Towards an automated system to detect disinformation in political media

1 Introduction

Identifying disinformation has been a challenge although there are many methods to check the spread of disinformation. The disinformation has been growing rapidly due to the heavy use of social media and high level of social engagement. Social media owners are finding ways to tackle the spread of disinformation online due to its spread and global impact. Existing research has focuses on two broad categories of disinformation- opinion-based which includes fake reviews in online platforms and fact-based which include fake news. The scope of this project is to develop methods and to experiment with existing methods to identify fake news. There are open source browser extensions for real time detection of misinformation. But the performance of such automatic detection engines have not been reliable due to the real challenge in defining what fake news is and also the limited availability of annotated data with gold standard labels. Human labeling also has an impact/bias of the misinformation because they filter the information they read (For example, people read/watch what they like and ignore what they don't).

2 Related Work

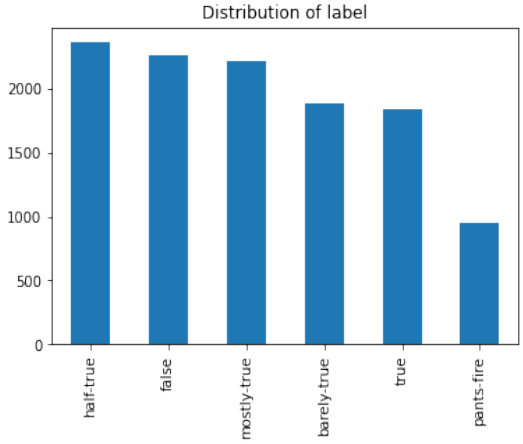
[1] and [2] develop methods to detect misinformation on Twitter, by identification of suspicious and malicious patterns by using supervised and unsupervised learning techniques. [3] aim to understand the reliability of the news by using surface level linguistic based methods that identify patterns in the writing style. Other research has based analysis by building real time analyzers to crawl data from social media sources such as twitter to collect the tweets regarding a claim to identify if it is true. [4] reviews how a user would perceive if a social media post is real or fake. They include fake reviews in e-commerce platforms, hoaxes on collaborative platforms, and fake news in social media. While a lot of research is focused on twitter [5] aims to identify features based on topic that help in early detection of misinformation. focused on personal messaging platforms such as WhatsApp. Research on Fact-checking WhatsApp rumors suggests that raising a signal of doubt on a claim can suppress the effect of its spread[6]. An interesting system used in social media to reduce circulation of misinformation is to rate the news sources based on their authenticity[7]. For instance, in reddit's Manchester united football club's fan community, they list 'tiers' for various news sources that report transfer news during the transfer season.

3 Dataset

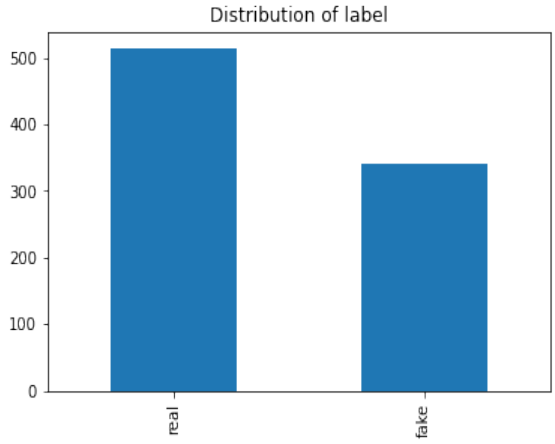
Existing benchmark datasets are generated by current systems such as non-partisan fact checking websites and heuristic based web browser extensions. We use 2 benchmark datasets for our analysis. **LIAR** [8], which consists of 12,836 human-labeled short statements, which are sampled from various

contexts from Politifact.com. Besides the short statements the dataset also contains meta data. The data dictionary is presented in table 1.

FakeNewsNet [9][10][11] The data consists of 854 News Articles from Politifact with lables 'real' and 'fake'. , The breakdown of 'Real' and 'Fake' labels are 513 and 341 respectively with a total vocabulary size of 16951 words



(a) distribution of label: Liar dataset



(b) distribution of label: FakeNewsNet dataset

Politifact.com provides detailed analysis report and links to source documents for each case. The labels for news truthfulness are 6 fine-grained classes: pants-fire, false, barely-true, half-true, mostly true, and true. The class description is presented below. The logic behind the labels is determined by truth-o-meter, a simple human annotated fact-checking process explained in [12]. The goal of the Truth-O-Meter is to reflect the relative accuracy of a statement. The category description is as below:

- *True* – The statement is accurate and there's nothing significant missing.
- *Mostly True* – The statement is accurate but needs clarification or additional information.
- *Half True*– The statement is partially accurate but leaves out important details or takes things out of context.
- *Mostly False* – The statement contains an element of truth but ignores critical facts that would give a different impression.
- *False* – The statement is not accurate.
- *Pants on Fire* – The statement is not accurate and makes a ridiculous claim.

We have also identified other annotated datasets to test the model for generalization. Each of the datasets come from different sources. We describe each of them below.

- **BSDetector dataset on Kaggle:** This has text and metadata scraped from 244 websites tagged as "bullshit" by the BS Detector, a Chrome Extension by Daniel Sieradski. Data presents 12,999 posts in total pulled using the webhose.io API. The label for each website is the output of the BS detector that includes labels such as 'bias', 'conspiracy', 'fake', 'hate', 'satire', rather than human annotations. Data sources that were missing a label were assigned a label of "bs".

- **Buzzfeed dataset** - Given an article, few hyper partisan pages on facebook share that article with a new title which derails from the actual discussion in the original article misinforming the reader.

4 Evaluation

4.1 experimental set-up

There are quite a few challenges when analysing the the short statements in LIAR. The statements are in isolation (without much of a context) and have limited surface level linguistics for any model to understand. The detailed description for the experiments is as follows. We tokenise the text data using the BERT [13] tokenizer. We build a baseline using a variety of representations modeling them with logistic regression and Naive Bayes classifiers. In the case of LIAR's logistic Regression model, we add the L2 regularization and solve using the Liblinear solver while for FakeNewsNet, we build a baseline model using logistic Regression with with L2 regularization and SAGA solver. For testing out the neural architecture based methods, we embed the input using the 300- dimension Google News word2vec representation.We add a textual entailment mechanism where we query the short statement using the Google search API and compare the short statement with the web page for overlaps. We measure such overlaps using the BLEU score. We use a feed-forward neural network to input the meta-data and the BLEU score into the neural architecture. We experimented with various neural architectures and present the experimental set-up in table 1. We will now discuss the results of the experiments.

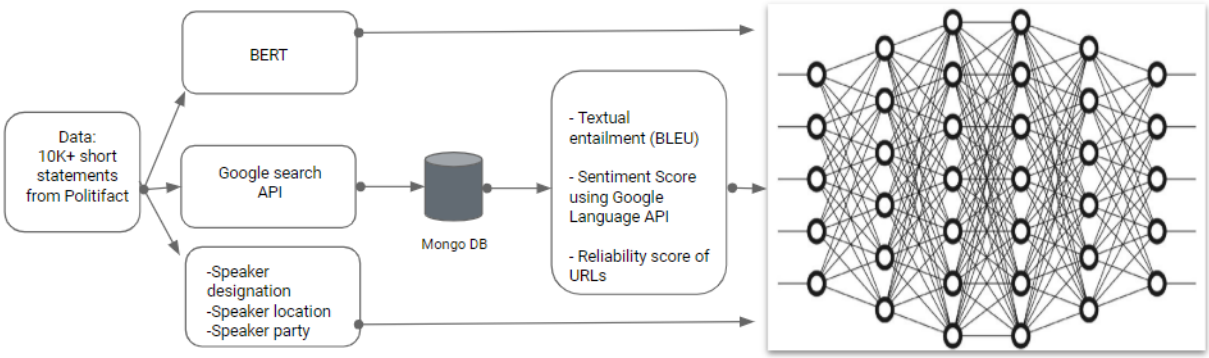


Figure 2: Proposed Architecture

Table 1: Experimental set-up

RNN, Bi RNN LSTM, Bi LSTM GRU, Bi GRU	Embedding	Google news word2vec 300dim
	Hidden vector size	2 x 200
	Batch size	200
	epochs	25
	Learning rate	0.001
	Activation	Sigmoid
	Optimizer	Adam
	Loss	Cross entropy
Feed forward Neural Network	Hidden vector size	20
	Batch size	50
	epochs	50
	Learning rate	0.001
	Activation	Sigmoid
	Optimizer	Adam
	Loss	Cross entropy

4.2 Discussion and Results

The reuslts for the models discussed are presented in Table 2.

Table 2: Performance results comparision

Dataset	Representation	Model	Validation Accuracy
LIAR	TF-IDF	Logistic Regression	**0.270**
	TF-IDF ngram(n=1,2)	Logistic Regression	0.267
	Binary	Logistic Regression	0.260
	TF-IDF	Naive Bayes	0.266
	TF-IDF ngram(n=1,2)	Naive Bayes	0.256
	TF-IDF	Neural Network	0.205
	word2vec	RNN	0.197
	word2vec	Bi-RNN	0.192
	word2vec	LSTM	0.192
	word2vec	Bi-LSTM	0.218
	word2vec	GRU	0.192
	word2vec	Bi-GRU	0.193
FakeNewsNet	TF-IDF	Logistic Regression	0.830
	TF-IDF ngram(n=1,2)	Logistic Regression	0.784
	Binary	Logistic Regression	0.918
	TF-IDF	Naive Bayes	0.632
	TF-IDF ngram(n=1,2)	Naive Bayes	0.591
	TF-IDF	Neural Network	**0.889**

We look at the misclassified samples for each class based on the above results. Since this is a multi class classification, we looked at one model's confusion matrix.

Following is the count of misclassified items per class by all the above models and one misclassified example per each label. The final test accuracy for LIAR dataset with Bidirectional LSTM was found to be **0.225**. In the case of FakeNewsNet the test accuracy using a feed forward neural network was found to be **0.825**

4.3 Error Analysis

LIAR dataset We evaluate the statements misclassified by Logistic regression and Bi-LSTM since they show highest performance.

- 1045 out of 1283 statements were misclassified by both the models : Bi LSTM and Logistic Regression

Table 3: Error analysis

Sno	Statement	Ground truth	Bi LSTM	Logistic Regression
1	Says 21,000 Wisconsin residents got jobs in 2011, but 18,000 of them were in other states.	false	mostly-true	mostly-true
2	Mitt Romney has proposed cutting his own taxes while raising them on 18 million working families.	mostly-true	half-true	half-true

- 219 statements were classified correctly by Bi LSTM but misclassified by Logistic Regression.
- For the first sentence above, ("Says 21,000 Wisconsin residents...") the actual fact was that that such a comparison does not exist since the numbers to compare were an apples to orange comparison as described from the politifact website below. *"the two figures are apples and oranges – you can't simply subtract the smaller number from the larger to determine the number of newly employed Wisconsin residents who got jobs out of the state."*

Table 4: Error analysis: Logistic Regression

Sno	Statement	Ground truth	Logistic Regression
1	The recent process of awarding $3 billion worth of airport vending contracts was the most open and transparent procurement process in the citys history.	half-true	true
2	On attacks by Republicans that various programs in the economic stimulus plan are not stimulative, "If you add all that stuff up, it accounts for less " than 1 percent of the overall package."	half-true	barely-true

- 275 statements were classified correctly by LR but misclassified by Bi LSTM

Table 5: Error analysis: Bi LSTM

Sno	Statement	Ground truth	Bi LSTM
1	Illegal immigration wasnt a subject that was on anybodys mind until I brought it up at my announcement.	false	half-true
2	Obama says his health care plan is "universal."	barely-true	false

FakeNewsNet We evaluate the statements misclassified by Feed forward neural network since it shows highest performance.

- 17 out of 171 statements which are real are misclassified as fake
- 3 out of 171 statements which are fake are misclassified as real

Below are the examples for misclassified statements by Feedforward NeuralNetwork model:

Table 6: Error analysis: Logistic Regression

Sno	Statement	Ground truth	Prediction
1	the state of the union 2012 we can either settle for a country where a shrinking number of people do really well while a growing number of americans barely get by ,or we can restore an economy where everyone gets a fair shot , and everyone does their fair share, and everyone plays by the same set of rules. what's at stake aren't democratic values or republican values , but american values . and we have to reclaim them.	Real	Fake
2	the vatican is under pressure to let more of its employees work from home after several offices remained open even after italy shut down all . . .	Fake	Real

Observation: Articles that were misclassified were found to be abnormally long.

5 Conclusion and future work

The results that we present are a proof of concept for designing better performing models. With more extensive experimentation it is possible to see an improvement in performance. Another direction could be to use a step-wise architecture where we combine the true classes and the false classes separately and first predict whether a given short statement is True or False and then move on to predict the finer classes (half-true, mostly-true etc).

References

[1] Julio Reis et al. "Supervised Learning for Fake News Detection". In: *IEEE Intelligent Systems* 34 (Mar. 2019), pp. 76–81. DOI: 10.1109/MIS.2019.2899143.

[2] Shuo Yang et al. "Unsupervised Fake News Detection on Social Media: A Generative Approach". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 5644–5651. DOI: 10.1609/aaai.v33i01.33015644.

[3] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. "Detecting Hoaxes, Frauds, and Deception in Writing Style Online". In: *Proceedings - IEEE Symposium on Security and Privacy* (May 2012), pp. 461–475. DOI: 10.1109/SP.2012.34.

[4] Srijan Kumar and Neil Shah. "False Information on Web and Social Media: A Survey". In: *CoRR* abs/1804.08559 (2018). arXiv: 1804.08559. URL: http://arxiv.org/abs/1804.08559.

[5] Michela Del Vicario et al. "Polarization and Fake News: Early Warning of Potential Misinformation Targets". In: *CoRR* abs/1802.01400 (2018). arXiv: 1802.01400. URL: http://arxiv.org/abs/1802.01400.

[6] Sumitra Badrinathan and Simon Chauchard. *Is there a way to counter fake news on WhatsApp?* https://www.hindustantimes.com/analysis/is-there-a-way-to-counter-fake-news-on-whatsapp/story-iC0Z5CG5ESy2YC9Q2SJlMI.html. [Online; Updated: Jan 28, 2020]. 2020.

[7] Diego Esteves et al. "Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web". In: *CoRR* abs/1809.00494 (2018). arXiv: 1809.00494. URL: http://arxiv.org/abs/1809.00494.

[8] William Yang Wang. ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: https://www.aclweb.org/anthology/P17-2067.

[9] Kai Shu et al. "FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media". In: *arXiv preprint arXiv:1809.01286* (2018).

[10] Kai Shu et al. "Fake News Detection on Social Media: A Data Mining Perspective". In: *ACM SIGKDD Explorations Newsletter* 19.1 (2017), pp. 22–36.

[11] Kai Shu, Suhang Wang, and Huan Liu. "Exploiting Tri-Relationship for Fake News Detection". In: *arXiv preprint arXiv:1712.07709* (2017).

[12] Angie Drobnic Holan. *The Principles of the Truth-O-Meter: PolitiFact's methodology for independent fact-checking*. https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/#Truth-O-Meter\%20ratings. 2018.

[13] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.