SpotFake: A Multi-modal Framework for Fake News Detection

Shivangi Singhal
IIIT-Delhi
Delhi,India
shivangis@iiitd.ac.in

Rajiv Ratn Shah
IIIT-Delhi
Delhi, India
rajivratn@iiitd.ac.in

Tanmoy Chakraborty
IIIT-Delhi
Delhi, India
tanmoy@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT-Delhi
Delhi, India
pk@iiitd.ac.in

Shin'ichi Satoh
NII
Tokyo, Japan
satoh@nii.ac.jp

***Abstract*—A rapid growth in the amount of fake news on social media is a very serious concern in our society. It is usually created by manipulating images, text, audio, and videos. This indicates that there is a need of multimodal system for fake news detection. Though, there are multimodal fake news detection systems but they tend to solve the problem of fake news by considering an additional sub-task like event discriminator and finding correlations across the modalities. The results of fake news detection are heavily dependent on the subtask and in absence of subtask training, the performance of fake news detection degrade by 10% on an average.**

To solve this issue, we introduce SpotFake- a multi-modal framework for fake news detection. Our proposed solution detects fake news without taking into account any other subtasks. It exploits both the textual and visual features of an article. Specifically, we made use of language models (like BERT) to learn text features, and image features are learned from VGG-19 pre-trained on ImageNet dataset. All the experiments are performed on two publicly available datasets, *i.e.,* Twitter and Weibo. The proposed model performs better than the current state-of-the-art on Twitter and Weibo datasets by 3.27% and 6.83%, respectively.

***Keywords*-Fake News Detection, Multimedia, Social Computing, Natural Language Processing, Deep Learning**

I. Introduction

"Fake news" is not new. The root of it existed in the society long back, but the damage done by it to mankind made it a serious issue to be solved by the research community. This term has now become a jargon, but the way it is defined is a bit different as compared to the earlier studies [1], [2]. Earlier, any kind of distinct content such as satires, hoaxes, news propaganda and clickbaits were termed as fake news. However, a recent study [3] define fake news *"to be news articles that are intentionally and verifiably false, and could mislead readers."* Moreover, such content is written with the intention to deceive someone. An example of such a false story is shown in Fig. 1. Interestingly, the image shown in the news is photo-shopped to make it look similar to the news that is generally featured on a popular news channel like CNN. This image made people believed that the news is real, but it was later quashed by the victim himself (see Fig. 2).

There can be various reasons for the spread of fake news. The first one is due to the lack of knowledge among people. The readers are unaware of the credibility of the sources

Figure 1: An example of fake news that claims that the actor Sylvester Stallone died due to prostate cancer.

[4], [5] and the authenticity of the news being read. This can have a huge negative impact on the masses. The second reason is lack of automated fact checking methods. Websites such as Politifact[1], Full Fact[2] and AltNews[3] make efforts in fake news detection but the time consuming manual method is too slow to prevent initial spread of a fake news.

This may be due to the fact that textual claims are not sufficient enough in detecting falsification. For instance, the text present in Fig. 3 (ii) says, *"the presence of sharks during Hurricane Sandy 2012"*, whereas deep analysis on the image concludes that it was spliced to show fake sharks in the image. Similarly, the text present in Fig. 3 (iii) says, *"picture of Solar Eclipse captured on March 20, 2015"*. However, the image is an artwork done so beautifully that it is hard to distinguish from reality. To authenticate this observation, we also did a survey on a sample population that consisted of people in the age group of 15-50 years. The observations obtained from the survey (see Section VI) confirms the fact that multi-modal features are more beneficial in detecting fake news as compared to uni-modal features. This lead us to propose **SpotFake** – a multi-modal framework for fake news detection. SpotFake takes into account the two modalities present in an article – text and image.

[1]http://www.politifact.com/
[2]http://www.fullfact.org/
[3]https://www.altnews.in/

978-1-7281-5527-2/19/$31.00 ©2019 IEEE
DOI 10.1109/BigMM.2019.00018

Figure 2: The reply from the actor after the spread of the news of his death.

The motivation to leverage multimodal information in SpotFake is as follows: (i) different modalities exhibit different aspects of a news, (ii) information derived from different modalities complement each other in detecting the authenticity of news, (iii) different sources manipulate different modalities based on their expertise (e.g., some people have experience in creating fake news by manipulating images and others may have experience in manipulating modalities such as text, audio and videos), and (iv) since real-world texts, photos, and videos are complex, contextual information is also important in addition to content information.

The main contribution of the paper is to design **SpotFake: a multimodal framework for fake news detection.** The propose architecture aims to detect whether a given news article is real or fake. It does not take into account any other sub-task in the detection process. The prime novelty of SpotFake is to incorporate the power of language models, i.e. Bidirectional Encoder Representations from Transformers (BERT) to incorporate contextual information. The image features are learned from VGG-19 pre-trained on ImageNet dataset. The representations from both the modalities are then concatenated together to produce the desired news vector. This news vector is finally used for classification.

The rest of the paper is organized as follows. Section II discusses the related work done in the domain of fake news detection with an emphasis on studies done using multimodal data. This is followed by a discussion of the proposed methodology in Section III and the different kind of dataset that is used in Section IV. In Section V, we present our experiment setup and detailed analysis of the observations. This is followed by Section VI, that discusses the details of public survey conducted to verify the importance of multiple modalities in detecting fake news. Finally, we conclude the paper with Section VII.

II. Related Work

Oxford Dictionary defines news to be a "newly received or noteworthy information, especially about recent events". There are various aspects present in a news article such as source, publisher, writing style followed, catchy headlines, content, image and fact-checking. Any changes made to any of these aspects lead to deceptive behaviour. Such a deceptive behaviour is often termed as "Fake News". Recent studies [3], [6], [7], [8] on fake news defines it to be " news articles that are intentionally and verifiably false, and could mislead readers."

To stop the proliferation of fake news, it is essential to detect sources that create such fake news. Various approaches to detect fake accounts include the use of cluster formation [9], random walks [10], steganalysis technique [11] and entropy minimization discretization [12]. Researchers also experimented with a mechanism that can generate an early warning to detect such accounts [13].

Majority of previous research done at news detection level was heavily dependent on text and user metadata features. Potthast et al. [7] showed how writing style, network connection and user reaction can lead to the detection of fake news. Moreover, Shu et al. [14] described how the writing style of an author impacts the views and opinions of people reading such articles. This plays a vital role in shaping the opinions of the masses.

To improve fact analysis in news content, Pan et al. [15] used knowledge graphs. Entity relation information extracted out of these graphs can be used to induce commonsense reasoning into text content. Recently, Lin et al. [16] used TransR model to generate knowledge graph embeddings (KGE) for the entity relation triplets extracted from news articles. The advantage of using TransR to get KGE is that, it builds entity and relation embeddings in distinct spaces. KGE learning is then done by projecting entities from entity space to corresponding relation space and then building translations between projected entities.

Forging fake images is a popular way to tamper news. To detect such incidents image splicing technique [17] was used that takes input as the EXIF metadata information and determines whether the image is self-consistent or not. Recently, Marra et al. [18] used GANs to detect fake images.

Though all the above mentioned uni-modal techniques were able to provide promising results, short and informal nature of social media data always becomes a challenge in information extraction. To overcome this limitation, the researchers started experimentation with features extracted from multiple modalities (i.e. text and image) and fused them together for richer data representation . Works [19], [20], [21] are the most notable studies in multimodal fake news detection.

Wang et al. [20] built an end-to-end model for fake news detection and event discriminator. This is termed as Event Adversarial Neural Networks for Multi-Modal Fake News Detection (EANN). Their model has two components. The text part took word embedding vector as input and generated text representation using a CNN [22]. Image representation was extracted from VGG-19 model pre-trained on ImageNet [23]. Finally, both of these representations were concatenated and fed in two fully connected neural network

classifiers, one for event discriminator and another for fake news classification.

Inspired by [20] architecture, Khattar et al. [21] built a similar architecture. They named it as Multimodal Variational Autoencoder for Fake News Detection (MVAE). The primary task was to build an auto encoder-decoder model. They used bi-directional LSTMs to extract text representation and image representation was again extracted from VGG-19. The latent vectors produced by concatenation of these two vectors were fed into a decoder for reconstructing the original samples. The same latent vectors were also used for a secondary task of fake news detection.

Though these multimodal systems perform well in detecting fake news, the classifiers have always been trained in tandem with another classifier. This increases training and model size overhead, increases training complexity and at times can also hinder the generalizability of the systems due to lack of data for the secondary task.

To solve such issues, we design SpotFake- a multimodal framework for fake news detection. It takes into consideration features from two different modalities and classifies the sample into real or fake without taking into account any other sub-task. Next, we highlight the details of the SpotFake.

III. Methodology

For this study, we conduct a public survey (see Section VI-A for details) to do an empirical analysis on human performance, difficulties of fake news detection, and the importance of multiple modalities especially the combination of text and image for fake news detection. Based on our survey results and previous literature, it is evident that a multimodal system is necessary for fake news detection. However, we wanted our system to be able to detect fake news independently without any other subtask, as seen in the current state-of-the-art systems. The fake news classifier of the current state-of-the-art system does not perform well by itself. However, performance significantly improves in the presence of a secondary task like sample reconstruction [21].

To this end, we propose SpotFake- a multimodal framework for fake news detection. SpotFake is divided into three sub-modules. The first sub-module is textual feature extractor that extract the contextual text features using a language model. The second sub-module is visual feature extractor that extract the the visual features from a post. Finally, the last sub-module is a multimodal fusion module that combines the representations obtained from different modalities together to form news feature vector. The complete outline of SpotFake is shown in Fig. 4.

Textual Feature Extractor

This is a sub-module of SpotFake that is responsible of extracting the contextual text features from the posts. It uses Bidirectional Encoder Representations from Transformers (BERT) [24] to represent words and sentences in a way that best captures underlying semantic and contextual meaning. We use BERT-base version that has 12 encoding layers (termed as transformer blocks). It takes as input a sequence of words that keep moving up the stack. Each layer applies self-attention, and passes its results through a feed-forward network, and then hands it off to the next encoder. A detailed description of the textual feature extractor is shown in Fig. 5.

In Fig. 5, [CLS] denotes the classification and w_i refers to the sequence of tokens that are placed as an input to the textual feature extractor sub-module. The features obtained from the second last output layer of the module are the desired contextual embedding of the post that are then passed through a fully-connected layer to reduce down to final dimension of length 32. These textual feature vectors are denoted as Tf.

Visual Feature Extractor

In general, the information in the visual form is logically learned and understood much faster by brain than in the textual form. Based on this intuition, we take into account the visual features of the posts. We employ the pre-trained VGG-19. We extract the output of the second last layer of VGG-19 convolutional network pre-trained on ImageNet dataset (denoted by $\mathrm{V_g}$) and pass it through a fully connected layer to reduce down to final dimension of length 32. The final visual representation (denoted as Vf) is obtained as follows:

$$Vf = \sigma(W.V_{\mathrm{g}}) \tag{1}$$

where $\mathrm{V_g}$ is the visual feature representation obtained from pre-trained VGG19 [23], and W is the weight matrix of the fully connected layer in the visual feature extractor.

Multimodal Fusion

The two feature vectors obtained via different modalities (i.e. Tf and Vf) are fused together using simple concatenation technique to obtain the desired news representation. This news representation is then passed through a fully connected neural network for fake news classification.

IV. Dataset

The training of SpotFake is performed on two publicly available datasets *i.e.,* Twitter and Weibo.

A. Twitter MediaEval Dataset

The dataset was released as the part of the challenge- The Verifying Multimedia Use at MediaEval [25]. The challenge aimed to find whether the information presented by the post sufficiently reflects reality. The dataset comprises of tweets and their associated images. It consists of 17,000 unique tweets related to various events. The training set consists of 9,000 fake news tweets and 6,000 real news tweets, and the test set containing 2,000 news tweets. A sample of fake images from the dataset is illustrated in Fig. 3.

Figure 3: (i) real photo of two Vietnamese siblings but being presented as it was captured during the Nepal 2015 earthquakes; (ii) photos of spliced sharks taken during Hurricane Sandy in 2012; (iii) a beautiful artwork portrayed as a picture of Solar Eclipse of March 20, 2015.

B. Weibo Dataset

In this dataset, the real news is collected from authoritative news sources of China, such as Xinhua News Agency and Weibo- a microblogging website in China [19]. The fake news is collected from Weibo in the time duration of May 2012-June 2016. The collected set of news is verified by the official rumour debunking system of Weibo. This system also allows masses to report suspicious posts which are then verified by a committee of trusted users. According to the previous work by [19] , the system also acts as the authoritative source for collecting rumor news in China by providing fact-checking evaluation results.

V. Experiment

In this section, we highlight our experiment setup and the model parameters used in training SpotFake. This is followed by a detailed analysis of the results of proposed model with the current SOTA.

A. Experimental Setup of SpotFake

To initiate, tweets and their respective images are filtered out from the dataset and pre-processed. For the text modality, only pre-processing step done is fixing the input length of the sequence. All the sentences above the decided length are trimmed, and anything below the decided length is padded with zeroes. The final length value is decided to be the one where 95% of the sentences are below it. This is 23 tokens for the Twitter dataset and 200 characters for Weibo dataset (Since Weibo is a Chinese dataset, tokenization is done at character level). For the image component of the model, all the images are resized to 224x224x3.

For our text feature extractor, we use pretrained BERT-Base available on tfhub. The padded and tokenized text is passed into the BERT model to receive word vectors of dimension 768. These vectors of length 768 are then passed through two fully connected layers of size 768 and 32, respectively. Similarly, the re-sized images are passed through VGG-19 pretrained on ImageNet, and a vector of length 4096 is extracted. This vector is then passed through two fully connected neural network layers of size 2742 and 32, respectively (For Weibo dataset, we have one fully connected layer of size 32).

The 32-dimensional vectors of both the modalities are concatenated and passed into a fully connected neural network classifier with a hidden layer of size 35 and classification layer of size 1 with sigmoid activation function.

Every fully connected layer in the model has a relu activation function and a dropout probability of 0.4. The model is trained on a batch size of 256 and Adam optimizer using early stopping done on validation accuracy.

B. Hyperparameter Tuning for SpotFake

From the number of hidden layers to the number of neurons, and the dropout probabilities, everything is configurable in our model. A full, exhaustive list of hyperparameters is given in TableI.

For selecting the correct permutation of hyperparameters, we perform iterations of random search on possible hyperparameter combinations. In each iteration, the number of possible permutations is reduced based on the performance of the previous iteration. For conducting random search and evaluating parameters in a random search, talos[4] library is used. Cycling learning rate proposed in [26] is used to find an optimal learning rate for our models.

C. Results

In this section, we are reporting the performance comparison of SpotFake with the current state-of-the-art EANN [20] and MVAE [21] on accuracy %, and precision, recall and F1-score of each class. The complete comparison results are shown in Table II.

Both EANN and MVAE have two models configuration each. EANN-/MVAE- is when fake news classifier is trained standalone. EANN/MVAE is when fake news classifier is trained in tandem with a secondary task. The secondary task

[4]https://github.com/autonomio/talos

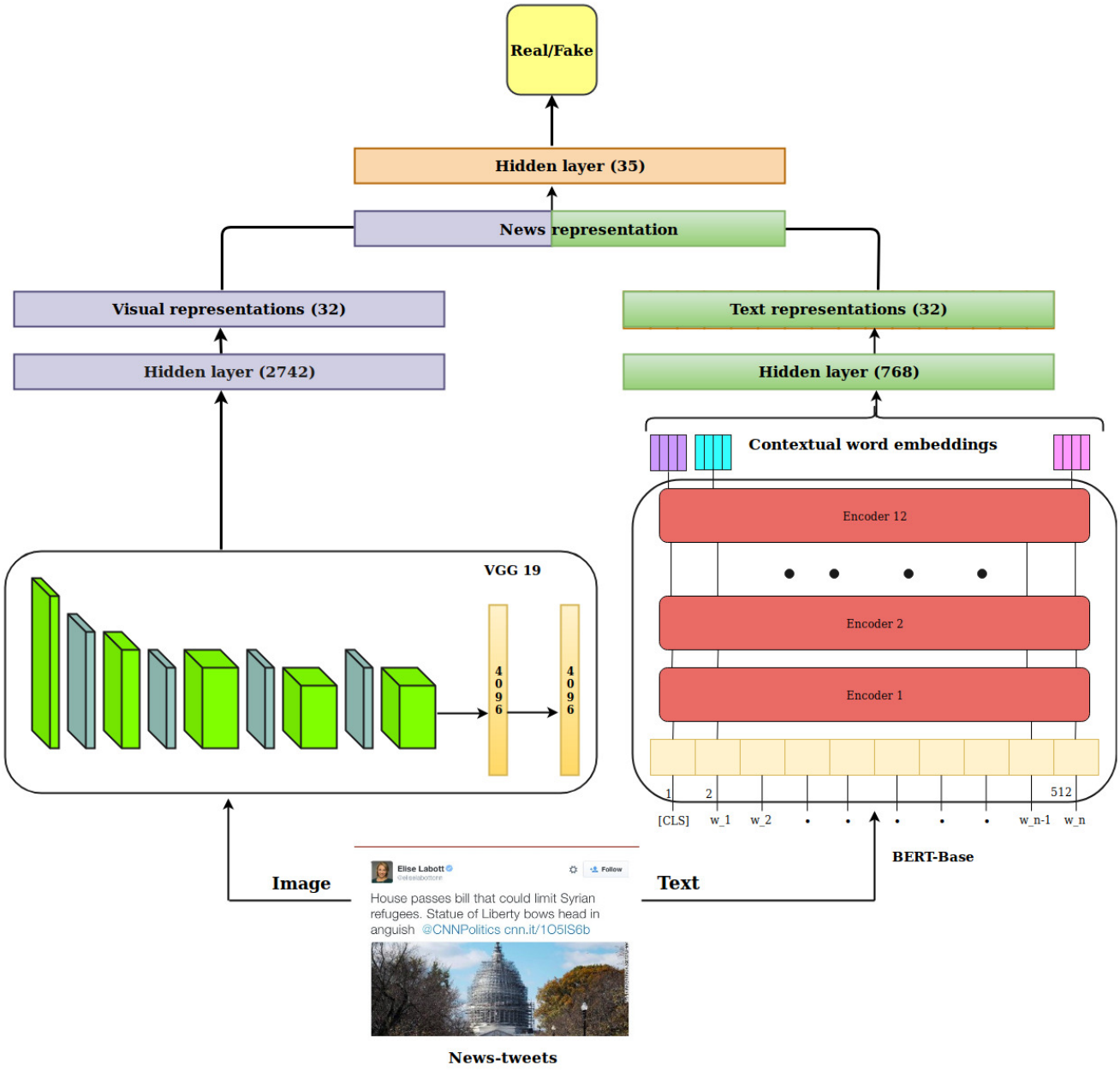

Figure 4: A schematic diagram of the proposed SpotFake model. Value in () indicates number of neurons in a layer.

in case of EANN is an event discriminator that removes the event-specific features and keep shared features among events. Whereas in MVAE, the secondary task is to discover the correlations across the modalities to improve shared representations.

Though SpotFake is a standalone fake news classifier, we still outperform both configurations of EANN and MVAE by large margins on both the datasets. On the Twitter dataset, SpotFake achieves 12.97% and 6.27% accuracy gain over EANN- and EANN respectively. Performance gain on Weibo dataset over EANN- and EANN is 9.73% and 6.53% respectively. The brief overview of the results is given in Table III. When compared to MVAE, on twitter dataset, we outperform MVAE-and MVAE by 12.17% and 3.27% respectively. On Weibo dataset, the performance gain is 14.93% and 6.83%. The brief overview of the results is given in Table IV.

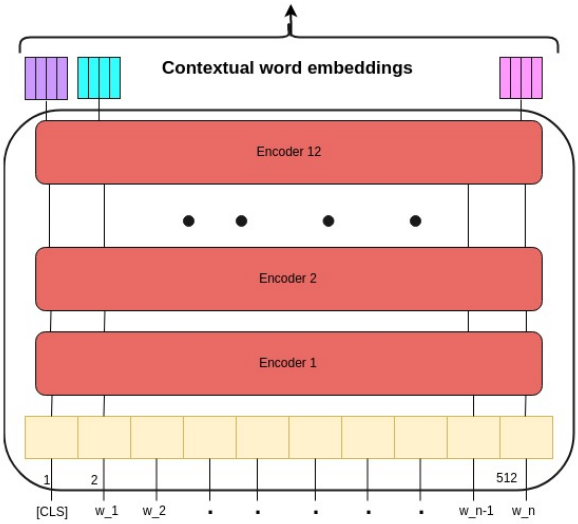

Figure 5: A high level diagram of textual feature extractor (BERT).

parameters	Twitter	Weibo
BERT trainable	False	False
VGG trainable	False	False
dropout	0.4	0.4
# of hidden layers (text)	2	2
# of neurons in hidden layer (text)	768,32	768,32
# of hidden layers (image)	2	1
# of neurons in hidden layer (image)	2742,32	32
# of dense layers (concatenation)	1	1
# of neurons in dense layer (concatenate)	64	64
text length	23 words	200 char
batch size	256	256
optimizer	adam	adam
learning rate	0.0005	0.001

Table I: An overview of hyper parameter setting used in SpotFake.

VI. Survey

The aim of conducting this survey is to get an understanding of the people's perception about fake news, sources that are more susceptible to the spread of fake news and the effect of multiple modalities on the human ability to detect fake news.

First few questions of the survey aimed at finding out people's awareness about the concept of fake news, the source they rely on reading news online with reasoning for doing the same and their ability to detect fake content. Next part of the survey contains questions that judge human ability to detect fake news in the presence of different modalities. Survey taker is asked to differentiate between fake news and real news via only text, only image or using both the modalities. Few samples of the questions asked are shown in Fig. 6 and 7.

The outcomes of this survey indicated that fake news detection performance can be improved in the presence of both modalities. More detailed results of the survey are discussed next.

A. Observations from Survey

The survey [5] consists of twenty questions. A total of 88 candidates participated in the survey, 63% of them were male. The majority of candidates (about 64%) were in the age group of 15-50 years.

All the findings from the survey are listed below.

- Our study shows that the veracity of the content is important to 98% of the people. 17% of the people are not really sure about what is termed as Fake News. This is a big concern because people should be aware of the existence of fake news and know what are the consequences of such news and how it can harm people.
- Approximately 91% respondents believe that traditional outlets are more trustworthy sources to get news in comparison to OSM. This questions the credibility of any information obtained via OSM.
- We find that 44.3% of people are not successful in identifying whether they are reading fake news. This means that fake news is crafted in such a way that users are not able to classify it as fake. This clearly indicates the necessity of designing an automated system to help detect fake news.
- Initially, around 71% of the people believed that it would be easy for them to distinguish between real and fake news when given two modalities. Then various questions regarding the same were asked. Finally, the importance of multiple modalities in detecting fake news was asked again. This time it was observed, that 81.4% of the people believed that they were able to distinguish between fake and real news when given two modalities (image and text) as compared to 38.4% in case of text only and 32.6% in case of image only.
  We see a jump of more than 40% when we give more modalities as compared to uni-modality for fake news detection. This clearly shows that having multiple modalities provides more information and makes it easier to detect fake news. This is shown in Fig. 8.

VII. Conclusion

Previous literature has attacked the problem of detecting fake news from different angles like natural language processing, knowledge graphs, computer vision and user profiling. It has been shown that for consistent results, a multimodal method is required. Where the current multimodal

[5]We will release the survey and its responses upon acceptance of this paper

Dataset	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1-Score	Precison	Recall	F1-Score
Twitter	textual	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	visual	0.596	0.695	0.518	0.593	0.524	0.7	0.599
	VQA [27]	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural Talk [28]	0.610	0.728	0.504	0.595	.534	0.752	0.625
	att-RNN [29]	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN- [20]	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	EANN [20]	0.715	NA	NA	NA	NA	NA	NA
	MVAE- [21]	0.656	NA	NA	0.641	NA	NA	0.669
	MVAE [21]	0.745	**0.801**	0.719	0.758	0.689	**0.777**	**0.730**
	SpotFake	**0.7777**	0.751	**0.900**	**0.82**	**0.832**	0.606	0.701
Weibo	textual	0.643	0.662	0.578	0.617	0.609	0.685	0.647
	visual	0.608	0.610	0.605	0.607	0.607	0.611	0.609
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	Neural Talk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN-	0.795	0.827	0.697	0.756	0.752	0.863	0.804
	EANN	0.827	NA	NA	NA	NA	NA	NA
	MVAE-	0.743	NA	NA	NA	NA	NA	NA
	MVAE	0.824	0.854	0.769	0.809	0.802	**0.875**	**0.837**
	SpotFake	**0.8923**	**0.902**	**0.964**	**0.932**	**0.847**	0.656	0.739

Table II: Classification Results on Twitter and Weibo datasets.

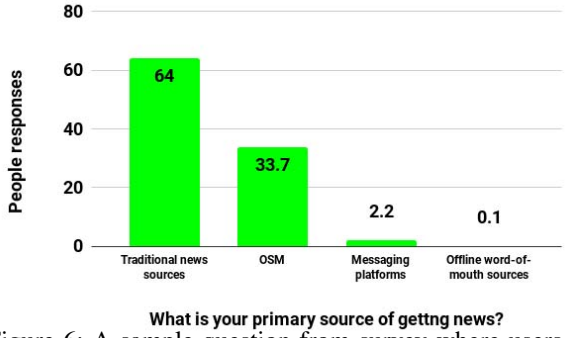

Figure 6: A sample question from survey where users were asked to tell their primary source of getting news.

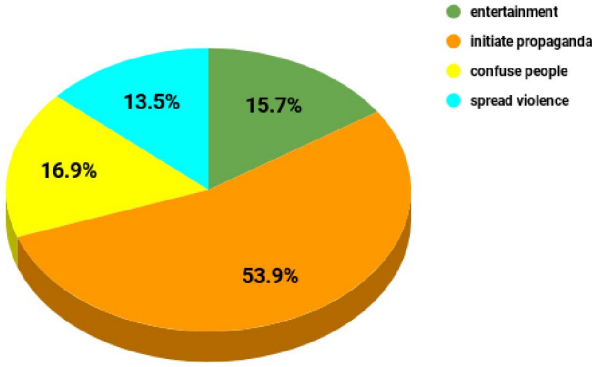

Figure 7: A sample question from survey where users were asked to voice their opinion on the reasons that they think lead to spread of fake news.

state-of-the-art suffers from a problem of not being able to learn from fake news detection problem as a primary task, SpotFake uses language transformer model and pre-trained ImageNet models for extraction and classifies using fully connected layer. It outperforms the baselines by a margin of 6% accuracy on an average. There is still room for improvement on longer length articles and more complex fusion techniques to understand how different modalities play a role in fake news detection.

	Accuracy	
Model	**Twitter**	**Weibo**
EANN- [20]	64.8 (12.97)	79.5 (9.73)
EANN [20]	71.5 (6.27)	82.7 (6.53)
SpotFake	**77.77**	**89.23**

Table III: Performance comparison of EANN with SpotFake in terms of %.

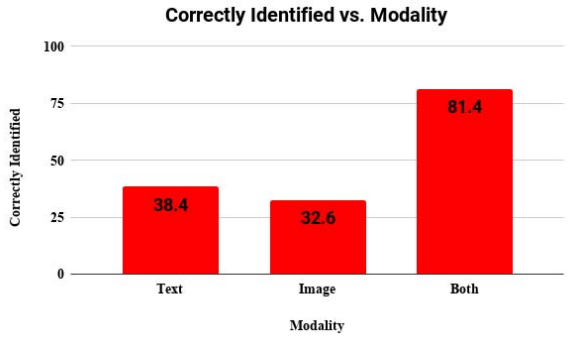

Figure 8: Percentage of people that were successful in identifying fake news when given different modalities.

	Accuracy	
Model	**Twitter**	**Weibo**
MVAE- [21]	65.6 (12.17)	74.3 (14.93)
MVAE [21]	74.5 (3.27)	82.4 (6.83)
SpotFake	**77.77**	**89.23**

Table IV: Performance comparison of MVAE with SpotFake in terms of %.

ACKNOWLEDGEMENT

Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi and ECRA Grant by SERB, Government of India. Shivangi Singhal is partly supported by the NII International Internship program and MIDAS Lab. The work is partially supported by DST, India (DST/INT/UK/P-158/2017, ECR/2017/001691), and Infosys Centre for AI, IIITD.

REFERENCES

[1] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as "false news"," in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, ser. WMDD '15. New York, NY, USA: ACM, 2015, pp. 15–19. [Online]. Available: http://doi.acm.org/10.1145/2823465.2823467

[2] S. De Sarkar, F. Yang, and A. Mukherjee, "Attending sentences to detect satirical fake news," in *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018, pp. 3371–3380. [Online]. Available: http://aclweb.org/anthology/C18-1285

[3] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, Spring 2017. [Online]. Available: https://ideas.repec.org/a/aea/jecper/v31y2017i2p211-36.html

[4] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 675–684. [Online]. Available: http://doi.acm.org/10.1145/1963405.1963500

[5] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: A real-time web-based system for assessing credibility of content on twitter," *CoRR*, vol. abs/1405.5490, 2014.

[6] E. Mustafaraj and P. T. Metaxas, "The fake news spreading plague: Was it preventable?" *CoRR*, vol. abs/1703.06988, 2017.

[7] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *CoRR*, vol. abs/1702.05638, 2017.

[8] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ser. ASIST '15. Silver Springs, MD, USA: American Society for Information Science, 2015, pp. 82:1–82:4. [Online]. Available: http://dl.acm.org/citation.cfm?id=2857070.2857152

[9] C. Xiao, D. M. Freeman, and T. Hwa, "Detecting clusters of fake accounts in online social networks," in *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, ser. AISec '15. New York, NY, USA: ACM, 2015, pp. 91–101. [Online]. Available: http://doi.acm.org/10.1145/2808769.2808779

[10] J. Jia, B. Wang, and N. Z. Gong, "Random walk based fake account detection in online social networks," in *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, June 2017, pp. 273–284.

[11] E. Ahmadizadeh, E. Aghasian, H. Pour Taheri, and R. Fallah, "An automated model to detect fake profiles and botnets in online social networks using steganography technique," *iosrjournals*, vol. 17, pp. 2278–661, 02 2015.

[12] B. Erahin, . Akta, D. Kln, and C. Akyol, "Twitter fake account detection," in *2017 International Conference on Computer Science and Engineering (UBMK)*, Oct 2017, pp. 388–392.

[13] H. Halawa, M. Ripeanu, K. Beznosov, B. Coskun, and M. Liu, "An early warning system for suspicious accounts," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, ser. AISec '17. New York, NY, USA: ACM, 2017, pp. 51–52. [Online]. Available: http://doi.acm.org/10.1145/3128572.3140455

[14] K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection," *CoRR*, vol. abs/1712.07709, 2017.

[15] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, "Content based fake news detection using knowledge graphs," in *International Semantic Web Conference (1)*, ser. Lecture Notes in Computer Science, vol. 11136. Springer, 2018, pp. 669–683.

[16] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI Press, 2015, pp. 2181–2187. [Online]. Available: http://dl.acm.org/citation.cfm?id=2886521.2886624

[17] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," *CoRR*, vol. abs/1805.04096, 2018.

[18] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, April 2018, pp. 384–389.

[19] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 2017, pp. 795–816. [Online]. Available: https://doi.org/10.1145/3123266.3123454

[20] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. 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*, ser. KDD '18. New York, NY, USA: ACM, 2018, pp. 849–857. [Online]. Available: http://doi.acm.org/10.1145/3219819.3219903

[21] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: ACM, 2019, pp. 2915–2921. [Online]. Available: http://doi.acm.org/10.1145/3308558.3313552

[22] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014. [Online]. Available: http://arxiv.org/abs/1408.5882

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[25] C. Boididou *et al.*, "Verifying multimedia use at mediaeval 2015."

[26] L. N. Smith, "No more pesky learning rate guessing games," *CoRR*, vol. abs/1506.01186, 2015. [Online]. Available: http://arxiv.org/abs/1506.01186

[27] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "Vqa: Visual question answering," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.

[28] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[29] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 795–816.