A Survey on Natural Language Processing for Fake News Detection

Ray Oshikawa
Natural Sciences I
College of Arts and Sciences
The University of Tokyo
ray.oshikawa@gmail.com

Jing Qian, William Yang Wang
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106 USA
{jing_qian,william}@cs.ucsb.edu

arXiv:1811.00770v1 [cs.CL] 2 Nov 2018

Abstract

Fake news detection is a critical yet challenging problem in Natural Language Processing (NLP). The rapid rise of social networking platforms has not only yielded a vast increase in information accessibility but has also accelerated the spread of fake news. Given the massive amount of Web content, automatic fake news detection is a practical NLP problem required by all online content providers. This paper presents a survey on fake news detection. Our survey introduces the challenges of automatic fake news detection. We systematically review the datasets and NLP solutions that have been developed for this task. We also discuss the limits of these datasets and problem formulations, our insights, and recommended solutions.

1 Introduction

Automatic fake news detection is the task of assessing the truthfulness of claims in news. This is a new, but critical NLP problem because both traditional news media and social media have huge social-political impacts on every individual in the society. For example, exposure to fake news can cause attitudes of inefficacy, alienation, and cynicism toward certain political candidates (Balmas, 2014). The worst part of the spread of fake news is that sometimes it does link to offline violent events that threaten the public safety (e.g., the PizzaGate (Kang and Goldman, 2016)). Detecting fake news is of crucial importance to the NLP community, as it also creates broader impacts on how technologies can facilitate the verification of the veracity of claims while educating the general public.

The conventional solution to this task is to ask professionals such as journalists to check claims against evidence based on previously spoken or written facts. However, it is time-consuming and costs a lot of human resources, For example, PolitiFact[1] takes three editors to judge whether a piece of news is real or not.

As the Internet community and the speed of the spread of information are growing rapidly, automated fact checking on internet content has gained plenty of interests in the Artificial Intelligence research community. The goal of automatic fake news detection is to reduce the human time and effort to detect fake news and help us to stop spreading them. The task of fake news detection has been studied from various perspectives with the development in subareas of Computer Science, such as Machine Learning (ML), Data Mining (DM), and NLP.

In this paper, we survey automated fake news detection from the perspective of NLP. Broadly speaking, we introduce the technical challenges in fake news detection and how researchers define different tasks and formulate machine learning solutions to tackle this problem. We discuss the pros and cons, as well as the potential pitfalls and drawbacks of each task. More specifically, we provide an overview of research efforts for fake news detection and a systematic comparison of their task definitions, datasets, model construction, and performances. We also discuss a guideline for future research in this direction. This paper also includes some other aspects such as social engagement analysis. Our contributions are three folds:

- We provide the first comprehensive review on Natural Language Processing solutions for automatic fake news detection;
- We systematically analyze how fake news detection is aligned with existing NLP tasks, and discuss the assumptions and notable is-

[1] https://www.politifact.com/

Name	Main Input	Data Size	Label	Annotation
LIAR	short claim	12,836	six-grade	editors, journalists
FEVER	short claim	185,445	three-grade	trained annotators
BUZZFEEDNEWS	FB post	2282	four-grade	journalists
BUZZFACE	FB post	2263	four-grade	journalists
some-like-it-hoax	FB post	15,500	hoaxes or non-hoaxes	none
PHEME	Tweet	330	true or false	journalists
CREDBANK	Tweet	60 million	30-element vector	workers
FAKENEWSNET	article	23,921	fake or real	editors
BS DETECTOR	article	-	10 different types	none

Table 1: A Summary of Various Fake News Detection Related Datasets. *FB: FaceBook.*

sues for different formulations of the problem;

- We categorize and summarize available datasets, NLP approaches, and results, providing first-hand experiences and accessible introductions for new researchers interested in this problem.

2 Datasets

A major challenge for automated fake news detection is the availability and the quality of the datasets. There exists a variety of datasets for fake news detection. We categorize them and discuss their characteristics.

2.1 One-or-Few-Sentences Datasets

2.1.1 Short Claims

A recent benchmark dataset for fake news detection is LIAR (Wang, 2017). This dataset includes 12,836 real-world short statements collected from PolitiFact, where editors handpicked the claims from a variety of occasions such as debate, campaign, Facebook, Twitter, interviews, ads, etc. Each statement is labeled with six-grade truthfulness. The information about the subjects, party, context, and speakers are also included in this dataset. Vlachos and Riedel (2014) and Ferreira and Vlachos (2016) are the first to study PolitiFact data, but LIAR is orders of magnitude larger and more comprehensive. However, note that the original LIAR paper does not include the editor's justification or evidence due to copyright concerns, and users will need to retrieve the justification/evidence separately using an API. Also, even though both the claims and the evidence are from real-world occasions, they are highly unstructured. Fact-checking remains relatively challenging for this dataset.

Fever (Thorne et al., 2018) is a dataset providing related evidences for fake news detection. Fever contains 185,445 claims generated from Wikipedia data. Each statement is labeled as *Supported, Refuted, or Not Enough Info*. They also marked which sentences from Wikipedia they use as evidence. Fever makes it possible to develop a system which can predict the truthfulness of a claim together with the evidence, even though the type of facts and evidence from Wikipedia may still exhibit some major stylistic differences from those in real-world political campaigns.

POLITIFACT, CHANNEL4.COM[2], and SNOPES[3] are three sources for manually labeled short claims in news, which is collected and labeled manually. Many datasets, such as Wang (2017) and Rashkin et al. (2017), are created based on these websites.

2.1.2 Posts On Social Networking Services

In addition to the websites mentioned above, posts on Social Networking Services (SNS), such as Twitter and Facebook, can also be a source of short news statements. There are some datasets for fake news detection focusing on SNS, but they tend to have a limited set of topics and can be less related to news.

BUZZFEEDNEWS[4] collects 2,282 posts from 9 news agencies on Facebook. Each post is fact-checked by 5 BuzzFeed journalists. The advantages of this dataset are that the articles are collected from both sides of left-leaning and right-leaning organizations, and they are enriched in

[2]https://www.channel4.com/news/factcheck/
[3]https://www.snopes.com/fact-check/
[4]https://github.com/BuzzFeedNews/2016-10-facebook-fact-check

Attributes	Value
ID of the statement	11972
Label	True
Statement	Building a wall on the U.S.-Mexico border will take literally years.
Subject(s)	Immigration
Speaker	Rick Perry
Speaker's job title	Governor of Texas
Party affiliation	Republican
Total Credit History Counts	30,30,42,23,18
Context	Radio Interview

Table 2: An Example Entry from LIAR. The ordered total credit history counts are {barely true, false, half true, mostly true, pants on fire}.

Potthast et al. (2017) by adding data such as the linked articles. BUZZFACE (Santia and Williams, 2018) extends the BuzzFeed dataset with the comments related to news articles on Facebook. It contains 2,263 news articles and 1.6 million comments. SOME-LIKE-IT-HOAX[5] (Tacchini et al., 2017) consists of 15,500 posts from 32 Facebook pages, that is, the public profile of organizations (14 conspiracy and 18 scientific organizations). This dataset is labeled based on the identity of the publisher instead of post-level annotations so that it may have imposed a strong assumption. A potential major pitfall for such dataset is that such kind of labeling strategy can result in machine learning models learning characteristics of each publisher, rather than that of the fake news.

PHEME (Zubiaga et al., 2016) and CREDBANK (Mitra and Gilbert, 2015) are two Twitter datasets. PHEME contains 330 twitter threads (a series of connected Tweets from one person) of nine newsworthy events, labeled as true or false according to thread structures and follow-follower relationships. CREDBANK contains 60 million tweets covering 96 days, grouped into 1,049 events with a 30-dimensional vector of truthfulness labels. Each event was rated on a 5-point Likert scale of truthfulness by 30 human annotators. They simply concatenate 30 ratings as a vector because they find it difficult to reduce it to a one-dimensional score.

As mentioned above, these datasets were created for verifying the truthfulness of tweets. Thus they are limited to a few numbers of topics and can include tweets with no relationship to news. Hence both datasets are not so much ideal for fake news detection so that they are more frequently used for rumor detection.

2.2 Entire-Article Datasets

There are several datasets for fake news detection focusing on fake news detection based on the entire article. For example, FAKENEWSNET (Shu et al., 2017a,b, 2018) is an ongoing data collection project for fake news research. It consists of headlines and body texts of fake news articles from BuzzFeed and PolitiFact. It also collects information about social engagements of these articles from Twitter.

BS DETECTOR[6] is collected from a browser extension named BS Detector, which indicates its labels are the outputs of BS Detector, not human annotators. BS Detector searches all links on a web page at issue for references to unreliable sources by checking against a manually compiled list of unreliable domains. Note that the major issue with using this dataset is that the machine learning models trained on this dataset are learning the parameters of the BS Detector.

Websites such as BLUFF THE LISTENER and THE ONION create sarcastic and humorous (Rubin et al., 2015a) fake news intentionally. Note that the types of fake news from these sources are limited. Moreover, it is relatively easy to classify them against traditional new media articles. A dataset consists of articles from various publishers can be better (Rashkin et al., 2017), though individual claims must be checked. We should also note that one must avoid using aggregate labels simply based on website source, as it adds more confounding variables and it is more of a website classification task.

[5]https://github.com/gabll/some-like-it-hoax

[6]https://github.com/bs-detector/bs-detector

3 Tasks

The general goal of fake news detection is to identify fake news. However, this task can be formulated in various ways.

3.1 Input

In this paper, we focus on fake news detection of text content. The input can be text ranging from short statements to entire articles. Additional information such as speakers' identity can be appended. Inputs are related to which dataset is used (see Section 2).

3.2 Output

In most studies, fake news detection is formulated as a classification or regression problem, but classification is more frequently used.

3.2.1 Classification

The most common way is to formulate the fake news detection as a binary classification problem. However, categorize all the news into two classes (fake or real) is difficult because there are cases where the news is partly real and partly fake. To address this problem, add additional classes is a common practice. There are mainly two ways of adding additional classes. One is to set a category for the news which is neither completely real nor completely fake. The other one is to set more than two degrees of truthfulness, like LIAR and CREDBANK. The latter method reflects human judgments more delicately. When using these datasets, the expected outputs are multi-class labels, and those labels are learned as independent labels with i.i.d assumptions (Rashkin et al., 2017; Wang, 2017).

3.2.2 Regression

Fake news detection can also be formulated as a regression task, where the output is a numeric score of truthfulness. This approach is used by Nakashole and Mitchell (2014). Formulating the task in this way can make it less straightforward to do the evaluation. Usually, evaluation is done by calculating the difference between the predicted scores and the ground truth scores, or using Pearson/Spearman Correlations. However, since the available datasets have discrete ground truth scores, the challenge here is how to convert the discrete labels to numeric scores.

3.2.3 Clustering

One of the conditions for fake news classifiers to achieve good performances is to have sufficient labeled data. However, to obtain reliable labels requires a lot of time and labor. Therefore, semi-supervised and unsupervised methods are proposed. The task is then formulated as a clustering problem instead of a classification one Rubin and Vashchilko (2012).

4 Methods

4.1 Preprocessing

Preprocessing usually includes tokenization, stemming, and generalization or weighting words.

To convert tokenized texts into features, Term Frequency-Inverse Document Frequency (TF-IDF) and Linguistic Inquiry and Word Count (LIWC) are frequently used. For word sequences, pre-learned word embedding vectors such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) are commonly used. Appropriate preprocessing is necessary for a better understanding of fake news. Mihalcea and Strapparava (2009) use LIWC and find there is a difference in word usage between deceptive language and non-deceptive ones, so using word classification may have significant meaning on detection.

When using entire articles as inputs, an additional preprocessing step is to identify the central claims from raw texts. Thorne et al. (2018) rank the sentences using TF-IDF and DrQA system (Chen et al., 2017). Solutions to the text summarization task can also be applied.

4.2 Collecting Evidences

The RTE-based (Recognizing Textual Entailment) method is frequently used to gather and utilize evidence. RTE is the task of recognizing relationships between sentences, which can be applied to fake news detection. By gathering sentences which is for or against input from data sources such as news articles using RTE method, we can predict whether the input is correct or not. RTE-based models need textual evidence for fact check; thus this approach can be used only when the dataset includes evidence, such as FEVER and Emergence. Besides, RTE models cannot learn correctly when a claim in a dataset does not have enough information as evidence. To address this

problem Thorne et al. (2018) develop a new approach which simulates training instances labeled as *Not Enough Info* by sampling evident sentences. Thus RTE models can use data without evidence.

4.3 Rhetorical Approach

Rhetorical Structure Theory (RST), sometimes combined with Vector Space Model (VSM), is often used for fake news detection (Rubin et al., 2015b). RST is an analytic framework for the coherence of a story. Through defining functional relations (e.g., Circumstance, Evidence, and Purpose) of text units, this framework can systematically identify the essential idea and analyze the characteristics of the input text. Fake news is then identified according to its coherence and structure.

To explain the results by RST, VSM is used to convert news texts into vectors, which are compared to the center of true news and fake news in high-dimensional RST space. Each dimension of the vector space indicates the number of rhetorical relations in the news text.

4.4 Machine Learning Models

As mentioned in section 3, the majority of existing research use supervised method while semi-supervised or unsupervised methods are commonly used. In this section, we mainly describe classification models with several actual examples.

4.4.1 Non-Neural Network Models

The most frequently used classification models in fake news detection are Support Vector Machine (SVM) and Naive Bayes Classifier (NBC). These two models differ a lot in structure thus comparing among them is meaningful. Logistic regression (LR) and decision tree such as Random Forest Classifier (RFC) are also used.

4.4.2 Neural Network Models

Many types of neural network models such as multi-layer perceptrons work for fake news detection, and many combinations of models are shown.

Recurrent Neural Network (RNN) is very popular in Natural Language Processing, especially Long Short-Term Memory(LSTM), which solves the vanishing gradient problem. LSTMs can capture longer-term dependencies. For example, Rashkin et al. (2017) set up two types of LSTM model, one put simple word embeddings initialized with GloVe into LSTM, and the other concatenate LSTM output with LIWC feature vectors before undergoing the activation layer. In both cases, they were more accurate than NBC and Maximum Entropy(MaxEnt) models, even though slightly.

Ruchansky et al. (2017) extract representations of both users and articles as low-dimensional vectors, and for representation of articles, they use LSTM for each article. Textual information of each social engagement for an article is processed by doc2vec and put in LSTM, and are integrated with the score of the user in the last layer to classify.

Convolutional neural networks (CNN) are also widely used since they succeed in many text classification tasks. Wang (2017) uses a model based on Kim's CNN (Kim, 2014). They concatenate the max-pooled text representations with the meta-data representation from the bi-directional LSTM. CNN also used for analyzation using a variety of meta-data. For example, Deligiannis et al. give graph-like data of relationships between news and publishers to CNN and assess news from them.

Karimi et al. (2018) proposed Multi-source Multi-class Fake news Detection framework (MMFD), in which CNN analyzes local patterns of each text in a claim and LSTM analyze temporal dependencies in the entire text. This model takes advantage of the characteristics of both models because LSTM works better for long sentences.

Attention mechanisms are often incorporated into neural networks to achieve better performance. Long et al. (2017) use attention model that incorporates the speakers name and the statements topic to attend to features first, then weighted vectors are fed into an LSTM. Doing this increases accuracy by about 3 % (shown in Table 3, id 3,4). Kirilin and Strube (2018) use a very similar attention mechanism.

Memory networks, which is a kind of attention-based neural network, also shares the idea of attention mechanism. Pham (2018) uses Single Layer Memory network to learn a different representation of words by memorizing the set of words in the memory. When judging, input sentences weight the words in memory by attention mechanism. Thus the model can extract related words from its memory.

5 Experimental Results

We compare empirical results on classification datasets via various machine learning models in this section. Table 3 summaries the results on four datasets: LIAR, FAKENEWSNET, FEVER, and PHEME.

In Table 3, we collect and compare the existing results of fake news classification research. For comparison, we use accuracy, which is a commonly used metric. Other evaluation metrics (Shu et al., 2017a) such as Precision, Recall, F-scores and ROC-AUC are also discussed.

6 Observations, Discussions, & Recommendations

6.1 Datasets and Inputs

Rubin et al. (2015a) define nine requirements for fake news detection corpus, and we agree: 1. Availability of both truthful and deceptive instances; 2. Digital textual format accessibility; 3. Verifiability of "ground truth"; 4. Homogeneity in lengths; 5. Homogeneity in writing matte; 6. Predefined timeframe; 7. The manner of news delivery; 8. Pragmatic concerns; 9. Language and culture.

Research on fake news detection has been progressing, and the situation has changed since these requirements were defined in 2015. As the performances on fake news detection are improved, the more reality-based and detailed detection becomes more realistic so that new datasets should be useful to develop models realizing such detection. Thus, we add three new recommendations for a new dataset based on cases found in previous research. Concerning developing more reality-based datasets, requirement 10 and 12 should be fulfilled, and concerning more detailed datasets, requirement 11 should be fulfilled.

10: Easy to create from raw data: Pragmatic fake news detection should be performed on emerging news, so models learned from datasets should not require much hand-crafted information. In order to imitate this and set a challenging task, datasets must not include too much information tagged by human except for true-or-false labels. For example, Karimi et al. (2018) supplement LIAR by adding the verdict reports written by label generators. When they do so, the attention score for that reports tends to be high as shown in Table 3 in this paper and raise accuracy by 4%. This could be the problem because verdict reports are highly related to answering and not generated in emerging news.

11: Fine-grained truthfulness: News or claims might be a mixture of true and false statements, so it is not practical to categorize them totally into true or false. When creating human annotators engaged in labeling news tend to believe what they read, shown in Buntain and Golbeck (2017). Besides, the binary classification has already achieved high accuracy around 90% even if inputs are restricted to textual sources Bhattacharjee et al. (2017) achieve over 96% accuracy (Table 3, id16) using only textual data of the claims themselves from LIAR, while 6-class classification is still a challenging task (id 1-14), Della Vedova et al. (2018) achieve almost 90% accuracy even when there is little social engagement data. In order to define a more challenging and practical task, the datasets should include more detailed truthfulness information.

12: Quote claims or articles from various speakers or publisher: When creating a new dataset, data should not be extracted from only one specific publisher, because a model will learn not fake news features but that of publishers. Moreover, when we choose which websites we use, we should be careful to what types of fake news it indicates (Hoaxes, Propaganda or Satire (Rubin et al., 2015a)). It is easier to use data from fact-checking sites such as PolitiFact, but the labels will rely on editors decision. In this way, we can avoid having confounding variables in the analysis that creates bias and complicates the study. For example, we strongly discourage anyone to use the **BS Detector** dataset, due to the lack of annotation and strong assumptions: This task is more like a classification of website types vs. fake news.

6.2 Models

First, we compare how each model process textual content based on NLP. Most models we shared in Table 3 used word embeddings, especially word2vec, for taking the meanings of each text. The key to applying machine learning to fake news detection is choosing efficient features from just text with redundant information because features differ among fake news and real news, not among news topics or publishers of the news, should be extracted.

Dataset	N	Author	Input	Base Model	Acc.
LIAR	6	Wang (2017)	Text	SVMs	0.255
			Text	CNNs	0.270
			Text+Speaker	CNNs (4.4.2)	0.248
			Text+Meta	CNNs	**0.274**
		Karimi et al. (2018)	Text	MMFD(4.4.2)	0.291
			Text+Meta	MMFD	**0.348**
		Long et al. (2017)	Text	LSTM+Att (4.4.2)	0.255
			Text+Meta	LSTM(no Att)	0.399
			Text+Meta	LSTM+Att	**0.415**
		Pham (2018)	Text+Meta	NN(Single Att)	0.276
			Text+Meta	NN(Dual Att)	0.373
			Text+Credit	MM(4.4.2)	**0.442**
		Kirilin and Strube (2018)	Text+Meta	LSTM(4.4.2)	0.415
			Text+Meta+Sp2C	LSTM	**0.457**
	2	Bhattacharjee et al. (2017)	Text	NLP Shallow	0.921
				Deep (CNN)	0.962
FAKE NEWS NET (2017b)	2	Shu et al. (2017b)	BuzzFeed	RST(4.3)	0.610
				LIWC(4.1)	0.655
				Castillo	0.747
				TriFN	**0.864**
		Della Vedova et al. (2018)		HC-CB-3(6.2)	**0.856**
		Deligiannis et al.		GCN(4.4.2)	**0.944**
		Shu et al. (2017b)	PolitiFact	RST	0.571
				LIWC(4.1)	0.637
				Castillo	0.779
				TriFN	**0.878**
		Deligiannis et al.		GCN	**0.895**
		Della Vedova et al. (2018)		HC-CB-3(6.2)	**0.938**
FEVER	3*	Thorne et al. (2018)	claim & evidences	Decomposable Att	**0.319**
	3				0.509
	3*	Yin and Roth (2018)		TWOWINGOS	**0.543**
	3				0.760
	3*	Hanselowski et al. (2018)		LSTM (ESIM-Att)	**0.647**
	3				0.684
	3*	UNC-NLP		(not yet announced)	**0.640**
	3			http://fever.ai/task.html	0.680
	3*	UCL Machine Reading			**0.623**
	3				0.674
	3*	Athene UKP TU Darmstadt			**0.613**
	3				0.652
PHEME	2	Kochkina et al. (2018)	9 events	NileTMRG	0.360
				branchLSTM	0.314
				MTL3	**0.405**
			5 major events	NileTMRG	0.438
				branchLSTM	0.454
				MTL3	**0.492**

Table 3: The Current Results for Fake News Detection. *Papers are sorted by the accuracy of the most accurate model. The highest result in each paper is in bold. "Att" is short for "Attention". Acc.: Accuracy. In* FEVER, $N = 3*$ *means that the task is combined with evidence collection, and strictly speaking, it is not a classification task but claims verification, but we put it for your information.*

There are some essential features to extract particularly in fake news detection. First, the psycholinguistic categories of words used in the fake news have been proven to be different in some researches since Mihalcea and Strapparava (2009) find characteristics of the word used in deceptive languages. Shu et al. (2017b) achieve 64% accuracy on FAKENEWSNET by only analyzing word usage in LIWC. Thus it is clear analyzation on word usage contributes much to detecting fake news. Second, the rhetorical features may differ in fake news. Rubin and Vashchilko (2012) show that there should be some differences in the structure of sentences in deceptive languages. In Table 3, RST (4.3) is the only framework to learn such features, and achieve 61% accuracy on FAKENEWSNET.

However, those hand-crafted features extraction may be replaced by neural networks. Rashkin et al. (2017) shows that adding LIWC did not improve the performance of the LSTM model but even harm it while Naive Bayes and MaxEnt models are improved. It may be because some neural network models like LSTM can learn lexical information in LIWC by themselves. There is no such a study on rhetorical features so we cannot conclude, but neural network models may also lean them, considering the RST model(id 17,23) achieve only low accuracies compared to other methods.

Hence it may be better to use automated learning methods. For Natural Language Processing, LSTM and attention based method such as attention attachments or memory network is often used. It is because they can analyze long-term and content-transitional information so that they can use the abundant word data of sentences and detect context. Actually, many research in Table 3 use attention methods (id 7,9-14,19,20,25,26,29,30,33,34) or LSTM (id 5-9,13,14,33,34,42,46) to learn textual models. A popular application of attention mechanism is to generate attention weights for hidden layers based on meta-data.

Second, considering additional information other than text in claims or articles, such as speaker credibility or social engagements data is the other efficient and practical method; thus most recent studies mainly focus on this method. Most studies on LIAR improve accuracy by changing the way to introduce not texts but speakers' information because it is difficult to detect a lie from short sentences. Kirilin and Strube (2018) improve accuracy by 21% through replacing the credibility history in LIAR's with a larger credibility source they launched named speak2credit[7] (id 13-14) They show that their attention model relies on speaker's credibility by 43%, much higher than 17% on a statement of claim, by case study.

However, the tendency to rely their judgments on speakers or publishers may cause some problem. Vlachos said that the most dangerous misinformation comes from the sources we trust , and upgrading or downgrading specific sources cause silencing minorities' voice(Graves, 2018).

Thus he developed new datasets FEVER including evidence so that it can be used for claim verification not only for classification. Such content-based approaches should be developed more in the future. For claim verification on FEVER, Yin and Roth (2018) improves precision rate to 45% from 10% (the benchmark score). The point is that considering the recall rate does not change that dramatically (from 46% to 50%), this model has less chance of verifying fake claim incorrectly. Research on FEVER is fewer than that on others because this dataset was published very recently and the accuracy, recall and precision rate are relatively low in most studies. There are very latest results in Table 3 (id 35-40), but their performances do not make much difference.

Social engagements data also shows to be effective. For example, in Shu et al. (2017b) the model using only social engagements data (id 19,25) defeated the model using only textual data (id 17,18,23,24). The same as using speakers credibility, we should think about the proper use of additional data as Della Vedova et al. (2018)(id 21,28) developed model which uses the content-based method when there are not enough social-engagements-based information and otherwise use mainly social-based one.

7 Related Problems

7.1 Fact Checking

Fact checking is the task of assessing the truthfulness of claims made by public figures such as politicians, pundits, etc (Vlachos and Riedel, 2014). Many researchers do not distinguish fake news detection and fact checking since both of

[7] https://github.com/akthesis/speaker2credit

them are to assess the truthfulness of claims. However, fake news detection usually only focuses on news events while fact checking is broader. Thorne and Vlachos (2018) provides a comprehensive review on this topic.

7.2 Rumor Detection

There is not a consistent definition of rumor detection. A recent survey (Zubiaga et al., 2018) defines rumor detection as separating personal statements into rumor or non-rumor, where rumor is defined as a statement consisting of unverified pieces of information at the time of posting. In other words, rumor must contain information that is worth verification rather than subjective opinions or feelings.

7.3 Stance Detection

Stance detection is the task of assessing whether the document supports a specific claim or not. It is different from fake news detection in that it is not for veracity but for consistency. Stance detection can be a subtask of fake news detection since it can be applied to searching documents for evidence (Ferreira and Vlachos, 2016).

7.4 Sentiment Analysis

Sentiment analysis is the task of extracting emotions, such as customers' favorable or unfavorable impression of a restaurant. Different from rumor detection and fake news detection, sentiment analysis is not to do an objective verification of claim but to analyze personal emotions.

8 Conclusion

We described and compared previous datasets and proposed new requirements for future datasets; 1.Easy to make from raw data in internets, 2.Have enough classes of truthfulness, 3.Quote claims or articles from different speakers or publishers. Besides, We compared the accuracy of many previous experiments and made some challenging task to our future fake news detection model; 1. More textual content-based method on multi-class fake news detection based on Natural Language Processing should be developed for realizing reliable detection. 2. We need a more logical explanation for fake news characteristics 3. There should be a limitation in language-based fake news detection in the case that there are not enough linguistic differences to improve detection accuracy to very high rate so that we should extend the way of verification with evidence as the content-based method. 4. Note that hand-crafted features extraction will be replaced by neural network models improvement.

References

Meital Balmas. 2014. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research*, 41(3):430–454.

Sreyasee Das Bhattacharjee, Ashit Talukder, and Bala Venkatram Balantrapu. 2017. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *Big Data (Big Data), 2017 IEEE International Conference on*, pages 556–565. IEEE.

Cody Buntain and Jennifer Golbeck. 2017. Automatically identifying fake news in popular twitter threads. In *Smart Cloud (SmartCloud), 2017 IEEE International Conference on*, pages 208–215. IEEE.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Nikos Deligiannis, Tien Huu Do, Duc Minh Nguyen, and Xiao Luo. Deep learning for geolocating social media users and detecting fake news.

Marco L Della Vedova, Eugenio Tacchini, Stefano Moret, Gabriele Ballarin, Massimo DiPierro, and Luca de Alfaro. 2018. Automatic online fake news detection combining content and social signals. In *2018 22nd Conference of Open Innovations Association (FRUCT)*, pages 272–279. IEEE.

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*, pages 1163–1168.

Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. *Factsheet*, 2:2018–02.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479*.

Cecilia Kang and Adam Goldman. 2016. In washington pizzeria attack, fake news brought real guns. *the New York Times*.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Angelika Kirilin and Micheal Strube. 2018. Exploiting a speakers credibility to detect fake news. In *Proceedings of Data Science, Journalism & Media workshop at KDD (DSJM18)*.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. *arXiv preprint arXiv:1806.03713*.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 252–256.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pages 258–267.

Ndapandula Nakashole and Tom M Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1009–1019.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Trung Tin Pham. 2018. A study on deep learning for fake news detection.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. 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.

Victoria L Rubin, Yimin Chen, and Niall J Conroy. 2015a. Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 83. American Society for Information Science.

Victoria L Rubin, Niall J Conroy, and Yimin Chen. 2015b. Towards news verification: Deception detection methods for news discourse. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, January*, pages 5–8.

Victoria L Rubin and Tatiana Vashchilko. 2012. Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 97–106. Association for Computational Linguistics.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806. ACM.

Giovanni C Santia and Jake Ryland Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *ICWSM*, pages 531–540.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017a. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2017b. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*.

Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. *COLING*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction.

In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

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

Wenpeng Yin and Dan Roth. 2018. Twowingos: A two-wing optimization strategy for evidential claim verification. *arXiv preprint arXiv:1808.03465*.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.