Text-Convolutional Neural Networks for Fake News Detection in Tweets

Harsh Sinha, Sakshi Kalra, and Yashvardhan Sharma

Birla Institute of Technology and Science, Pilani, INDIA
{h20130838,p20180437,yash}@pilani.bits-pilani.ac.in,

Abstract. With the widespread use of online social networking websites, user generated stories and social network platform have become critical in news propagation. The web portals are being used to mislead users for political gains. Unreliable information is being shared without any fact-checking. Therefore, there is a dire need for automatic news verification system which can help journalists and the common users from misleading content. In this work, the task is defined as being able to classify a tweet as real or fake. The complexity of natural language constructs along with variegated languages makes this task very challenging. In this work, a deep learning model to learn semantic word embeddings is proposed to handle this complexity. The evaluations on the benchmark dataset (VMU 2015) show that deep learning methods are superior to traditional natural language processing algorithms.

Keywords: social media, twitter, fake news

1 Introduction

Social networking websites have become an integral part of news information collection. This can be attributed to the fact that such social networking platforms provide an open stage to express opinions and share content. Moreover, spreading information has become very easy. Any article can be swiftly shared among all friends and followers. With the massive increase in social interactions on online social networks, there has also been an increase of misleading activities that exploit such infrastructure. Such rapid dissemination of misleading opinions can also have devastating effects especially during natural calamities like hurricanes, or terrorist attacks. The inaccurate information is generally written to mislead community for political gains. However, fake information can also be spread in ignorance as there is no fact checking. Communities share information not on its credibility but just by reading headline.

Fake news refers to any multimedia content which contains misleading information about the event it is associated with. For example, a user may post an image out of context. It is also seen that users share pictures which are concerned with some other similar event. A user may manually edit or morph an image as a form of amusement. It is found that fake news in general contains a false claim and an associated image. During crisis such as hurricanes, explicit

fake images generate fear and numerous people fall for it. On the other hand, there are genuine or real news which represent the event truthfully. Such posts are helpful for the community to make the community aware and safe. There are another class of posts which are fake but are propagated in a sarcastic manner. In this paper, a deep learning approach is employed for supervised learning for benign classification of tweets as real or fake. Fake news detection is arduous as fact-based checking for news is not feasible. To solve the problem, the approach must be able to learn latent representation of fake news data. Therefore, the paper proposes a Convolutional Neural Network (CNN) based deep learning approach to learn specific latent representation for accurate classification.

Prior works have used techniques such as graphs[2][11][9] and anomaly detection [12]. However, this work focuses on extracting, higher level representations from raw input text. With advent of GPUs, there has been significant development in deep learning especially Convolutional Neural Networks (CNNs). The primary cause which lead to proliferation of CNNs across domains is its agility in reducing variations and extracting spatial correlations.

Owing to widely accepted success of CNNs across several domains, this work investigates whether CNNs can be used efficiently for text classification. This work explores the ability of CNNs to learn correlations in natural language constructs while being invariant to individual characteristics of a user.

Researchers model text as sequences. Neural architectures such as Recurrent Neural Networks (RNNs) and Long Short Term Memory networks (LSTMs) are employed for sequence processing. In this work, 1-dimensional CNNs are used as CNNs are competent in extracting space-invariant features. RNNs are useful in predicting the areas such as machine translation or image captioning. However, CNNs are superior in classifying a sentence. CNNs can extract latent features pivoted for accurate classification. Moreover, CNNs are superior as they are extremely efficient and fast in comparison to RNNs. This work focuses on learning an optimal CNN that can be successfully applied for text classification.

Fig. 1. Framework for CNN based fake news detection

The rest of the paper is arranged accordingly. Section 2 describes the prior techniques used for rumor detection. An overview of the proposed methodology is available in Section 3. In Section 4 details the experimental set-up and preprocessing methods . A discussion of the outcomes and relative performance

measures is presented in Section 5. The article concludes with a short overview of the work in Section 6.

2 Related Work

Rumors over the internet have been addressed specifically in the domains of Web Spam detection. The initial taxonomy was proposed by Gyongyi et al. [7]. Further, to combat web spam Castillo et al. [2] used a graph based structure to infer link-based and content-based dependencies between web pages. Seo et al. [11] developed a way to study how rumors spread on social networks. It was found that misleading information is generated from a few sources which is re-posted by several other users. Similarly, the dissemination of tweets was studied extensively by Mendoza et al. [9]. They present tweets specifically propagated during crisis and emergencies. Researchers have also tried to employ user characteristic information to model a binary classification problem of users as spammers and non-spammers [1]. A similar study on user information was carried out by Stringhini et al. [12] by specifically studying anomalous behaviour.

Gupta et al. [5] presents that during Mumbai Terrorist Attacks, there was a greater dominance of misleading information. They propose an automated classification framework to calculate credibility of tweets [6]. Cheong et al. [3] analyze Twitter users and their interaction during natural calamities like floods to find the propagation of information on social networking platforms during crisis.

In this work, a deep learning approach is proposed to learn latent features that credibly differentiates a real news from fake tweet. Similar studies have been conducted by Gupta et al. [6] which detect fake news during Hurricane Sandy. However, the difference lies in obtaining a realistic estimate by conducting experiments on benchmark datasets which are not restricted to a particular crisis event.

3 Proposed Methodology

The block diagram of the proposed methodology is shown in Figure 1. The major components of the proposed framework include preprocessing, embedding matrix generation and classification.

3.1 Data Preprocessing

Tweets crawled from Twitter API contains non-ascii characters which have to be removed for efficient classification. First of all the HTML tags were removed as they cannot be converted to text. Secondly, in conversational tweets there are user mentions by '@' symbol. It is useful to build a fake tweet propagation model. However, the proposed model attempts to learn difference in latent natural language constructs to classify a tweet as fake or real. Thus, it doesn't add value

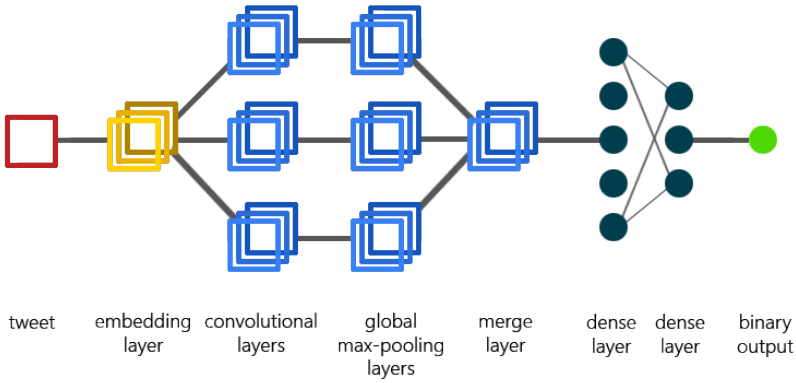

Fig. 2. Proposed CNN architecture

for text based classification. Several other preprocessing steps include removal of urls, UTF-8 BOM, and hashtags. Once the text was perfectly clean without any special characters, the text corpus was used to learn an embeddings matrix.

3.2 Embedding Matrix

Embedding matrix is a conceptual idea to model text as points in a n-dimensional hyperspace or more conveniently, as a 'one-hot' encoded vector. The n-dimensional vector represents different words in the sentence. Assuming one word per dimension can lead to a very high dimensional space, so the vectors are transformed into lower vector subspace.

In the n-dimensional hyperspace, product of one-hot encoded vector with an embedding matrix results in a word embedding. The embedding matrix can be represented as $W \epsilon \mathbb{R}^{e \times n}$ where e represents the embedding dimensionality and n is size of vocabulary. This reduces the input size and avoids over-fitting. Finally, each word can be represented by a e-dimensional vector and every tweet is composed of several words.

3.3 Classification

The embedding matrix representing a tweet can be used for classification using a Convolutional Neural Network (CNN).

The first layer of proposed CNN represents the input layer which feeds tweets as matrices as explained in Section 3.2. These high dimensional representation is reduced to a subspace using an embedding layer. The output matrix from

the embedding layer $I \epsilon \mathbb{R}^{w \times h \times c}$ where w, h and c are matrix width, height, and number of channels respectively. A filter matrix $K \epsilon \mathbb{R}^{k \times k \times n}$ is convoluted with the matrix I, which results in n activation maps. A convolution layer is followed by a global max pooling layer which sub-samples the resultant activation maps of the previous layer. A pooling layer is used to obtain an efficient representation of data, while rejecting unimportant spatial information. A pooling layer with filter size $k \times k$ computed on a matrix of size $I \epsilon \mathbb{R}^{w \times h \times c}$ generates a matrix of size $P \epsilon \mathbb{R}^{\frac{w}{k} \times \frac{h}{k} \times c}$. However, a global max pooling layer is used in proposed methodology which outputs a matrix of size $P \epsilon \mathbb{R}^{\frac{w}{k} \times \frac{h}{k} \times 1}$. A global max pooling layer is very helpful in language processing domains. As dense layers are prone to overfitting, global max-pooling layer preserves the spatial information while reducing the number of parameters for accurate classification. In the proposed architecture, three different convolutional layers are used with their respective global max-pooling layers as shown in Figure 2. Finally, all the matrices are concatenated and fed for classification using 2 dense layers. In order to reduce overfitting, every convolutional layer and dense layer is succeeded by a dropout layer.

The final softmax loss was replaced with binary-crossentropy loss. The softmax loss is explained in (1).

$$\begin{aligned} L(x_a, x_b) &= -log\left(\frac{e^{f(x_b)}}{\sum_{j=1}^{m} e^{f(x_b)}}\right) \\ &= log\left(\sum_{j=1}^{m} e^{f(x_b)}\right) - f(x_b) \end{aligned} \tag{1}$$

where L denotes the softmax loss between two samples x_a and x_b, f denotes the linear transformation of sample ($f(x) = W_i \cdot x + b$), using W as the weight matrix and b as bias).

The binary crossentropy loss is depicted in (2)

$$L(x_a, x_b) = (y_b \cdot log(f(x_b) + (1 - y_b) log(1 - f(x_b)) \tag{2}$$

where y_b denotes the true label of the sample. Unlike the softmax loss which depends on Boltzmann's distribution, binary cross entropy loss depends on Shannon's information entropy. The binary cross entropy loss takes into account each component independently.

The goal of proposed deep neural architecture is to learn suitable filters using back-propagation for accurate classification of tweet as fake or real.

4 Experiments

The following section explains various datasets, experiments and different hyperparameters used for performance evaluation of the proposed methodology.

The proposed CNN is evaluated in terms of average recognition accuracy.

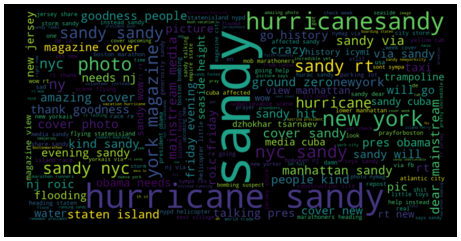

Fig. 3. Wordcloud depicting the frequency of words in tweets used for training

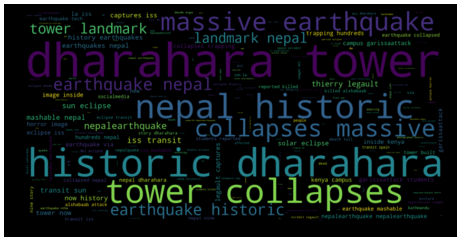

Fig. 4. Wordcloud depicting the frequency of words in tweets used for testing

4.1 Dataset

The Verifying Multimedia Use (VMU) 2015 dataset contains a corpus of tweets classified as real and fake with their images shared on the social networking website Twitter. The dataset was used in Verifying Multimedia Use Workshop 2015 [4].

The dataset consists of tweet ID, tweet text, user ID, associated image ID, associated username, timestamp and label as real or fake. The training dataset collects tweets associated with events such as Boston Marathon, Columbian Chemicals, Hurricane Sandy, Malaysia Airlines MH-370 and Sochi Olympics. However, the test dataset contains tweets associated with events such as Garissa Attacks and Nepal Earthquake. The illustration presents the datasets as wordclouds in Figure 3 and Figure 4.

4.2 Feature Extraction

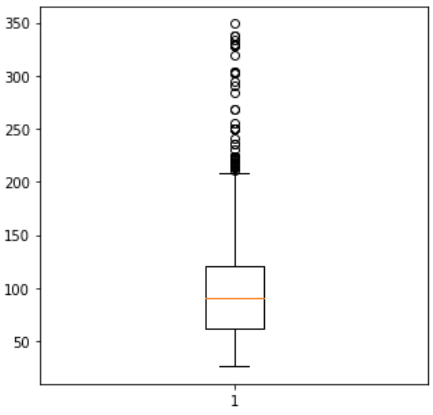

Fig. 5. Boxplot depicting distribution of length of tweets (number of characters pr tweet)

The distribution of length of tweets is shown in Figure 5. The distribution is not correct as twitter's character limit is 140 characters. Thus, the tweet text is cleaned such that it contains ascii characters only. The several preprocessing steps included removal of HTML tags, UTF-8 BOM, non-english characters, and urls. Any special symbols such as hashtags and '@' mentions were also removed.

Further, for classification using CNN each tweet is represented by a vector, where each word is represented by a natural number using a tokenizer word index. Each vector representing a tweet is padded with zeros such that all the tweets have equal length.

5 Results and Discussion

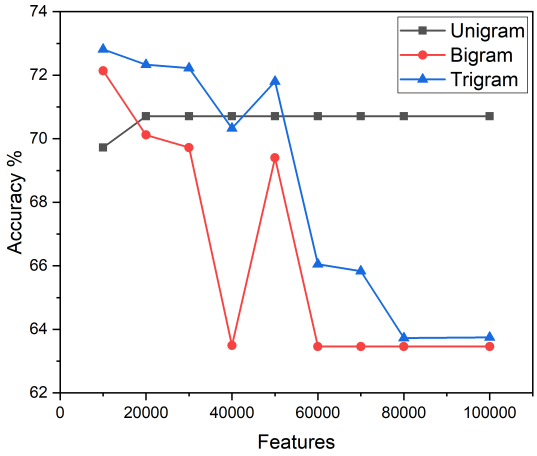

Fig. 6. Classification accuracy of different n-gram models with Logistic Regression

In order, to learn the baseline accuracy for classification, the several n-gram models are developed. The classification accuracy by logistic regression in conjunction with n-gram based feature extraction is presented in Figure 6. The plot depicts classification accuracy for uni-gram, bi-gram and tri-gram models evaluated w.r.t. feature vector size. The best accuracy is attained by trigram model for a feature vector size of 30,000 achieving an accuracy of 72.2%.

The dataset was also evaluated using Doc2Vec [8] models namely Distributed Bag of Words(DBOW), Distributed Memory (PV-DMM) Mean, Distributed Concatenated (PV-DMC). The respective accuracies obtained are shown in Table 1. However, the accuracy obtained are inferior as the tweets can barely be considered as documents.

Table 1. Average classification accuracy for Doc2Vec models trained on different n-gram features

	Uni-gram	Bi-gram	Tri-gram
DBOW	66.78%	67.4%	67.91%
PV-DMM	64.29%	64.98%	64.98%
PV-DMC	63.46%	65.6%	66.36%

The Doc2Vec model extends the idea of Word2Vec [8]. As the dataset contains tweets which are analogous to sentences, a Word2Vec model can learn representations of tweets at the word-level. Further, a 2 layer artificial neural network (ANN) is trained on Word2Vec embeddings. The models have three variants as shown in Figure 7. First of all, the neural network model is trained on pre-trained Word2Vec embedding, restricting re-training of Word2Vec embedding. Secondly, the model is started from Word2Vec model, but it is allowed to fine tune embedding values to achieve higher classification accuracy. Finally the ANN model was allowed to learn embeddings from scratch. We observe that Fine-tune Word2Vec model achieves superior test accuracy of 84.82%. Moreover, learning embeddings from scratch has a greater tendency to overfit the training data.

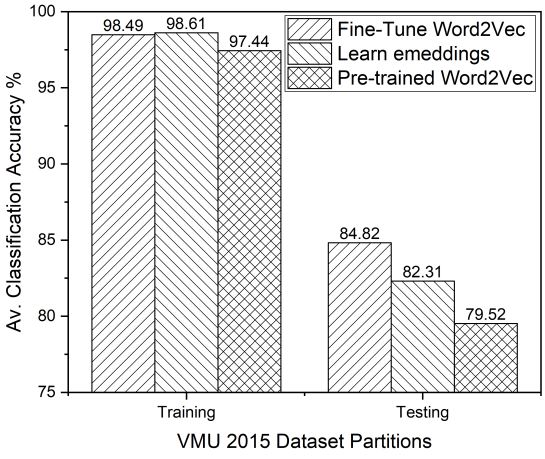

Fig. 7. Classification performance of ANN with Word2Vec on VMU 2015 dataset

Finally, a CNN is trained on the VMU 2015 dataset to achieve an average accuracy of 87.43% which is better than all the models presented in the study. Although, CNNs are primarily used for image domains, the proposed methodology makes use of 1-dimensional convolutions in order to learn a CNN for text data. The three convolutional layers uses kernels of size 2×1, 3×1 and 4×1 to simulate n-gram models in a CNN architecture.

The comparative performance of the proposed CNN architecture is presented in Figure 8. The proposed CNN achieves better accuracy than all other approaches achieving 87.43%. The proposed methodology is also compared by UoS-ITI [10] which use a semi automatic approach of tokenization, POS tagging, named entity recognition and relational extraction through regex patterns. The proposed approach achieves a superior accuracy as it aims to learn the latent features which discriminate a fake tweet from a real tweet. Most of the techniques used for text classification are based on handcrafted features such n-gram feature extraction. However, neural networks combine feature extraction and classification in a single algorithm eliminating human biases. This leads to better accuracy as evident by ANN and CNN models which achieve 84.82% and 87.43% respectively.

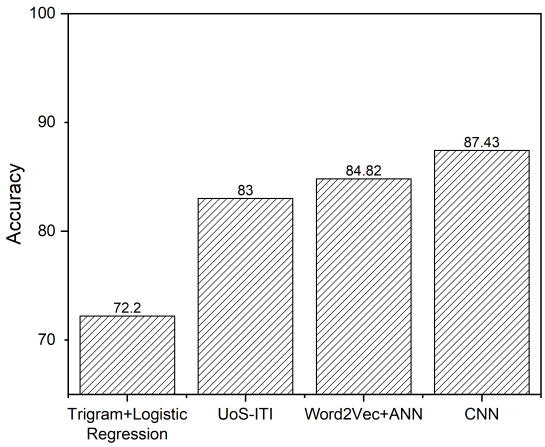

Fig. 8. Comparative Classification performance of proposed methodology on VMU 2015 dataset

6 Conclusion

In this work, a practical and efficient deep learning approach is presented which could discriminate misleading and credible tweets by representing text in n-dimensional vector spaces. The approach achieves an acceptable accuracy of 87.43% which is superior to traditional handcrafted techniques. It establishes that 1-dimensional CNNs can be promising in text classification. The high accuracy mends from the fact that CNN is effective in learning a latent representation of the text embedding data by combining feature extraction and classificationn in a single pipeline, pivoted for accurate classification.

References

1. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on twitter. In: Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). vol. 6, p. 12 (2010)
2. Castillo, C., Donato, D., Gionis, A., Murdock, V., Silvestri, F.: Know your neighbors: Web spam detection using the web topology. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 423–430. ACM (2007)
3. Cheong, F., Cheong, C.: Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods. PACIS 11, 46–46 (2011)
4. Detection, visualization of misleading content on Twitter: Boididou, christina and papadopoulos, symeon and zampoglou, markos and apostolidis, lazaros and papadopoulou, olga and kompatsiaris, yiannis. International Journal of Multimedia Information Retrieval 7(1), 71–86 (2018)
5. Gupta, A., Kumaraguru, P.: @ twitter credibility ranking of tweets on events# breakingnews. Tech. rep. (2012)
6. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on World Wide Web. pp. 729–736. ACM (2013)
7. Gyongyi, Z., Garcia-Molina, H.: Web spam taxonomy. In: First international workshop on adversarial information retrieval on the web (AIRWeb 2005) (2005)
8. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196 (2014)
9. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: Can we trust what we rt? In: Proceedings of the first workshop on social media analytics. pp. 71–79. ACM (2010)
10. Middleton, S.: Extracting attributed verification and debunking reports from social media: mediaeval-2015 trust and credibility analysis of image and video (2015)
11. Seo, E., Mohapatra, P., Abdelzaher, T.: Identifying rumors and their sources in social networks. In: Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III. vol. 8389, p. 83891I. International Society for Optics and Photonics (2012)
12. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th annual computer security applications conference. pp. 1–9. ACM (2010)