentiality. This annotation methodology suggests that nested posts/replies play an important role in rumours propagation. Tolmie et al. [14] compare the microblogging with sociological aspects of conversations and highlights the importance of analyzing threads of discussions instead of individuals tweets due to the characteristics of these platforms.

3. METHODOLOGY

Previous work has shown that misleading posts can be detected with good results through syntax or author's writing style analyses, especially if sentiment words are taken into account, as negative emotions can indicate guilt [8, 15]. Despite the power of these methods, the fake text generators are continuously sophisticating their strategies to induce readers as much as possible [16]. In this case, we need to explore more reliable and robust data, such as semantics, fact-checking information, and the reader's response in order to avoid losing important information.

The problem statement can be defined as: Let T_i^E be a social media post related to a particular event E, consisting of a set of keywords $K(T_i^E) = \{k_1, k_2, ..., k_n\}$. Let $C(T_i^E) = \{c_1, c_2, ..., c_n\}$ be a set of user comments, corresponding to a post T_i^E, and $N^E = \{n_1, n_2, ..., n_n\}$ a set of reliable news related to the event E retrieved using $K(T_i^E)$. Different to prior research, we aim at learning a model $\mathcal{F}$ that can label a single post in one of following three categories: false ($y = 0$), true ($y = 1$), or unknown ($y = 2$), taking into account not only the post itself but also the comments and the reliable news piece. Our approach treats the input sample as a unidimensional signal as we did in our ICASSP 2019 work [8]). The final feature vector is formed by the concatenation of this data:

$$\mathcal{F} : (T_i^E \cdot C(T_i^E) \cdot N^E) \implies y. \tag{1}$$

As we will show in the next section, one key challenge of this model is how to use the reliable news piece as side information to correctly classify the target post.

3.1. Input Signal Transformation

To deal with the three different text inputs, we first represent each text piece as a unidimensional signal (sequence of characters). We then project each transformed signal onto a latent semantic space to obtain a more representative characterization of its content and semantics before using it in the actual classification model. For the semantic projection, we leverage the multilingual model developed by Yang et al. [17] based on Transfomers [18] learned in a multitask training scenario. This model presented superior or competitive performance in several NLP tasks as semantic retrieval, translation pair bitext retrieval, and retrieval question answering.

After this projection, each text signal (which naturally has different lengths) is transformed into a 512-dimension semantic vector. The purpose of this projection is to come up with semantic representations from all the textual input (post, comments, and news piece) that can be further used by the classification model to highlight inconsistencies presented in the target publication.

3.2. Network Architecture

Misleading posts are usually associated with pieces of checked information in order to get shared more often. We hypothesize that there exists a substantial correlation between the post message, their comments, and past trustworthy news linked with a specific event that highlights whether a post is one of the classes mentioned earlier.

Our model is based on a convolutional neural network, and its topology is detailed in Fig. 1. The input consists of three textual documents fed in parallel: a single social media post, all the associated comments concatenated, and a piece of checked news related to the event reported on the post. We project these documents individually onto the adopted multilingual transformer space (named 'Encoder' in Fig. 1), and afterwards we apply five one-dimension (1D) convolutional filters over the projections to standardize the processing operations. The results of all these convolutions are flattened and concatenated to assemble the input of the final convolutional layers in order to capture short and long-range semantic similarities.

After the concatenation layer, there is a sequence of 1D convolution, max-pooling, and fully-connected layers. The rationale behind this structure is to capture intrinsic syntactic and semantic patterns among the three input sentences that are far apart from each other. In the following, to enrich the feature vector representation, there are two sets of 1D convolutions and fully-connected layers before the final classification layer, a softmax, which gives the output probability distribution over the three classes. To complete the network topology description, there is a dropout (p = 0.5) between layers Conv3 and FC2, and a ReLU activation layer after the FC4 layer.

The convolution filters are connected only to a local region of the input, while the fully-connected layer neurons are completely connected with the three adjacent layers. The objective of this model-

Authorized licensed use limited to: MKSSS CUMMINS COLLEGE OF ENGINEERING FOR WOMEN. Downloaded on November 21,2020 at 07:31:08 UTC from IEEE Xplore. Restrictions apply.

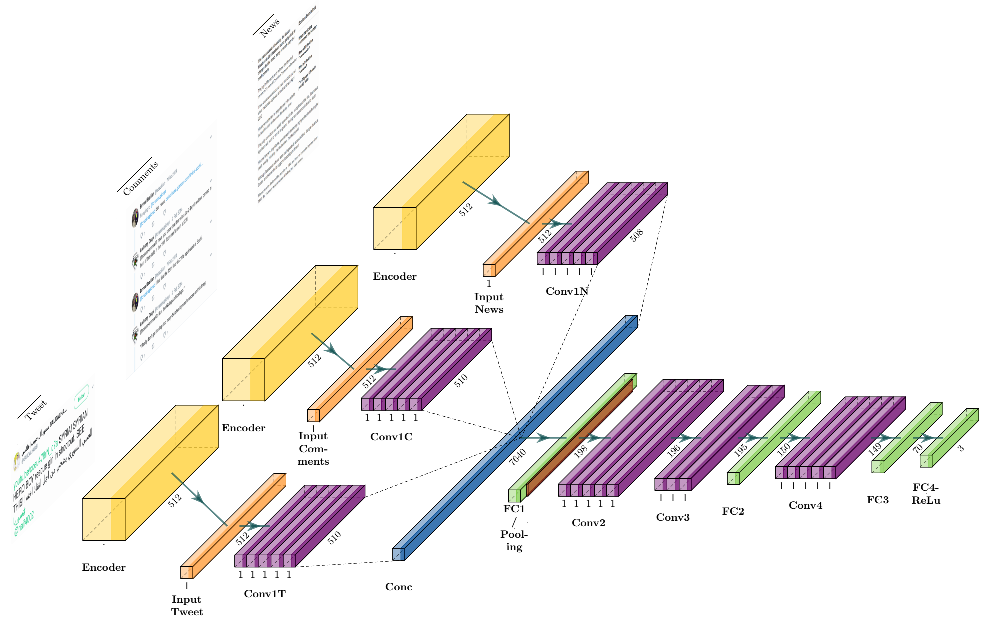

Fig. 1. The post (represented as a Twitter® tweet), the concatenated comments, and the news piece are individually fed to the adopted multilingual encoder transformer that outputs 512-dimension information-richer feature vectors. For each of these three vectors, a five-filter one-dimension convolution is applied, and the results are flattened and concatenated in a 7,640-dimension feature vector. Ultimately, this vector goes through a series of convolutional, max-pooling, and fully-connected layers until the final decision-making stage.

ing is to localize first and second-order features, such as agreements and disagreements among the merged text messages. In Table 1, we show the hyper-parameters settings of this proposed CNN model.

Layer ID	**Description**	**Hyper-parameters**			
		# filters	**filter size**	**stride**	**# units**
Conv1T	1D Convolutional	5	(1 x 5)	1	–
Conv1C	1D Convolutional	5	(1 x 5)	1	–
Conv1N	1D Convolutional	5	(1 x 3)	1	–
FC1	Fully-connected	–	–	–	200
Pooling	Max-pooling	–	(1 x 3)	1	–
Conv2	1D Convolutional	5	(1 x 3)	1	–
Conv3	1D Convolutional	3	(1 x 2)	2	–
FC2	Fully-connected	–	–	–	150
Conv4	1D Convolutional	5	(1 x 5)	1	–
FC3	Fully-connected	–	–	–	70
FC4-ReLU	Fully-connected	–	–	–	3

Table 1. Convolutional Neural Network hyper-parameters.

3.3. Classification Ensemble

To improve model predictions, at test time, we adopt an ensemble method organized in two steps: (1) obtain all vectors after the embedding layer (Input Tweet, Comments and News), and get the average for each class (on the training set); and (2) sum-up the values of the test post vector with the mean feature vector for each of the three classes of interest in this work (false, true or unknown). In this case, for every test sample, we generate three others. The intuition is to aggregate the knowledge of general linguistics characteristics from the encoder model of all categories.

The final prediction label is obtained through majority voting over the three classifications for each test sample and the classifications for the same post associated with different news pieces.

4. EXPERIMENTAL RESULTS

To effectively evaluate the proposed framework, we present some experiments to answer the following questions:

Q1 How text embeddings from a multilingual encoder and the usage of news pieces improve the identification of misleading content on social media?

Q2 How the comments contribute to improving classification performance?

Q3 How the ensemble method capture general information to better model the test set?

4.1. Dataset

To evaluate the proposed fake news detector, we adopted two publicly available datasets that were collected from Twitter®, and are part of two social media verification challenges held in 2015 [19] and 2016 [20]. The only difference between them is that the first one was created considering two categories (fake and real), while the other also has the unknown class. According to Boididou et al. [21], these datasets consist of tweets related to 17 events, but analyzing the image captions, we discovered more than 35 distinct events covering six languages. Since we are using a multilingual dataset, instead

Authorized licensed use limited to: MKSSS CUMMINS COLLEGE OF ENGINEERING FOR WOMEN. Downloaded on November 21,2020 at 07:31:08 UTC from IEEE Xplore. Restrictions apply.

of translating the tweets or add noise through projecting the samples onto a monolingual space, we take advantage of multilingual encoders [17]. As such, we have a linguistic invariant solution covering 16 languages, including the six presented in the dataset. As the original test dataset did not provide tweets classified as unknown, we randomly selected (and removed) 600 samples from the training set and moved them to the test set, totalling 30 percent of the amount.

We performed a pre-processing task removing the following items from the data: retweets; user references (@user); URLs, and; the # character from hashtags (in cases where the *CamelCase* pattern was applied, we added spaces between the concatenated words).

To obtain the tweet's comments (reply in Twitter® jargon), we developed a web-scraping application, which uses the tweet ID to find the required post and recover the target replies/comments. Although only three percent of the tweets are still available, it was enough for our experiments. For those tweets that are not available or there are no user replies, we set the input volume as zero.

To recover the related trustful news, we used the events keywords to retrieve the news from BBC and Reuters[1] websites. To prevent training and testing overlap of news samples, we used the BBC news in the training set and the Reuters news in the test set. The test set also contains 36 new events unseen in the training.

In our experiments, we associate each tweet with each retrieved news related to the same event E. In this way, the total number of samples is $S = \sum_E (|T_i^E||N^E|)$, where $|T_i^E|$ is the number of tweets related to an event E in the dataset, and $|N^E|$ is the number of retrieved news related to the same event E. Table 2 presents the original and augmented dataset distribution.

Class	Train	Test	Augmen. Train	Augmen. Test
Real	4314	2200	9304	22000
Fake	6690	3732	23026	36262
Unknown	1416	600	2361	600

Table 2. Dataset breakdown before and after data augmentation. All used data are freely available on Github.

4.2. Experimental Setup

We devised four different experiment scenarios to answer the three questions posed earlier. First, to assess the correlation between the post text and the news content, we performed two experiments: (1) considering only posts and news pieces related to the same event (CN – checked news); and (2) using samples that associate posts and news pieces with different events (UCN – unchecked news). In these analyses, we did not use the comments.

In the third experiment (CNC – checked news and comments), we added the comments to the previous CN scenario to evaluate the contribution of this information to the classification performance. Finally, the last experiment (CNCE – checked news, comments, and ensemble) included the classification ensemble strategy described in Section 3.3 over the CNC scenario.

We used the Tensorflow framework to build the EMET architecture. All the Python code developed to implement these models is freely available[2]. The machine specifications used to conduct the experiments are Intel®(R) 3.4GHz with 12 CPU cores, 128 GB of RAM, and 1 GPU NVIDIA® TITAN X.

4.3. Results and Discussion

We defined as baselines the works of Wang et al. [10] (EANN), Khattar et al. [11] (MVAE), and Qi et al. [12] (MVNN), despite some differences already discussed in Section 2. Table 3 presents the comparison between our results and the baselines over the same dataset, and the following paragraphs will discuss them.

Method	Domain	Accuracy	Precision	Recall	F1
EANN [10]	Text	53.2	59.8	54.1	56.8
	Multi	71.5	82.2	63.8	71.9
MVAE [11]	Text	52.6	52.7	53.9	53.2
	Multi	74.5	74.5	74.8	74.4
MVNN [12]	Visual	89.7	93.0	87.2	90.0
EMET	UCN	76.43	76.95	76.43	75.20
	CN	92.92	92.99	92.92	92.94
	CNC	93.47	**93.91**	**93.47**	**93.61**
	CNCE	**94.08**	91.31	91.21	91.26

Table 3. Results of baseline approaches vs. our proposed method.

To answer **Q1**, we compare our method (UCN and CN scenarios) with EANN and MVAE over a textual domain as the authors performed their experiments in a monolingual scenario and did not use any auxiliary textual data to detect misleading content. We see that the addition of trustworthy news content improves the accuracy significantly and that our two scenarios beat the baselines. We believe this can be easily explained by the fact that the added news content provides us with more information to perform the classification. To associate a piece of event-related credible news with the post, allows the network to capture better semantic cues to infer the reliability of the social media publication.

To answer **Q2**, we incorporate the tweet's comments into our model in scenario CNC. When compared with CN, we observed a significant improvement in all metrics, indicating that this information provides more discriminative features, showing complementary knowledge to improve the model performance.

The last scenario aims to answer **Q3**, experimenting over the ensemble strategy. The findings show that the ensemble method slightly outperforms the already mentioned experiments in terms of accuracy, presenting a modest reduction in precision, recall, and F1 metrics.

5. CONCLUSIONS AND FUTURE WORK

The problem of fake news detection can involve information on multiple domains, for instance, textual and visual. In this research, we worked only with textual data represented as signals, but all baselines also worked with multiple domains (text + images).

Despite the challenge in comparing approaches working over different types of data, for the task of fake news detection, our results suggest that additional textual data provides more information and better results than visual clues, at least in current social media publications.

In closing, we believe our presented method helps addressing the problem of fake news detection on social media platforms in a multilingual scenario. The EMET model achieved significant results on event-based publications, based on prior information and user interaction related to social media posts. For further studies, we will explore the addition of multidomain data since prior research demonstrated that it can boost the model performance. Different augmentation strategies along with the inclusion of the images as a side information also show promise for further refining decision-making.

[1] http://www.bbc.com/news and http://www.reuters.com

[2] https://github.com/stephanefschwarz/EMET

Authorized licensed use limited to: MKSSS CUMMINS COLLEGE OF ENGINEERING FOR WOMEN. Downloaded on November 21,2020 at 07:31:08 UTC from IEEE Xplore. Restrictions apply.

References

[1] Pew Research Center. Social media fact sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/, 2019. [Online; accessed on October 10th, 2019].

[2] Samantha Bradshaw and Philip N. Howard. 2019 global inventory of organised social media manipulation. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf, 2019. [Online; accessed on October 10th, 2019].

[3] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[4] Sabrina Tavernise. As fake news spreads lies, more readers shrug at the truth. https://www.nytimes.com/2016/12/06/us/fake-news-partisan-republican-democrat.html, 2016. [Online; accessed on October 10th, 2019].

[5] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[6] Anders Edelbo Lillie and Emil Refsgaard Middelboe. Fake news detection using stance classification: A survey. *arXiv preprint arXiv:1907.00181*, 2019.

[7] Dean Pomerleau and Delip Rao. Fake news challenge. http://www.fakenewschallenge.org, 2019. [Online; accessed on October 10th, 2019].

[8] Antônio Theóphilo, Luís AM Pereira, and Anderson Rocha. A needle in a haystack? harnessing onomatopoeia and user-specific stylometrics for authorship attribution of micro-messages. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2692–2696. IEEE, 2019.

[9] Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. Pheme: Computing veracity—the fourth challenge of big social data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session (ESCW-PN)*, 2015.

[10] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 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*, pages 849–857. ACM, 2018.

[11] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921. ACM, 2019.

[12] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. Exploiting multi-domain visual information for fake news detection. *arXiv preprint arXiv:1908.04472*, 2019.

[13] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.

[14] Peter Tolmie, Rob Procter, Mark Rouncefield, Maria Liakata, and Arkaitz Zubiaga. Microblog analysis as a program of work. *ACM Transactions on Social Computing*, 1(1):2, 2018.

[15] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511. IEEE, 2019.

[16] The Guardian. New ai fake text generator may be too dangerous to release, say creators. https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction, 2019. [Online; accessed on October 10th, 2019].

[17] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*, 2019.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[19] MediaEval Multimedia Benchmark. 2015 verifying multimedia use dataset. http://www.multimediaeval.org/mediaeval2015/verifyingmultimediause/, 2015. [Online; accessed on October 10th, 2019].

[20] MediaEval Multimedia Benchmark. 2016 verifying multimedia use dataset. http://www.multimediaeval.org/mediaeval2016/verifyingmultimediause/, 2016. [Online; accessed on October 10th, 2019].

[21] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

Authorized licensed use limited to: MKSSS CUMMINS COLLEGE OF ENGINEERING FOR WOMEN. Downloaded on November 21,2020 at 07:31:08 UTC from IEEE Xplore. Restrictions apply.