Truth be Told: Fake News Detection Using User Reactions on Reddit

Vinay Setty
University of Stavanger, Norway
vsetty@acm.org

Erlend Rekve
University of Stavanger, Norway
erlend.rekve@stud.uis.no

ABSTRACT

In this paper, we provide a large dataset for fake news detection using social media comments. The dataset consists of 12,597 claims (of which 63% are labelled as fake) from four different sources (Snopes, Poltifact, Emergent and Twitter). The novel part of the dataset is that it also includes over 662K social media discussion comments related to these claims from Reddit. We make this dataset public for the research community. In addition, for the task of fake news detection using social media comments, we provide a simple but strong baseline solution deep neural network model which beats several solutions in the literature.

CCS CONCEPTS

• **Information systems** → **Clustering and classification**.

KEYWORDS

Fake news detection; Reddit comments; Deep neural networks

ACM Reference Format:
Vinay Setty and Erlend Rekve. 2020. Truth be Told: Fake News Detection Using User Reactions on Reddit. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3417463

1 INTRODUCTION

Online news sites and social networks have become a major source of news, information and knowledge for the vast majority of people. Unfortunately, these platforms also help misinformation and false claims to spread faster and deeper in social networks [13]. To address this issue, many popular social networks and news aggregators such as Facebook and Google news are leveraging either crowdsourcing or fact checking services such as snopes.com, politicat.com and emergent.info. While these solutions are effective and important, they rely on expert analysis and manual effort inhibiting timely detection and limiting the scale.

Following this trend, many automated claim verification and fake news detection techniques have been proposed in the literature which rely on linguistic features, source domain and speaker information [6–10]. However, according to some studies, misinformation is very difficult to detect even for critical human readers [1]. Moreover, it is not hard to write high quality news articles conveying false facts. For example, a professionally written news article claiming that "The Queen was threatening to abdicate should Britain leave the EU" was published in "Yournewswire.com" and shared in Facebook over 23,000 times even though it is a false claim[1]. Since these articles are written by humans, it is hard to come up with a specific set of features to verify their truthfulness. Another widely used feature by these methods are source-based attributes of the news such as domain name, page rank and author of the claim. While these features boost the recall of detecting false claims, they result in poor precision since all news articles from a specific website tend to be classified as true or false. Moreover, depending on the way these features are represented, it might introduce a bias for the downstream tasks. For example, a news article from forbes.com which has a high reputation and page rank reports that "That Scientific Global Warming Consensus...Not!"[2] which is debunked by politifact.com. Moreover, it is not difficult to mask the source of the article using blogs in reliable domains such as "wordpress.com". Finally, combining many features including social media comments using sophisticated models such as Conditional Random Fields (CRFs) have also been proposed [9].

Other works consider temporal patterns of the response received for news articles and model them using deep neural networks such as LSTMs [12, 14]. However, these works do not consider the news article contents rather only focus on the textual content of the reactions. Models to jointly consider the textual content of the news article as well as the textual social media reactions to detect the false claims are limited in the literature. Another challenge is the lack of social media data to conduct research.

In this paper, we make following contributions:

(1) We create a large-scale dataset with social media comments for the task of fake news detection and make it available for public use.
(2) We propose neural network models to represent textual content of news articles and social media comments jointly to detect false claims.
(3) Our solution purely depends on textual content rather than any handcrafted features or domain information.

2 RELATED WORK

There are many works which consider manually engineered linguistic and source-based features [2, 7–9]. More recent works propose

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3417463

[1] http://www.bbc.com/news/av/world-us-canada-38794905/fake-news-this-is-a-war-on-alternative-media
[2] https://www.forbes.com/sites/larrybell/2012/07/17/that-scientific-global-warming-consensus-not/#2d60d3b83bb3

deep neural networks with attention mechanims [6, 10]. As we showed in our experiments, news article contents alone are not always sufficient and none of these approaches consider additional information such as social media discussions to detect false claims.

For modelling social media, existing works in literature manually hand-craft features such as Facebook likes, number of shares, and user demographics to train classifiers for detecting rumors [5, 15]. Modeling rumor cascades in social networks as Recurrent Neural Networks (RNNs) have been proposed recently [4, 14]. Hybrid models which consider the temporal sequence of textual content of social media comments have also been proposed [12]. However, these models do not consider the news article text and social media comments holistically. Reddit data is different from existing datasets with subreddit information and user discussion threads.

3 PROBLEM DESCRIPTION

Given a claim c in textual format, our goal is to automatically detect whether it is 'true' or 'false'. For this purpose, we consider news articles N_c which mention the claim and corresponding social media response S_c they received online. For example, given the claim "Obama caught forging his birth certificate", we have a corresponding news article reporting the false claim "Video evidence of Barack Obama admitting forging his birth certificate" and discussion thread of reactions it received in Reddit. While many in online forums tend to believe this article, some reddit comments disagree and provide proof for it. For example, one of the reddit comments says "Sorry but this is just batshit Alex Jones crazytown level stupid.". Our goal is to model a neural network which only relies on textual content of the news articles and the online reactions.

4 PROPOSED MODEL

In this section we give the details of the models we propose. As shown in Figure 1, the model consists of two main parts: (1) *News Article Model:* The content of the news articles are represented using two Convolutional Neural Networks (CNNs) for the claim text and text of the news articles. (2) *Social Media Model:* Since social media comments have a different vocabulary and document structure, we represent them by training a vector using Doc2Vec [3]. Finally, before classifying the claim as true or false using a softmax classifier, they are concatenated and passed onto dense layers.

4.1 News Article Model

Our first model is responsible for learning a representation of the content of the news articles. The sentences in the news articles are represented as a sequence of k-dimensional vectors $x_i \in \mathcal{R}^k$ of length n. The sequences are padded if the sentences are of length less than n. The weights for these vectors are initialized using the weights from the corresponding pre-trained word vectors (e.g, word2vec or Glove). Any missing words are initialized with random weights. These sequences are fed into a fully connected embeddings layer so that their weights are also trainable. Concatenating the n word vectors of k dimensions, forms a $n \times k$ matrix as an input. Then further features are generated using 1-D convolution operations and are performed using the filter activation functions such as "ReLU" on a window of fixed number of words.

Table 1: detailed statistics of dataset

Labeled by	Articles	Claims	Reddit comments	False claims
Snopes	36,271	3,096	352,708	73%
Politifact	11,019	9,453	289,424	80%
Emergent	250	48	20,073	36%
Total	47,540	12,597	662,205	63 %

$$c_i = f\left(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b\right)$$

We train two independent CNN models; one for the claim text and one for the body of the news article respectively (Fig. 1). The weights from these two models are concatenated and jointly trained.

4.2 Social Media Discussion Model

Since social media messages are known to often contain informal language (slang) and emojis, we train an independent model (see Figure 1 for an example comment). Instead of relying on pre-trained word vectors, we train a Doc2Vec model for the social media comments. Each social media comment is represented using a vector of d dimensions. More formally, given a sequence of context words $w_1, w_2, ..w_{k-1}, w_{k+1}, .., w_T$ around a word w_k, from social media comments, our goal is to maximize the average log probability of predicting the word w_k:

$$\max \frac{1}{T} \sum_{i=k}^{T-k} logP(w_k | w_{i-k}, ..., w_{i+k})$$

In addition, the quality and reliability of the comment also relies on the forum in which it is published. For example, comments from specific sub-reddits such as "The Donald" have poor reliability. Therefore, we also include an additional feature representing the comment's subreddit as a one-hot encoded vector.

5 EXPERIMENTAL EVALUATION

Now we describe the datasets along with Reddit comments we crawled. The statistics of the datasets are summarized in Table 1.

Rumor Data: This dataset contains claims labeled by snopes.com, politifact.com, and emergent.com. To get relevant articles, similar to [9] we use a search engine to search for articles related to a claim. We also included the search results from [9] into our own data. This method results in a lot of irrelevant data, i.e., articles that do not relate to the claim searched after. To reduce the noise, we filtered out weblinks from well known fact-checking websites and popular social media networks. In addition, we used a pre-trained stance detection [11] to filter out articles not relevant to a claim. This model was created during the fake news challange[3] to determine if an article is unrelated, for, against or discussing its headline.

Reddit Data: This dataset contains comments from the popular news aggregation forum Reddit. Where users can share and discuss almost any topic, a user can create a post that includes a title and a link to other web content. Other users can then make comments and discuss the various topics that are linked. Posts are organized into "Subreddits" which are smaller communities that discuss topics

[3] http://fakenewschallange.org

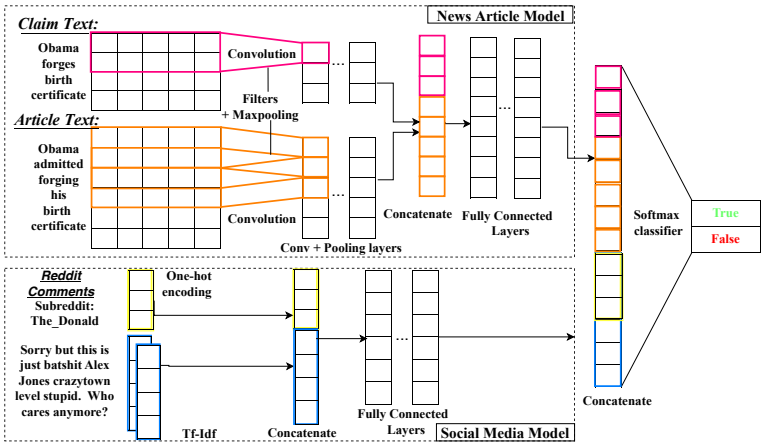
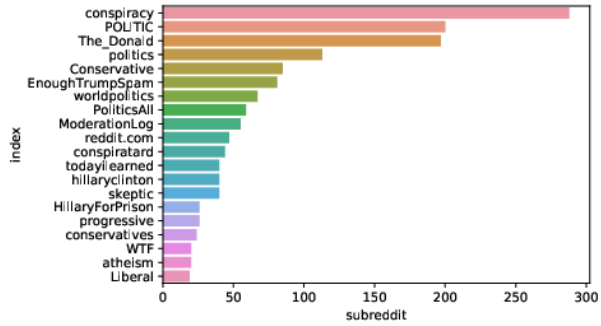
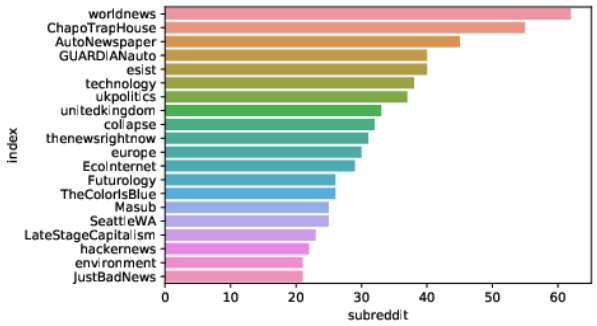


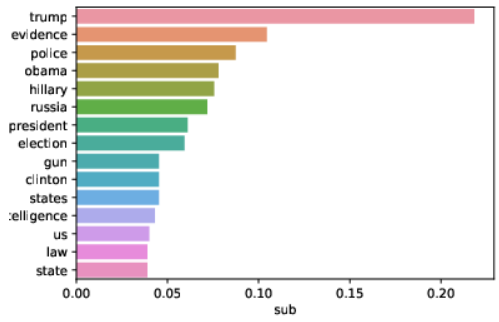
Figure 1: Model detecting false claims. News article model uses CNNs to combine claim and article and Doc2Vec to represent comments



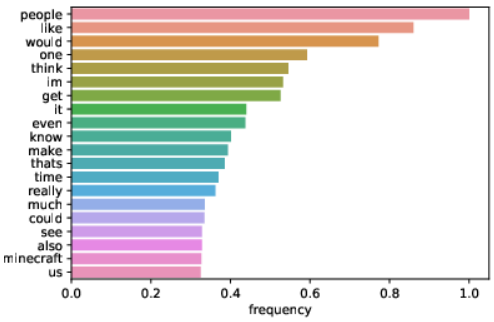
(a) Top sub-reddits for false news



(b) Top sub-reddits for true news



(c) Word distribution for false news



(d) Word distribution for true news

Figure 2: Reddit dataset analysis

of interest. One such forum is "r/news" which discusses current news headlines. To collect comments discussing articles from the above datasets we used the official Reddit API and queried for posts discussing articles in our dataset. In order to avoid any contamination of truthfulness knowledge from fact-checking sites we remove comments that reference the fact-checking websites. Moreover, we also omit the sub-tree of comments which appear after the debuking comments referencing fact-checking websites. In Figure 2 we show the distribution of subreddits and words for the fake and real news. The subreddit and vocabulary distribution information can be potentially be used in future fake news detection tasks. Our dataset can be downloaded from the popular service kaggle[4].

All the datasets are then merged and used as training data for our models. All the claims are linked to multiple articles and associated Reddit comments. As seen in Table 1 the dataset is unbalanced

[4]https://www.kaggle.com/deepnews/fakenews-reddit-comments/

Table 2: Evaluation of false claim detection with baseline comparison and ablation test. All results in bold are statistically significant ($p < 0.05$)

Approach	Accuracy		False Claims			True Claims			AUC
	Micro	Macro	Prec.	Rec.	F1	Prec.	Rec.	F1	
SVM	0.793	0.796	0.739	0.748	0.791	0.754	0.843	0.796	0.884
LSTM	0.735	0.726	0.690	0.734	0.762	0.829	0.558	0.667	0.835
CRF [9]	0.618	0.610	0.630	0.740	0.680	0.594	0.460	0.520	0.611
Reddit Doc2Vec	0.657	0.647	0.651	0.740	0.692	0.667	0.567	0.613	0.686
News Article CNN	0.802	0.798	0.777	0.873	0.822	**0.838**	0.724	0.777	0.885
Joint (News Article + Reddit)	**0.826**	**0.825**	**0.837**	**0.895**	**0.866**	0.814	**0.817**	**0.812**	**0.909**

Table 3: Confidence scores of example claim

News Article	Reddit	Joint	CRF	Label
18% False	99% False	58% False	71 % False	False

towards fake claims and is a consequence of the fact-checking services nature to research claims that seem likely to be false.

5.1 Setup

Our model is implemented on Keras with tensorflow. To tune the model's hyperparameters we employed grid search. After the hyperpameter tuning we chose dropout of 0.5, 64 hidden layers and filter size 300 and three kernels of size 2, 4 and 5. As baselines we use simple SVM with tf-idf features, LSTM, and a Conditional Random Field (CRF) model [9]. Then we evaluate our models independently with Reddit comments, news article model and the joint model.

5.2 Results and Discussion

To combine the news article model and social media model into one, the output of the penultimate layer of each model is concatenated and fed into a series of dense hidden layers, before going into a softmax classifier. As a baseline we use the CRF model presented in [9], with premission from the author we ran a pre-trained model on our data. It can be seen from Table 2 that there is a performance gain by concatenating the two models, a reason for it being so small is that we do not have Reddit comments from all the articles either by there not existing or limitations of the Reddit API. If we take away all the articles not containing Reddit comments, the performance of the joint model over the news article model increases by 2.5%. We can also see that the joined model outperforms the CRF model, and proves that neural networks can be an essential part of detecting false claims on the web. [9] reports an macro-accuracy of 80%, a reason for the CRF model preforming badly in our experiments is that we only tested on one fold of their trained models.

5.3 Anectoral Example

To see how the model is evaluating a specific claim: "The European Scientific Journal, a peer-reviewed academic publication, concluded that the collapse of the Twin Towers and World Trade Center Building 7 on 11 September 2001 was the result of a controlled demolition".[5]. The Table 3 shows the confidence scores of this claim form the various models. It can be seen that news article model labels this claim as a true. The social media model labels based on the comments from Reddit forums this claim as false.

[5]https://www.snopes.com/fact-check/journal-endorses-911-conspiracy-theory/

6 CONCLUSIONS AND FUTURE WORK

Through this paper we release a large-scale dataset for fake news detection using the popular social media forum Reddit. We crawled the Reddit comments for the labeled data from several fact checking sources. We also presented a simple neural network model to jointly consider the claim text, news article body and the social media comments to classify fake news. We show that our model is simple and yet outperforms several strong baselines in the literature. In future work, we will do more analysis of the Reddit dataset and exploit network features to improve fake news detection.

REFERENCES

[1] www.npr.org/sections/thetwo-way/2016/11/23/503129818/. [Online; accessed 21-Oct-2019].
[2] Benjamin D Horne and Sibel Adali. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*, 2017.
[3] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014.
[4] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, pages 3818–3824, 2016.
[5] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *CIKM*, pages 1751–1754, 2015.
[6] Rahul Mishra and Vinay Setty. SADHAN: hierarchical attention networks to learn latent aspect embeddings for fake news detection. In *ICTIR*, pages 197–204, 2019.
[7] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *CoRR*, abs/1708.07104, 2017.
[8] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *CIKM*, pages 2173–2178, 2016.
[9] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *WWW*, pages 1003–1012, 2017.
[10] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *EMNLP*, pages 22–32, 2018.
[11] Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR*, abs/1707.03264, 2017.
[12] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A hybrid deep model for fake news detection. In *CIKM*, pages 797–806, 2017.
[13] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
[14] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *WSDM*, pages 637–645, 2018.
[15] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *WWW*, pages 1395–1405, 2015.