Stance Detection & Fake News NLG Proposal

Bernice Brown **Brian Chen** **Tahya Weiss-Gibbons** **Nicholas Kobald**

1 Introduction

The term *Fake News* has gained popularity following the 2016 United States presidential election and the vote for the United Kingdom to exit the European Union (Rose, 2017)(Kucharski, 2016). Fake News refers to articles that meet poor journalistic standards, and contain incorrect or misleading information. It's suggested that these articles, and their tendency to be shared on social media had discernible effect on the events of the USA election, and Brexit (Allcott and Gentzkow, 2017).

Determining whether or not a news article is fake is difficult. A Stanford study shows students from middle school through college have difficulty distinguishing real news articles from advertisements (Wineburg et al., 2016). As a a result, attempts have been made to automate the detection of fake or misleading news articles (Conroy et al., 2015)

The purpose of this research is to apply natural language processing and machine learning techniques to analyzing the validity of news articles. In particular, we will begin by following the outline presented by Fake News Challenge (Fake News Challenge, 2017).

The *stance* of a text is the attitude it expresses towards a particular target (Augenstein et al., 2016). The first step of the Fake News challenge is to categorize the stance of the body against the stance of the heading of the article. The Fake News Challenge organization provides an implementation, which we will use as our baseline.

2 Previous Work

2.1 Data Requirements

Rubin et. al. showed data used to investigate rumors and deception need to have the following characteristics. There must be both truthful and deceptive news within the data set, the format must be accessible, the data must be verifiable, and there must exist data points of comparable lengths and writing styles(Rubin et al., 2015).

2.2 Stance Detection

In Emergent: a Novel Data Set for Stance Classification (Ferreira and Vlachos, 2016), stance detection was used to classify claims in relation to news articles in the Emergent data set. Emergent is a data-set derived from a digital journalism project for rumour debunking. Consisting of 300 claims and 2,595 associated news articles, the Emergent project contains labelled data that can be used in a variety of NLP tasks. (Ferreira and Vlachos, 2016). Headlines could be classified as either for, against or observing a claim, where observing a claim merely mentions the claim without giving a stance.

A headline of a news article and an associated claim were considered. Certain features were extracted, considering first the headline alone and then the headline and claim together. Headline features were extracted using the bag-of-words representation, as well as whether the headline ended in a question mark. The idea of the bag-of-words representation is to quantize each extracted key point into one of visual words, and then represent each image by a histogram of the visual words. For this purpose, a clustering algorithm (e.g., K-means), is generally used for generating the visual words /citezhang2010understanding. Features for the distance from the root word to

any refuting words (e.g deny) and any reporting words (e.g claim, presumably) were also added.

For considering the headline and the claim together, their approach involved constructing a graph, such that each word in the headline and the claim was vertex. For every word pairing between the claim and the headline an edge was created and assigned a score. If the stems of the words are identical, it is given the max score. If the words are paraphrases, which is determined using the Paraphrase Database (PPDB) (Ferreira and Vlachos, 2016), they are given the maximum paraphrase score. If neither they are given the minimum score. The Kuhn-Munkres algorithm was run on the graph, which found the maximum scoring word alignment (Ferreira and Vlachos, 2016). Features were added for negating words and subject-to-verb objects. Vectors were used for comparing the claims to headlines, looking at the cosine similarity of the subject-to-verb objects.

Using the headline and headline-claim methods, two overlap thresholds were defined, minimum for and maximum against. If the overlap is higher than the minimum for threshold it is labelled for, if it is less than the maximum against it is labelled against. If it hits neither threshold it is labelled observing. An accuracy of 73% was achieved on the Emergent test data set using this method. This method was challenged though in detecting observing stances, due to the similarities between the headlines and the claims when a headline is observing. This lead to a mislabelling of the observing claims as for claims.

Other work done on stance detection includes Stance Detection with Bidirectional Conditional Encoding (Augenstein et al., 2016). This paper looked at the stances of tweets, given sparse training data or where the target is not explicitly mentioned in the text. Two baselines were used, in a manner of treating stance detection as sentence level sentiment analysis. One was implemented using a Support Vector Machine Classifier and the other with long-short term memory networks (LSTM). LSTM networks were first proposed by Hochreiter and Schmidhuber and it is a gradient based method of learning to store information over extended time intervals (**?**). Their paper found LSTM networks to work best and were used for most of the encodings.

Initially, the text and targets were independently encoded as a k-dimensional dense vector space, using two different LSTM networks. The model learned target-independent representation for the tweets, with the initial training encoding the tweets independant of any target. It relied on the nonlinear projection layer to incorporate the target in the stance prediction (Augenstein et al., 2016).

The paper also tested a conditional encoding method. First the target was encoded as a fixed length vector using one LSTM network. The tweet was then encoded using another LSTM network with its initial state using a representation of the target. This encoding was adapted to use bidirectional conditional encoding. One encoding was achieved by reading the target and the tweet from left-to-right, and then another by reading them from right-to-left. This allowed for the context on either side to be considered, in target dependant encoding method.

To deal with the small amount of training data, unsupervised pretraining was used, by initialized word embeddings used in the LSTM with a trained word vector model. These embeddings were only used for initialization and were optimized with further training. The paper found that conditional encoding was well suited for learning how to fit a targets with generalized encodings and that bidirectional encoding performed best overall, especially where the target was not explicitly mentioned in the tweet.

3 Deception Detection

Conroy et. al. discusses several approaches for automatic fake news detection using natural language processing. This is described as categorizing news on a spectrum based on their level of certainty as well as their veracity (the intention to mislead). The automatic detection of fake news is centered around predicting the chances that any news item is intentionally misleading based on the content of the news item. According to the research, the two main approaches currently being used are linguistic methods and network methods. Both of these methods make use of machine learning on their training data set (Conroy et al.,

2015).

3.1 Linguistic methods

These methods utilize knowledge of speech patterns that are able to identify truthfulness and deception more accurately than most humans. Under this approach, a basic way that text is analyzed is considering all words in a block of text as equally significant units. Using natural language processing, this technique would be implemented using *n-grams* to analyze word frequencies and find indicators of deception. This method might also involve tagging the lexical cues of words (also called shallow syntax) or frequencies of words which can uncover linguistic patterns of deception. The techniques under this approach rely heavily on the analysis of the usage of language. They also work very well when combined with other approaches (Conroy et al., 2015).

3.2 Network methods

These methods make use of a network of associated information (like metadata) to predict the level of veracity of the content. It is also pointed out that the use of networks of data can provide a means to check the validity of a news item due to the presence of findable truths in the network. This involves making queries on existing knowledge to measure the truthfulness of new news items (Conroy et al., 2015).

The conclusion drawn was that both methods are very accurate in classifying news items. This gives rise to the use of a hybrid methods that takes into account both approaches to automating fake news detection. Such hybrid methods would have a linguistics-based analysis process that takes into account lexical analysis. It would also be able to perform efficiently in place of a strictly linguistic or network based approach. These techniques should be created with the intent of complementing the processes performed by a researcher in detecting fake news, as opposed to replacing them (Conroy et al., 2015).

4 Proposed Approach

4.1 Data Available

The fake news dataset is sourced from the Emergent Dataset described by Ferreira and Vlachos (Ferreira and Vlachos, 2016). This development dataset is formatted as triples of headline text, body text and labels. The headlines and bodies are primarily plain text, but some contain emoji or other UTF-8 symbols. The labels map to the four classes defined in the fake news challenge, where each class maps to one of the following stances (Fake News Challenge, 2017):

1. **Agrees:** The body text agrees (expresses the same viewpoint) with the claim in the headline.
2. **Disagrees:** The body text does not agree with the headline.
3. **Discusses:** The body text and the headline pertain to the same topic, but the body text neither agrees nor disagrees with the headline.
4. **Unrelated:** The body text and the headline cover different claims or topics.

There are approximately 50,000 articles in the development dataset. The distribution of articles by class (stance) is as follows (Fake News Challenge Github, 2017b):

unrelated	discuss	agree	disagree
0.73131	0.17828	0.0736012	0.0168094

When creating testing data from the entire data set, the labels are removed to leave pairs of headlines and body text.

4.2 Baseline Implementation

The Fake News Challenge competition provides an example baseline for stance classification. The default baseline classifier utilizes gradient boosting with 200 estimators and a hand-picked set of features (Fake News Challenge, 2017):

- **Word overlap**: The proportion of unique words shared between the headline and body, calculated using the formula $\frac{|types_{headline} \cap types_{body}|}{|types_{headline} \cup types_{body}|}$
- **Refuting**: If the headline contains a keyword from a set of refuting words (e.g. fake, fraud, hoax)
- **Polarity**: If there is a negative sentiment expressed in the headline and/or body. This feature also uses the set of refuting words.

- **Binary Co-occurrence**: How often tokens in the headline appear in the body.
- **N-gram overlap**: How often n-grams from the headline appear in the introductory paragraph of the body and the entire body text. The baseline implementation tests n-grams of size 2-6.
- **Char-gram overlap**: How often character sequences (char grams) from the headline appear in the introductory paragraph of the body and the entire body text. The baseline implementation tests char grams in powers of two from 2-16.

However, the implementation should be compatible with most other general classifiers such as Support Vector Machines.

Two methods are employed for splitting the data. A hold-out set defaulting to 20 percent of the dataset is first extracted for use in scoring and evaluation. Ten-fold cross validation is subsequently used for the training process. Both of these methods enable the scoring process, which calculates the accuracy of the classifier on the development dataset. The default classifier and feature set score 79.53% using this metric (Fake News Challenge Github, 2017a).

4.3 Potential Improvements

For our work on the fake news challenge problem, we have created a 3-stage plan for investigating and improving on the existing baseline:

1. Explore different classifiers with the given feature set or tune the existing classifier. Given the relative simplicity of the baseline, tuning the default gradient boosting classifier or substituting a similar classifier may allow for swift feedback without revising the existing set of features.
2. Investigate alternative features and encodings (such as those used by Augenstein et. al).
3. Explore alternative general methods such as long-short term memory networks (Augenstein et al., 2016).

We also plan on extending the existing scoring system to display metrics such as precision and recall for the classifiers used. We hope this will allow for a finer understanding and more specific method of ranking performance among the various implementations to be tested.

5 Timeline

May 18 - Proposal done (all members)
June 6 - Testing basic modifications to baseline (all members)
June 15 - Midterm, run the baseline, presentation (all members)
June 29 - Testing of first alternative model and features (all members)
July 6 - Final feature set complete (all members)

This initial timeline will be elaborated and more fine-grained roles assigned as the project progresses and we gain more knowledge of the problem space.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election .

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464* .

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52(1):1–4.

Fake News Challenge. 2017. Fake news challenge stage 1 (fnc-i): Stance detection. http://www.fakenewschallenge.org/.

Fake News Challenge Github. 2017a. Baseline fnc implementation. https://github.com/FakeNewsChallenge/fnc-1-baseline.

Fake News Challenge Github. 2017b. Stance detection dataset for fnc-1. https://github.com/FakeNewsChallenge/fnc-1.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL.

Adam Kucharski. 2016. Post-truth: Study epidemiology of fake news. *Nature* 540(7634):525–525.

Jonathan Rose. 2017. Brexit, trump, and post-truth politics.

Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology* 52(1):1–4. https://doi.org/10.1002/pra2.2015.145052010083.

S Wineburg, S McGrew, J Breakstone, and T Ortega. 2016. Evaluating information: The cornerstone of civic online reasoning.