Fake News Detection
Statistical NLP Project Proposal

Daniele Imperiale
Computer Science Department
New York University
daniele.imperiale@nyu.edu

Yizhi Li
Computer Science Department
New York University
yizhi.li@nyu.edu

Abstract

Fake news detection has proven to be a particularly important and challenging problem over the last few years, particularly after the advent of social media and the rise of alternative news sources. For our final project we intend to add to the existing literature on this field by leveraging some meta data that we believe has not been considered so far. In particular, we plan to adopt a multi-source neural network based framework that has been proved effective in fake news detection, while expanding on the linguistic and contextual features that can be used to discern true statements from false ones. We decide to conduct our experiments on the *LIAR* dataset, which has abundant baselines and offers the possibility of studying varying degrees of truthfulness. To better explore the intuition behind the task, our proposed project mainly focus on how to obtain universal and domain-specific features of fake news detection. Meanwhile, we are looking for a good design of experiment settings that can reveal the effect of our proposed methods.

1 Introduction

Fake News Detection is a popular natural language processing research topic since the Internet greatly advocates freedom of speech and decentralization of news media, which unfortunately has come at the cost of spreading of information that often times is completely unvetted. Statements in textual format can be modified and broadcasted easily, many among which are the news containing misleading information entirely or partially. Moreover, the increase in information accessibility has also accelerated the spread of fake news. Misleading fake news usually play a pivotal role in events of great impact for the society as a whole, such as presidential election and the COVID-19 pandemic. In light of these considerations, distinguishing fake news from true ones has become indispensable for the healthy functioning our society.

This is a particularly challenging problem especially because there exist various shades and degrees of truth, which makes the detection task even more difficult. To reflect this concept of ambiguity of truthful pieces of information, many works on fake news detection formulate it as a multi-classification task to detect the degree of truthfulness of the piece of news. Among the detection methods, automatic detection approaches are extremely essential because of the huge quantity of the spreading fake news text, and have been widely embraced by social media platforms lately.

The great progress made on recent works of Machine Learning, especially Deep Learning LeCun et al. (2015), contribute a lot to many NLP downstream tasks. In our project, we propose a modified multi-source neural network based framework Karimi et al. (2018) that already proved effective on fake news detection. The framework contains two parts: extractors of unified-structure for features from different sources, and a features combiner implemented with attention mechanism. Since the feature extractors all use same neural model, the input from different sources such as news statement and speaker's truthfulness history need to be design carefully. Under this framework,

Preprint. Under review.

we can research and merge features from different prospective into existing models safely with well-designed classification loss. Oshikawa1 and Yang Wang (2020) offers as complete survey on the state of the art of the current methods that have been used for fake news detection.

There are mainly two types of automatic fake news detection methods. The *fact-checking based* methods focus on comparison between political statements from politicians and pundits and known facts, as analyzed by professional journalists and editors. The other type of *linguistic pattern based* methods pay attention on the linguistic information in the text such as phrase-level pattern or word-level sentiment analysis. We mainly investigate automatic fake news detection based on surface-level linguistic patterns, by trying to answer the following two questions:

- Based on surface-level linguistic structures, how well can machine learning algorithms classify a short statement into a fine-grained category of truthfulness?
- Can we design a deep neural network architecture to integrate the speaker's information and other related meta-data with text to enhance the performance of fake news detection?

For the first question, we are planning to leverage knowledge from "traditional" fakeness detection of texts. We will follow previous works to help our model discover general *Psychological and Linguistic* fake news patterns that may be easily transferred into other datasets, and possibly expand on the few linguistic structures we have seen being used so far. For the second question, we decide to conduct a domain-based speaker meta analysis on the *LIAR* dataset. Features including *Topic Polarization* and *Long-term Topic Preference* are expected to be beneficial to the detection task. Although the newly mined features could be hard to transfer to datasets in different domains, the idea of the feature mining pipeline can be useful for reproduction.

2 Data Set

For this binary/multiclass task, a recent benchmark dataset for fake news detection is LIAR, introduced by Wang (2017). This dataset collected a decade-long, 12.8K manually labeled short statements in various contexts from Politifact, as Vlachos and Riedel (2014). Each statement is labeled with six-grade truthfulness and contains information about the subjects, party, context, speakers home state, and credit history in the form of historical counts of inaccurate statements for each speaker. These statements are sampled from various of contexts/venues, and the top categories include news releases, TV/radio interviews, campaign speeches, TV ads, tweets, debates, Facebook posts, etc.

Rashkin et al. (2017) also published large datasets, by augmenting the Politifact dataset with more articles from PunditFact, which focuses on statements from professional pundits. In this project we decide to focus on the LIAR dataset also because it has been more extensively used by other works in this literature.

3 Related Work

This section describes some recent research on Fake News Detection in binary or multiclass classification settings where the news article are explicitly labeled either as truthful or fake (binary), or with various degrees of truthfulness.

The majority of existing research uses supervised methods, while semi-supervised or unsupervised methods are less commonly used. Among the non-neural networks models, Support Vector Machine (SVM) and Naive Bayes Classifier (NBC) are frequently used for this classification problems (Conroy et al., 2015; Khurana and Intelligentie, 2017; Shu et al., 2018). However, to best of our knowledge, most of recent works focus on Neural Networks.

Wang (2017) introduces a new reference dataset (LIAR) and uses a model based on Kim's CNN (Kim, 2014), concatenating the max-pooled text representations with the meta-data representation from the bi-directional LSTM. In his work, CNN outperforms various other non-neural networks models, and he obtains significant improvements by combining meta-data with text, like information on the speaker, the context of the statements, and other information.

Rashkin et al. (2017) analyzes linguistic patterns across different types of articles by sampling standard trustworthy and untrustworthy news sources. To characterize differences between these

news types, they also applied various lexical resources, like the Linguistic Inquiry and Word Count (LIWC), and estimate the use of strongly and weakly subjective words with a sentiment lexicon (Wilson et al., 2005) and intensifying lexicons. Their LSTM model outperforms simpler MaxEnt and Naive Bayes when only using text as input; however the other two models improve substantially with adding LIWC features. Interestingly, the LIWC features do not improve the neural model much, indicating that some of this lexical information is perhaps redundant to what the model was already learning from text.

Attention mechanisms are often incorporated into neural networks to achieve better performance. Long (2017) used an attention model that incorporates the speaker's name and the statement's topic to attend to features first, then weighted vectors are fed into an LSTM. Their LSTM without the speaker's profile does not perform better than the CNN of Wang (2017). However, speaker profile information like credit history, speaker's location, party affiliation and job title can improve fake news detection by an extra 2-3%, which leads to the interpretation that speaker's intention to speak the truth or fake it largely depends on his/her, profiles, especially his/hers credit history.The best results are obtained when all these attributes are considered together, with performance sitting at around 40% accuracy. Kirilin and Strube (2018) also leverages the speakers metadata of lie history in the detection task.

Karimi et al. (2018) addresses an end-to-end neural network framework to recognize different levels of fake news. This Multi-Source Multi-Class Fake News Detection (MMFD) framework combines CNN and LSTM as feature extractor and uses the attention mechanism to combine features from different sources of the same training sample. The sources here refers to the news statement, the speaker information, the fake news stated history of the speaker and corresponding verification reports from experts. This work shows that using DNNs to merge features is plausible and combing multiple meaningful features benefits the downstream fake news detection task. We propose to follow the MMFD framework using DNNs to conduct feature combination, which will help us exploit the potential of proposed fake news detection features.

Table 1: A summary of current results for LIAR on the 6-class label fake news prediction problem. +All means including all meta-data in LIAR.

Author	**Meta-Data**	**Base Model**	**Acc**
Wang		SVMs	0.255
		CNNs	0.270
	+ Speaker	CNNs	0.248
	+ All	CNNs	0.274
Karimi		MMFD	0.291
	+ All	MMFD	0.348
Long		LSTM + Attn	0.255
	+ All	LSTM (no Attn)	0.399
	+ All	LSTM + Attn	0.415
Kirilin	+ All	LSTM	0.415
	+ Sp2C	LSTM	0.457
Bhattacharjee	2-class label	NLP Shallow	.921
		Deep (CNN)	0.962

4 Methodology

4.1 Preliminaries

We propose to follow the MMFD architecture addressed in Karimi et al. (2018), which consists two parts:

Feature Extractor is a mixture of CNN and LSTM networks that are meant to extract information from sequential data. The extractor can handle various data such as the text of statement and the speaker's lying history.

Feature Combiner is addressed for merge the features obtained from different sources. After extraction, the feature vectors will be merged by the combiner with attention mechanism.

In addition to the designed neural network architecture, Karimi et al. (2018) also provide a well-defined multi-level fakeness detection loss, which leverage both the intra-class and inter-class distance of the training samples.

4.2 Fake News Detection Feature

Our proposed approach entails augmenting the set of features used for fake news detection used by the relevant literature that we described above by a variety of different approaches that, to the best of our knowledge, have not been studied extensively till now:

Psychological and Linguistic Patterns. Another angle we plan to investigate is the addition of a few more features from the field of psychology and linguistics that have not been considered so far. This point is more work in progress and we have not yet defined some specific linguistic structures that we believe could be leveraged, but we found that what has been attempted in the existing literature was relatively simple (e.g. leveraging standard emotional and cognitive dictionaries like LIWC, as described above) and this probably offers some area for improvements. These patterns are expected to be universal and can be easily transferred to other similar tasks.

Topic Polarization. We believe that the polarization of different political statements on the same topic could be modelled more explicitly and be leveraged to make more informed conclusions about the veridicity of opposing statements. Clearly, if two opposing statements by two different speakers on the same topic are issued around the same time, to some extent they cannot be both true. To the best of our knowledge, while the topic is generally considered when trying to tackle this problem, there has been no attempt at directly informing the truthfulness of the statement by leveraging other "competing" statements.

Long-term Topic Support. We plan to add the politicians' long term views on specific topics to inform how likely it is that a current statement is truthful. Past statements have only been incorporated in the form of "credit history", which only considers how likely the speaker was to be truthful in the past up to that point, and does not provide information on his/her long term view. We have not seen long term views being directly incorporated when assessing the truthfulness of a political statement. The reason why we believe there will be value in incorporating this source of information is that speakers with an long history of supporting a given viewpoint, political statement, policy or bill are more likely to be truthful on the same topic, and less prone to be deceiving for the sole purpose of gaining in the short term. Practically speaking, we plan to add this dimension by looking at the speaker's Wikipedia pages to see whether they supported/opposed certain rights/bills/topics in the past.

5 Experiments

The design of experiments is not finalized yet. The news statements in *LIAR* dataset are categorized into 6 level according to their fakeness. Clearly, formulate the task into a binary real-fake classification seems easier for the fake news detection models, which may have more baselines. However, if we stick on the original 6-class labeling, we may discover more tactic effect brought by different designed features. It's a trade-off dilemma since we both want to consolidate incremental gain on the baselines and explore the effect of linguistic patterns on fake news detection.

6 Results Evaluation

7 Conclusion

References

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1546–1557, 2018.

Angelika Kirilin and Micheal Strube. Exploiting a speaker's credibility to detect fake news. In *Proceedings of Data Science, Journalism & Media workshop at KDD (DSJM'18)*, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Yunfei Long. Fake news detection through multi-perspective speaker profiles. Association for Computational Linguistics, 2017.

Qian Jing Oshikawa1, Ray and William Yang Wang. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, page 6086–6093, 2020.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.

Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, page 18–22, 2014.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.