Social Network Analysis and Mining (2019) 9:37
https://doi.org/10.1007/s13278-019-0580-z

ORIGINAL ARTICLE

Understanding archetypes of fake news via fine-grained classification

Liqiang Wang[1,2] · Yafang Wang[2] · Gerard de Melo[3] · Gerhard Weikum[1]

Received: 11 December 2018 / Revised: 30 May 2019 / Accepted: 1 June 2019

Abstract
Fake news, doubtful statements and other unreliable content not only differ with regard to the level of misinformation but also with respect to the underlying intents. Prior work on algorithmic truth assessment has mostly pursued binary classifiers—factual versus fake—and disregarded these finer shades of untruth. In manual analyses of questionable content, in contrast, more fine-grained distinctions have been proposed, such as distinguishing between hoaxes, irony and propaganda or the six-way truthfulness ratings by the PolitiFact community. In this paper, we present a principled automated approach to distinguish these different cases while assessing and classifying news articles and claims. Our method is based on a hierarchy of five different kinds of fakeness and systematically explores a variety of signals from social media, capturing both the content and language of posts and the sharing and dissemination among users. The paper provides experimental results on the performance of our fine-grained classifier and a detailed analysis of the underlying features.

Keywords Fake news · Unreliable content · Social media · Fine-grained classification

1 Introduction

A recent study of information spread in Twitter has shown that fake news is disseminated substantially faster, farther, deeper and more broadly than reliable content on comparable topics (Vosoughi et al. 2018). The ability of fake news and doubtful claims to outpace serious reporting and verified facts gives it an undue advantage in influencing public opinions. This big societal problem has motivated researchers to develop largely automated methods to assess the "truth" (i.e., factuality and authenticity) of news and statements, leading to tools for fact checking, credibility assessment and trust analysis (see e.g., Conroy et al. 2015; Li et al. 2015; Popat et al. 2017; Rashkin et al. 2017; Wang 2017).

These methods are based on a variety of powerful data mining and machine learning techniques, with training data from manually labeled collections such as Snopes[1] or PolitiFact[2]. While assessing the absolute truth of news and claims remains an elusive goal, supervised classifiers can provide insights on the credibility of online contents and the nature of misinformation. However, prior work has mostly focused on binary classification: labeling an article as either fake or credible. Such binary classifiers thus neglect the diversity of fake news. This motivates the work in this paper, to go beyond binary classification and systematically explore the finer shades of misinformation.

Berghel (2017b) proposed a taxonomy of fake news according to the sources of news, including social satire sources, disclosed sources of fake news, anonymous sources and bogus sources. At the same time, fake news typically comes with a specific intent, such as for business profit or for political purposes. Rubin et al. (2015) distinguish three types of fake news based on the degree of deception: serious fabrication, large-scale hoax and humorous fake. The recent SHPT scheme by Rashkin et al. (2017) distinguishes satire, hoax, propaganda and trusted news.

✉ Yafang Wang
yafang.wang@sdu.edu.cn

Liqiang Wang
lwang@mpi-inf.mpg.de

Gerard de Melo
gdm@demelo.org

Gerhard Weikum
weikum@mpi-inf.mpg.de

[1] Max Planck Institute for Informatic, Saarbrücken, Germany

[2] Department of Computer Science, Shandong University, Jinan, China

[3] Department of Computer Science, Rutgers University, New Brunswick, NJ, USA

---

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

[2] http://www.politifact.com/.

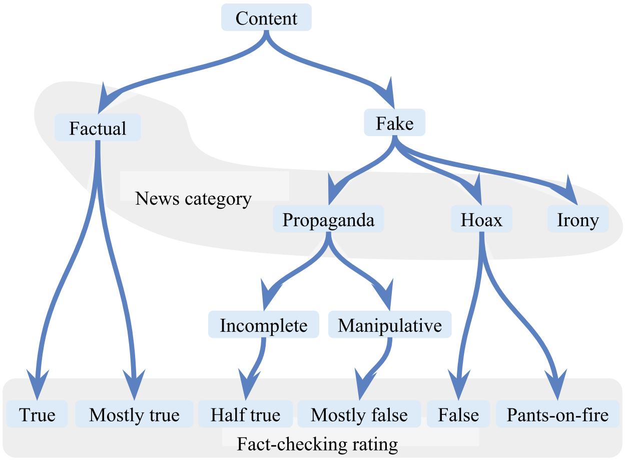


Fig. 1 Classification hierarchy of fake content

PolitiFact, the most reputed online community providing manual assessments of claims, relies on a six-way "Truth-O-Meter" rating system with labels "true", "mostly true", "half true", "mostly false", "false" and "pants-on-fire". Their guidelines[3] crisply explain each of these categories, particularly the ones in the middle, which reflect different degrees of misinformation. The "half true" label is for content that leaves out important context, and the "mostly false" label is assigned when content is severely misleading by omitting critical information or exaggerating a certain message.

To reflect such finer-grained classifications, we devised a taxonomic hierarchy that captures both the SHPT scheme and the PolitiFact ratings. This is shown in Fig. 1. To reconcile the SHPT and PolitiFact categorizations, we organize their labels as a tree, which leads to five major categories of fake news: factual, propaganda, hoax and irony and two refinements of propaganda to distinguish incomplete context from manipulative statements.

Figure 2 depicts our framework of fake news detection and analysis under the introduced taxonomy. This finer-grained classification system provides a new starting point for us to more deeply understand the patterns in fake contents, as well as how it spreads and influences users. To this end, we develop a hierarchical classifier that labels doubtful news or statements with one of our five "shades of untruth" (Wang et al. 2018). In contrast to prior work, we tap into a variety of signals from social media such as Twitter. Each news article or statement that we question is automatically expanded by gathering a large set of tweets that specifically relate to the claim. Although the tweets themselves are fairly noisy and susceptible to topic drift, our classifier leverages their aggregated signals for fairly accurate predictions. In addition to the initial classifier introduced in our short paper (Wang et al. 2018), which is based on logistic regression, we here present further results. First, to compare the classification effectiveness of different types of classifiers, we devise another classifier using deep neural network techniques. Second, we provide a deeper analysis of how misinformation is expressed including an in-depth analysis of linguistic evidence. In addition, previously unexplored aspects are investigated to better understand the different kinds of misinformation, including the sharing and spread of misinformation on Twitter, sentiment and subjectivity cues, the credibility of content creators, content drift between news and tweets and the stance of users in social networks. To the best of our knowledge, this comprehensive feature space and classification methodology has not been systematically studied in prior work on fake news detection.

The paper's salient contributions are:

1. We introduce a taxonomic hierarchy and present fine-grained classifiers of questionable news and statements, harnessing features from social media contents and dynamics.
2. We compare different classifiers, including a newly devised neural network, and systematically study their feature spaces.
3. We provide an analysis of different kinds of fake contents, considering both linguistic characteristics of user posts and the sharing dynamics in Twitter.
4. We conduct a series of experiments, to obtain insights on how various kinds of misinformation are expressed. Our datasets will be made publicly available to support further research.

2 Related work

2.1 The taxonomy of fake news

Two key factors that define fake news are the authenticity (or, the lack thereof) and intent. These are widely adopted in recent studies (Shu et al. 2017a) and serve as the foundation for the taxonomy by Rashkin et al. (2017). A number of alternative taxonomies have been put forth, based on other definition criteria. Berghel (2017b), for instance, propose four categories of fake news, based on the source of the news, and further introduce the alt-facts and post-truth fact checking categories as additional cases distinct from the fake news ones (Berghel 2017a). Campan et al. (2017) consider a taxonomy of fake news that distinguishes clickbait, propaganda, commentary/opinion and

---

[3] http://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/.

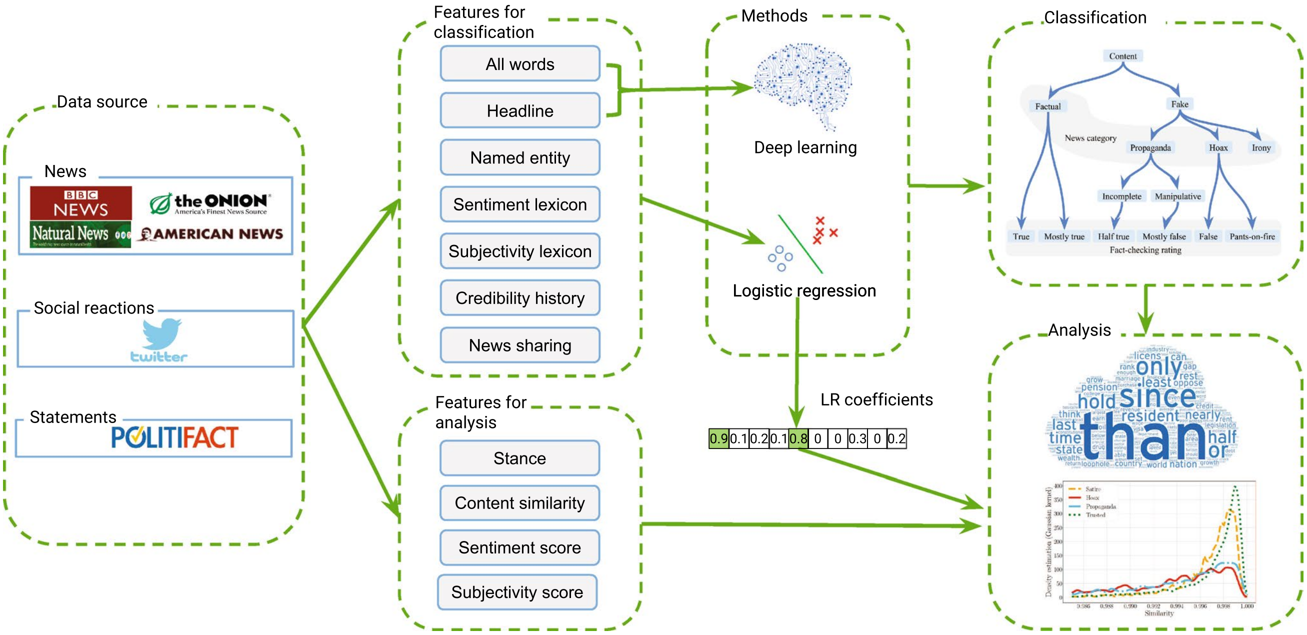


Fig. 2 Framework of fake news detection and analysis under the hierarchical taxonomy

humor/satire. Rumors are another special form, related to fake news, which is particularly prominent on social media (Rath et al. 2017; Yu et al. 2017). These divergent classification schemes have inspired us to present our finer-grained taxonomy.

2.2 The detection of fake news

Given the increasing proliferation and societal effects of fake news, in recent years, substantial research has focused on the task of detecting fake news, i.e., distinguishing it from genuine news. There are three main approaches to this. The first is to rely on intrinsic linguistic cues (Bourgonje et al. 2017; Jin et al. 2016; Potthast et al. 2017; Rashkin et al. 2017; Singhania et al. 2017) in the content of the article, tweet, claim or statement, drawing, for instance, on the occurrence of specific words or on stylistic cues in the headlines of articles. Another approach is to consider social media features (Farajtabar et al. 2017; Kim et al. 2017; Rath et al. 2017; Wu and Liu 2018; Yu et al. 2017), such as fake news spreading cascades, user profiles, user actions, the timeline of dissemination. Finally, some studies combine the two kinds of features for more reliable predictions (Del Vicario et al. 2018; Popat et al. 2017; Ruchansky et al. 2017; Shu et al. 2017b). However, among the above works, few provide any comprehensive feature analysis combining the news and tweet comments. Hence, we consider several features from these works, including the use of tweet comments, in our fake news detection system and provide a more in-depth analysis of the contributions of different signals.

2.3 Fake news analytics

In terms of analytics with respect to fake news, past work has fallen into two main categories. One line of work has sought to shed light on the characteristics of fake news. In this regard, research from a sociological vantage point has often attempted to understand the phenomenon of fake news based on particular theories, such as the *third-person effect* (Jang and Kim 2018) and the *filter bubble effect* (DiFranzo and Gloria-Garcia 2017). Fourney et al. (2017) assess the influence of fake news during the 2016 US presidential election based on website visit statistics. Rashkin et al. (2017) provide a contrastive analysis of the language style adopted in genuine news as opposed to several sorts of fake news, including hoaxes, satire and propaganda. They adopt the linguistic inquiry and word count (LIWC) resource, which is another kind of lexicon different from ours.

A further line of work has targeted the question of how to intervene so as to mitigate the harmful effects of fake news. Based on a fake news detection model with user exposure, Kim et al. (2017) introduce different strategies to decide the time to start the fact-checking process so as to reduce the impact. Spivey (2017) first simulates the dissemination of rumors on random formation graphs and scale-free networks and then attempts to infer solutions to control the influence of fake news based on the simulations. Farajtabar et al. (2017) present a model to reduce the influence of fake news via a point process-based intervention. Del Vicario et al. (2018) argue that confirmation bias and the polarization of society make it possible to identify at an early stage which

Table 1 SHPT datasets statistics

Type	Source	# of doc	# of shares	# of comments	Date
Satire	The onion	5000	1,800,295	578,433	Aug. 2013 ~ Mar. 2018
Hoax	American news	5000	109,228	14,371	Feb. 2016 ~ Mar. 2018
Propaganda	Natural news	5000	230,352	15,315	May. 2017 ~ Mar. 2018
Trusted	BBC news	5000	2,124,903	596,940	Aug. 2016 ~ Mar. 2018

Table 2 PolitiFact dataset statistics

Type	True			False		
	True	Mostly true	Half true	Mostly false	False	Pants-on-fire
6-class	12 %	19%	21%	18 %	18%	12%
4-class	Factual		Incomplete	Manipulative	Hoax	
Statements	6,096					
# Shares	124,215					
# Comments	38,963					
Time period	Jan. 2014~Mar. 2018.					

topics are most susceptible to being used for misinformation purposes. Rony et al. (2017) present a clickbait detection model, based on which they provide a deeper analysis of the topic, headline and impact. These sorts of intervention approaches may benefit from the more fine-grained distinctions provided by our taxonomy and classification models.

3 Dataset

The experiments in this paper are based on two collected datasets.[4] This section introduces how the data was collected and preprocessed.

3.1 SHPT dataset

A recent online report[5] identified the most prominent websites and corresponding Facebook follower counts for several categories of fake news. From this report, we select the most popular website for each type of fake news as listed in Table 1. For trusted news, we sampled articles from reputable news sources as provided by the STICS service (Hoffart et al. 2014).

To account for the spread of fake news on social media, we obtain auxiliary data from Twitter. For this, we first extract the headline of articles, as these are often posted along with the respective URLs. The headline is decomposed into keywords, which are connected with the logical "AND" operator to query Twitter. To avoid noisy results, only headlines with no less than five words are considered.

Although some of the postings thus obtained also contain some user commentary, the majority of them consist of just the headline and a link. We then also crawl the comments appearing in the conversation thread for each news sharing posting. Overall, our data thus consist of three different parts: the original news content, news sharing postings on social media and comment postings on social media. The dataset statistics are given in Table 1.

3.2 PolitiFact dataset

From the PolitiFact site, we crawled 6096 statements from January 2014 through March 2018. For each assessed statement, the site provides an article explaining the pertinent background and details, as well as the rationale for giving the statement its truthfulness rating. It is via these articles that PolitiFact content is typically shared on social media. Hence, we crawl the explanation articles for each statement and again query Twitter via the headline. We also again obtain the associated comment threads, as before for the SHPT dataset. Statistics about this dataset are provided in Table 2.

3.3 Data preprocessing

To eliminate noise and biased results, we preprocess the data as follows:

- We remove hyperlinks appearing in the news article and tweets.

---

[4] https://www.dropbox.com/sh/7mkgd2k85dk391l/AABN6ktTVNWB3P_4uD6xuM5_a?dl=0.

[5] https://www.usnews.com/news/national-news/articles/2016-11-14/avoid-these-fake-news-sites-at-all-costs.

- To avoid classification biases in the news features, we remove relevant sensitive terms from the SHPT dataset, including “bbc”, “onion”, “american”, “natural”, “news”.
- The Stanford CoreNLP tool (Finkel et al. 2005) is used for sentence splitting, part-of-speech labeling, lemmatization and named entity recognition. All words are lemmatized and lower-cased.
- Punctuation marks and most non-character symbols are removed.

4 Method

This section introduces our method for fine-grained classification of both news and statements.

4.1 Features and computations

We devise and study a number of informative features, which are later used to provide a detailed analysis of the data. We describe the kinds of features we propose to investigate and subsequently explain the actual feature computation.

Token-based classification features For a feature type f and a corresponding lexicon L^f, we have a $|L^f|$-dimensional vector $\mathbf{v}_d$ for each document (or statement, tweet) d, in which each factor $\mathbf{v}^f_{d,i}$ is computed via the following equation:

$$\begin{aligned}\mathbf{v}^f_{d,i} &= \mathrm{tfidf}(d,w) \quad w = i^{th} \text{ word} \in L^f \\ f &\in \{\text{allWords}, \text{excludeEntities}, \text{entities}, \\ &\qquad \text{sentiment}, \text{subjectivity}\}\end{aligned} \tag{1}$$

Here, $\mathrm{tfidf}(d,w)$ refers to the TF-IDF weighting of a word w in a document d, which we rely upon due to its effectiveness in selecting salient words with a high importance within a given document.

Named entities While different news domains exhibit various kinds of named entities that are mentioned, we conjecture that named entity statistics may also provide some signal with regard to the truthfulness of the content. We rely on the Stanford NLP tools (Finkel et al. 2005), which emit 12 types of named entities as labels.

Headline The headline of an article plays an important role in attracting the attention of a reader. Certain categories of articles may exhibit specific patterns, such as clickbait headlines.

Sentiment Sentiment polarity cues can be an important signal to distinguish reliable from unreliable content, based on the assumption that unreliable content tends to be more emotional than reliable content. The sentiment feature is based on a widely used lexicon, the extended ANEW (Warriner et al. 2013).

In the extended ANEW dictionary $\mathcal{D}_{\text{sen}}$, each word is rated on a nine-point scale ranging from 1 to 9 with respect to its sentiment intensity. The rated sentiment score is based on several dimensions, but in our experiments, we only consider the sentiment valence attribute. The score from 1 to 9 reflects the polarity and degree of sentiment, from negative to positive. Upon collecting the ratings from human judges, the authors also provide the mean rating μ and standard deviation σ. Hence, we compute the probability that the word’s rating falls exactly at the mean with σ as a weight t for this word’s rating μ. This means that if a rating is with a higher σ, the rating will be less reliable. The weight is computed by the probability density function (PDF) of a normal distribution $t = \frac{1}{\sqrt{2\pi\sigma^2}}$. Finally, we take the sentiment rating for a document d as

$$\sigma(d \mid \mathcal{D}_{\text{sen}}) = \frac{\sum_{w\in d; w\in\mathcal{D}_{\text{sen}}} \mu_w t_w}{\sum_{w\in d; w\in\mathcal{D}_{\text{sen}}} t_w} \tag{2}$$

Subjectivity Another pertinent assumption is that unreliable content tends to use more subjective or extreme words to convey an particular perspective. We thus rely on the MPQA subjectivity lexicon, as used in previous work (Wilson et al. 2005), for subjectivity cues in our experiment.

The subjectivity dictionary ($\mathcal{D}_{\text{sub}}$) merely provides a label (“strongsubj” or “weaksubj”) for each word. Hence, we assign each word w a rating r_w based on its label: If the label of w is “strongsubj”, then $r_w = 1$. Otherwise, $r_w = 0.5$. Based on this, the subjectivity score for document d is computed as:

$$\sigma(d \mid \mathcal{D}_{\text{sub}}) = \frac{\sum_{w\in d; w\in\mathcal{D}_{\text{sub}}} r_w}{\left|\{w \mid w \in d \ \wedge \ w \in \mathcal{D}_{\text{sub}}\}\right|} \tag{3}$$

Credibility history For statements or claims, the speaker or author may have preferences for specific kinds of unreliable contents or styles of presenting claims. In some forums, such as PolitiFact, the community annotates previous statements of speakers by degrees of credibility. Therefore, the credibility history of the speaker may be an interesting feature to aid in detecting unreliable content.

On the statement dataset, we use the historical distribution of different kinds of statements as the credibility feature $\mathbf{c}$ for a speaker s as follows:

$$\mathbf{c}^s_i = \frac{f_i}{\sum_{i=1}^{n} f_i} \qquad n = 4 \ or \ 6 \tag{4}$$

where f_i provides the number of statements from the speaker belonging to the ith category of fact-check ratings. The categories vary between 4-class and 6-class settings. Ultimately, this feature is used for the detection together with the token-based features by concatenation as $[\mathbf{v}_d, \mathbf{c}^s]$.

News and statement sharing in Twitter The number of times an article has been shared on Twitter is a strong signal that reflects on the popularity of the article. Through the timeline, we also get the life cycle information of different kinds of news. In this paper, we first provide an analysis of news sharing distribution and of the relevant timelines. Then, we validate the effectiveness of this feature in the classification task.

Content similarity Within a comment thread about a piece of potentially fake news, we often observe the phenomenon of content drift, as the discussion moves to less closely related and extended topics. We use the content similarity between the original news article and the tweet comments to measure the degree of drift.

Due to the advantage of being trained on Google News data, we rely on the publicly available 300-dimensional *word2vec* embeddings[6] to quantify the content similarity between a news article **A** and the corresponding tweet comment set **B** via cosine similarity. We use the averaged word embedding of a news article or a tweet to represent its semantic content. Although this is a simple method, it still serves as a widely used baseline in many works because of its reasonably strong effectiveness (Dai et al. 2015; Le and Mikolov 2014).

$$\mathrm{sim}(\mathbf{A}, \mathbf{B}) = \cos(E_{\mathbf{A}}, E_{\mathbf{B}})$$
$$E_{\mathbf{A}} = \frac{\sum_{w \in W(\mathbf{A})} e_w}{|W(\mathbf{A})|}, \qquad E_{\mathbf{B}} = \frac{\sum_{w \in W(\mathbf{B})} e_w}{|W(\mathbf{B})|} \tag{5}$$

where $W(\cdot)$ is the set of words in a news article or a set of tweets and e_w denotes the embedding of the word w.

Tweet comments Tweet comments reflect on the social dimension of a potential fake news story. In contrast to simple news sharing, comment tweets are replete with personal opinion and discussion. We include tweet comment features based on the assumption that the social reactions are different for different kinds of articles.

Stance We attempt to characterize the stance that social media users take. This is important on reflecting the potential influence of fake news articles. An article may derive its impact not simply from the mere fact that many readers are discussing it, but based on whether it succeeds in influencing those readers.

We use a pre-trained model (Riedel et al. 2017) to detect the stance of tweet comments. It provides four kinds of stance labels, namely “agree”, “disagree”, “discuss” and “unrelated”. However, due to the high degree of relevance between the news articles and tweet comments, we ignore the “unrelated” stance. We use the agreement score metric to reflect the degree of support among the readers, which is computed as the ratio of the “agree”-stance amount and the total stance amount. To improve its reliability, we ignore the value of the absolute agreement or disagreement stance score.

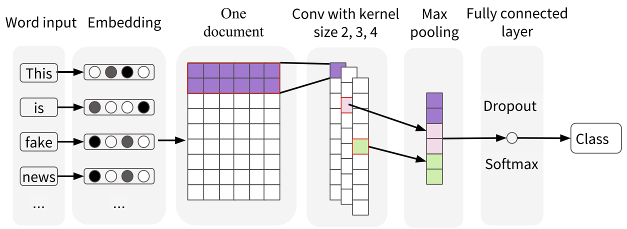


Fig. 3 CNN model representation

4.2 Logistic regression method

Our goal is to learn a model that is interpretable so as to enable detailed analyses. We hence rely on a logistic regression (LR) model to address not only the task of accurately classifying the items, but also to conveniently analyze the contribution of features. We invoke the one-versus-rest strategy for multi-class classification. The configuration of the LR multi-class model is as follows:

- We use fivefold cross-validation for parameter tuning, which involves randomly selecting 20% of the training documents as the validation set and the rest as the genuine training set in each training iteration.
- As the length of articles, statements and tweets is short, the features are very sparse. Therefore, we rely on L2 regularized logistic regression to capture more details.
- The parameter cost C is tuned according to the validation set results, which is supported by the toolkit with the exponential step length. The parameter e of the tolerance as the termination criterion is set as 0.0001.
- A Newton-type method is invoked to optimize the training objective function of the logistic regression. The LR implementation we use is from the LIBLINEAR (Fan et al. 2008) toolkit.

4.3 Deep learning method

We additionally present two deep learning classifiers for long text and headlines, respectively. The two classifiers are based on the concepts of convolutional neural networks (CNN) and long short-term memory (LSTM) networks.

CNN classifier In order to classify long texts such as news articles and a set of tweets, we introduce a CNN-based

---

[6] https://code.google.com/archive/p/word2vec/.

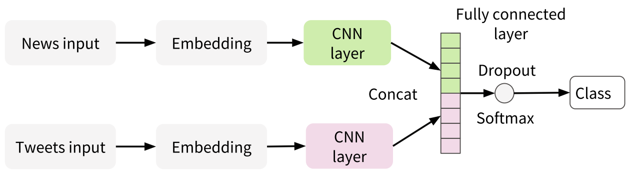


Fig. 4 Model combining news and tweets

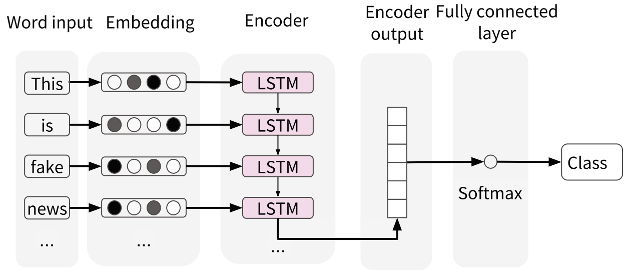


Fig. 5 LSTM model representation

classifier. CNNs are considered more effective at capturing n-gram features and other contextual relationships (Yin et al. 2017; Zhou et al. 2015). As illustrated in Fig. 3, we first convert the long input text into the embedding space as $\mathbf{x}$. Then, the CNN layer processes the input by means of a convolution operation with different kernel sizes, to capture n-gram features. Subsequently, a max-pooling layer down samples the features to reduce their dimensionality. To avoid overfitting, we apply the dropout technique. Finally, the output layer with softmax activation function yields the classification probabilities. The computations for this process are as follows:

$$\mathbf{y}_c = \mathbf{x} \otimes \mathbf{W}_c + \mathbf{b}_c \tag{6}$$

$$\mathbf{h}_c = \mathrm{MaxPooling}(\mathbf{y}_c) \tag{7}$$

$$\mathbf{h}_{\mathrm{drop}} = \mathrm{dropout}(\mathbf{h}_c, p_{\mathrm{drop}}) \tag{8}$$

$$\hat{y} = \mathrm{softmax}(\mathbf{h}_{\mathrm{drop}} + \mathbf{b}_a) \tag{9}$$

Here, $\mathbf{W}_c$ is a learned weight matrix, $\mathbf{b}_c$ is the learned bias term within the hidden layer, and p_{drop} is the dropout probability.

To combine news and comment features, we use a concatenation operation, as shown in Fig. 4. The tweet comments for a news article are also considered as a single document, to which the CNN layer is applied in Fig. 3.

LSTM classifier LSTM networks excel at capturing sequential correlations, especially for short text (Yin et al. 2017). Accordingly, we devise an LSTM classifier for detecting fake news within headlines and individual statements. The network is composed of LSTM cells as in Fig. 5. Each cell maintains a state $\mathbf{c}_t$ and the output $\mathbf{h}_t$ at time t. The current cell state can be fed as the initial state of the cell at next time $t+1$. The cell also comprises three gates: the input gate $\mathbf{i}_t$, forget gate $\mathbf{f}_t$ and output gate $\mathbf{o}_t$. The states and gates are represented as follows:

$$\mathbf{f}_t = \sigma(\mathbf{x}_t \mathbf{W}_f + \mathbf{h}_{t-1} \mathbf{U}_f + \mathbf{b}_f) \tag{10}$$

$$\mathbf{i}_t = \sigma(\mathbf{x}_t \mathbf{W}_i + \mathbf{h}_{t-1} \mathbf{U}_i + \mathbf{b}_i) \tag{11}$$

$$\mathbf{o}_t = \sigma(\mathbf{x}_t \mathbf{W}_o + \mathbf{h}_{t-1} \mathbf{U}_o + \mathbf{b}_f) \tag{12}$$

$$\hat{\mathbf{c}}_t = \tanh(\mathbf{x}_t \mathbf{W}_{cc} + \mathbf{h}_{t-1} \mathbf{U}_{cc} + \mathbf{b}_{cc}) \tag{13}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t \tag{14}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{15}$$

In our model, we use the final output $\mathbf{h}_t$ at time t as the input features for the output layer with softmax activation function.

Setup For classifications based on news articles and tweets, we apply the CNN model plotted in Fig. 3. The LSTM classifier in Fig. 5 is used to classify the news headlines in “SHPT” and statements in “PolitiFact”. For the combined feature of news articles and tweets, we apply the ensemble model in Fig. 4. In “PolitiFact”, we incorporate LSTM and CNN layer to extract the statements and tweets feature, respectively, and then combine both features in the same way plotted in Fig. 4. We adopt pre-trained 100-dimensional embeddings from GloVe (Pennington et al. 2014), which are trained on Wikipedia (as of 2014) and Gigaword 5. The categorical cross-entropy loss is adopted for this multi-class classification task. The model is trained in 500 iterations, and the optimal hyperparameters values are obtained by fivefold cross-validation.

5 Analytics

This section details our findings in the experiments.

5.1 Classification performance analysis

Table 3 provides the fivefold cross-validation results for the accuracy and Macro F_1 score of the classifiers on the SHPT and PolitiFact datasets. For the latter, we consider both the 4-way and 6-way target classification schemes (cf. Table 2). We evaluate the different feature set variants discussed earlier for the LR method: all words, all words excluding entity names, only entity names, only words from the sentiment lexicon, only words from the subjectivity lexicon. In addition, the deep learning method is also evaluated, using all words as features. The results can be summarized as follows:

1. The combination of content (articles or statements) and social media information can improve the prediction

Table 3 Classification accuracy and Macro F_1 score on SHPT and PolitiFact datasets (6-class and 4-class labeling)

Dataset	Input	All words		All words without entities		Entity words only		Sentiment words only		Subjectivity words only		Deep Learning with all words	
		Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
SHPT- 4 class	Headline	0.791	0.789	0.739	0.737	0.440	0.422	0.686	0.683	0.504	0.493	0.902	**0.902**
	Articles	0.975	**0.975**	0.966	0.966	0.857	0.857	0.942	0.942	0.847	0.846	0.959	0.959
	Tweets	0.601	0.574	0.592	0.563	0.493	0.438	0.534	0.501	0.501	0.452	0.627	**0.611**
	Both	0.981	**0.981**	0.975	0.975	0.881	0.881	0.954	0.954	0.871	0.871	0.969	0.969
Politi- 6 class	Statements	0.274	0.259	0.269	0.241	0.238	0.209	0.257	0.239	0.214	0.193	0.335	**0.333**
	Tweets	0.257	0.237	0.256	0.238	0.215	0.183	0.249	0.225	0.236	0.213	0.522	**0.519**
	Both	0.306	0.284	0.311	0.294	0.251	0.219	0.290	0.274	0.249	0.232	0.483	**0.482**
Politi- 4 class	Statements	0.420	0.285	0.413	0.284	0.372	0.295	0.408	0.254	0.303	0.264	0.441	**0.389**
	Tweets	0.339	0.320	0.339	0.324	0.286	0.269	0.332	0.308	0.320	0.300	0.651	**0.597**
	Both	0.458	0.344	0.450	0.354	0.397	0.298	0.434	0.330	0.367	0.297	0.633	**0.595**

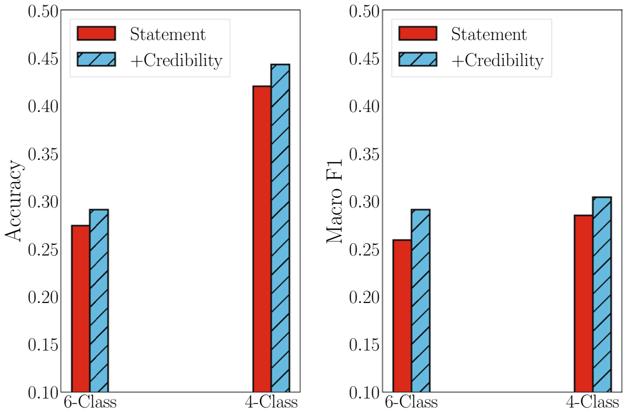
The best performance for each input is bolded considering F_1 score

quality, especially on the PolitiFact dataset, establishing the effectiveness of the tweet comments feature.

2. It is substantially more difficult to classify the individual statements from PolitiFact as opposed to the news articles in the SHPT dataset. There are multiple reasons for this, including that the length of news articles is longer and the content is rich in details.
3. The all-tokens feature version outperforms other alternatives. However, when excluding entities, one can achieve a close level of accuracy, which may also lead to a model with better out-of-domain generalization.
4. We observe that on the SHPT dataset, the tweets feature performs much worse than the news-based features. One reason is that for some news articles, no tweet comments were found on Twitter.
5. The deep learning method outperforms the LR variant on most experimental configurations, except on the SHPT dataset. The reason is that for short text such as headlines and statements, the features are fairly sparse for the LR method.
6. The deep learning method is much more effective than the LR method when applied to tweets. One reason may be that the tweets are concatenated with the document, despite not inherently being as correlated and relevant as the sentences within the news article. In this case, the CNN classifier can detect the pertinent parts within the content, while the LR method may lose focus when all the tweets are included.
7. From the Macro F_1 score results, we observe that the effectiveness of LR decreases slightly when the dataset is imbalanced, especially on "Politi-4 class", while the deep learning classifier is not affected by this. Recall that we adopt the one-versus-rest policy in the multi-class LR classifier for model efficiency. In addition, the F_1 results are acceptable, and this does not impact the feature analysis of each class within the LR classifier.

5.2 Credibility history

We rely on the credibility history feature to classify the statement ratings in the PolitiFact dataset. The results are plotted in Fig. 6. The classifier with only statement content words serves as the baseline, which is labeled as "Statement" in Fig. 6. The classifier combining the statement words and credibility history feature (as introduced in Sect. 4.1) is labeled as "+Credibility". We can observe that the credibility history feature succeeds in improving the classification, but with unsubstantial gains. While authors may exhibit a preference for a specific type of statement, these preferences may evolve over time. Hence, this feature potentially helps the classifier, but with limited contribution.



(a) Accuracy **(b)** Macro F_1

Fig. 6 Classification accuracy and Macro F_1 with and without credibility history feature

Table 4 Top feature words on SHPT news excluding entities

Type	News	Tweets
Satire	Advertisement, here, unemployed, reportedly, add, confirm, press, what, systems, tip, pretty, analyst, announce, sure, tester, whatever	Funny, whatdoyouthink, satire, rt, satirical, actually, humor, movie, joke, funnier, hahaha, guy, laugh, comedy, real, hahah, idk, idiotic, literally
Hoax	Comment, this, statement, be, accord, liberal, president, we, below, do, everybody, recent, presidency, argue, continue, administration	He, she, president, fake, flag, remove, potus, crookedhillary, scale, pray, rating, judge, muslim, enough, ass, welfare, idiot, admiral, tired, click
Propaganda	Relate, republish, writer, article, leftwing, vaccine, senior, leftist, ranger, socalled, dailymailcouk, more, saidhe, editorinchief, rrb	Food, poison, health, vaccine, left, eat, gmo, qanon, organic, remedy, vodka, pharma, chemical, brain, commie, shot, federal, depopulation
Trusted	Copyright, caption, image, panel, external, mr, say, playback, related, getty, share, close, unsupported, programme, link, prof	Tory, realise, bloody, mum, labour, mp, programme, council, minister, yawn, fossil, ff, lady, sort, brexit, pupil, licence, daft, comment, rubbish

Table 5 Top feature words on PolitiFact 6-class statements excluding entities

Rating	Top feature words
True	Resident, last, only, marriage, elect, hold, population, grow, death, sector, rent, county, loophole, think, decline, growth, nearly
Mostly true	World, pension, largest, earn, dollar, twothird, nation, provide, time, together, than, licens, oppose, since, percentage, country
Half true	Investment, add, lower, leadership, agree, large, row, put, access, million, combine, corporation, proposal, under, premium
Mostly false	Preexisting, option, thing, patient, congressman, stop, workforce, teacher, expansion, assault, pass, bank, decide, watch, same
False	Protect, much, road, abortion, story, cause, nothing, deal, essentially, attack, ever, insurance, subject, lie, ground, prove, fix
Pants-on-fire	Arrest, president, order, show, victim, cancel, protester, dead, hurricane, after, fire, cancer, fraud, flag, just, remove, virus, find

Table 6 Top feature words on PolitiFact 4-class statements excluding entities

Type	Statements	Tweets
Factual	Than, since, only, hold, nearly, resident, time, pension, half, state, nation, least, country, licens, loophole, legislation, oppose, last, or	Effective, largest, culture, bern, building, eliminate, trend, party, economy, govt, title, illegally, notice, none, against, economic, capitalism, childish
Incomplete	Investment, add, lower, leadership, agree, large, row, put, access, million, combine, corporation, proposal, under, premium, debate	Near, analysis, server, rating, judge, dozen, money, pal, live, whole, deportation, deductible, convenient, grant, logic, criticize, chance, reform
Manipulative	Preexisting, option, thing, patient, congressman, stop, workforce, teacher, expansion, assault, pass, bank, decide, watch, same, stay	Disqualify, mouthpiece, equally, lie, strange, head, conclusion, tire, killing, release, data, retire, blue, limit, obtain, ss, expert, unbiased, behind, truck
Hoax	Arrest, president, show, protect, order, tell, road, ever, either, cause, attack, liberal, bad, just, after, hurricane, vaccine, presidential, fire	Believe, lie, fake, meme, spew, darn, rumor, interview, wish, obviously, fool, lying, hear, debunk, main, propaganda, shut, birther, satire, falsely, headline

5.3 Linguistic analysis

It is instructive to inspect the models and determine which particular words were most informative in this finer-grained classification. After training the LR model, the linear coefficients for each word serve as interpretable weights that reveal the contribution of each feature toward the classification. Hence, we sort the feature words by their weight in different categories. As named entities may lead to biased models overfitting to particular topics, events, persons etc., we exclude named entities from consideration, so as to obtain a more linguistic analysis. Through the lists of feature words in Tables 4, 5 and 6, we observe that the linguistic styles differ between different kinds of news or statements.

1. The SHPT feature words reflect the writing style of different types of news articles, e.g., trusted news prefers a formal register, while satire and hoax news use more

Table 7 Top feature words of the news headlines on SHPT excluding entities

Type	Top feature words
Satire	Study, report, nation, tips, god, timeline, Americans, pro, fucking, introduce, works, assure, new, offering, linked, authorities, self, users
Hoax	Breaking, lsb, rsb, this, sickening, she, claims, instantly, trump, hillary, muslim, disgusting, furious, muslims, America, unthinkable, wants
Propaganda	Ranger, prepper, medicine, accord, leftwing, health, fisa, toxic, propaganda, remedy, vaccinate, insanity, vaccine, glyphosate, cryptocurrency
Trusted	Briefing, brexit, mp, election, say, mum, jail, boss, accuse, appeal, criticise, pm, probe, pupil, ni, sack, migrant, malaria, inquiry, deal, budget

informal words. For propaganda news, there are more words referring to other articles and websites.

2. Compared with the SHPT dataset, on PolitiFact we find more topical differences between ratings. For instance, pants-on-fire statements are more about disasters, disease and emergencies, while mostly true and half true statements focus more on economy and business topics.
3. Table 6 provides the tweet comments features for PolitiFact statements. We find that on different kinds of statements, the reactions of users are different. For most statements, the users appear able to infer the real intent, e.g., the words “rumor”, “satire”, “propaganda” for hoax statements, versus “mouthpiece”, “lie”, “strange” for manipulative statements. Considering that hoax statements may have propaganda and satirical goals, our taxonomy on the PolitiFact statements still appears to work. In addition, regarding the incomplete statements, the users appear to be more at ease when discussing such statements, which may reflect the goal of propaganda to deceive the readers and steer public opinion in a less overt way.
4. In Table 4, we also notice some tweet words that reveal the intent of news, e.g., “satire”, “humor”, “comedy” and “joke” for satirical news, and “fake” and “click” pointing out the clickbait intent for some hoax news. Combined with the tweets for PolitiFact statements, we conclude that the tweet comments convey more cues regarding the public opinion and reactions to unreliable content. Overall, the combined signal is stronger in discerning the real intent of the authors.

5.4 Headline analysis

The effectiveness of the classification based on news headlines is given in Table 3. Although the prediction accuracy is lower when using only entity, sentiment or subjectivity words, the level of accuracy remains strong in the 4-way classification setting.

In addition, as our main objective is to distinguish the writing styles of headlines, it is instructive to consider top feature words in the headlines, as shown in Table 7. We observe that the headlines of satirical news often refer to other targets, e.g., other “report”, “study”, “timeline”. To gain attention, hoax news use more sensational words such as “breaking”, “sickening”, “disgusting”, “furious”, “unthinkable” in their headlines. For propaganda news, the headlines usually point out the target objects directly, such as “medicine”, “leftwing”, “vaccine”. In summary, different kinds of news exhibit specific characteristics in their headline words.

5.5 Content similarity analysis

It may easily occur for tweet comments on a news topic to drift from the original topic to a more broadly related one. Different types of fake news exhibit different patterns in this regard. The content similarity between news and the corresponding tweets are plotted in Fig. 7. From these observations, we can conclude that “satire” and “trusted” news are more consistent with the tweet comments, showing a small degree of drift from the original content. Thus, on these kinds of news, people tend to really focus on the topic introduced by the original news article. In contrast, “hoax” and “propaganda” articles tend to have more comments with smaller content similarity. This suggests that for these two

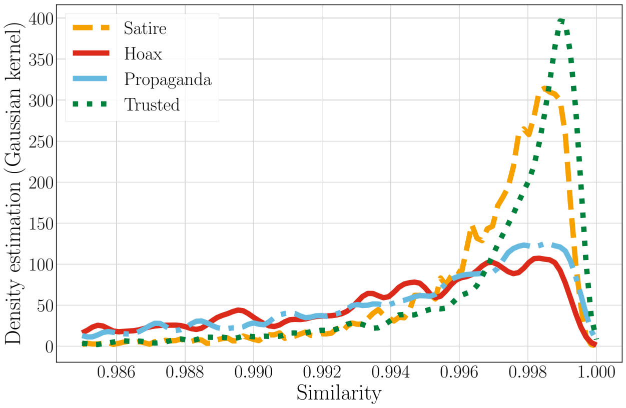


Fig. 7 Content similarity between news and the corresponding Tweets

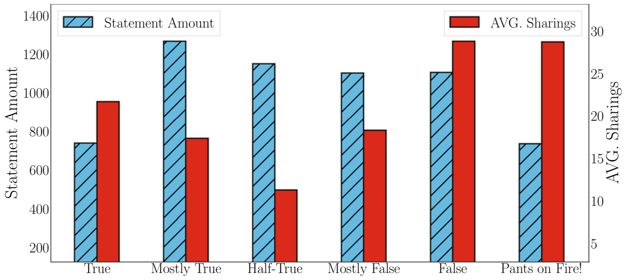


Fig. 8 Statement volume and averaged sharing volume on Twitter for different ratings

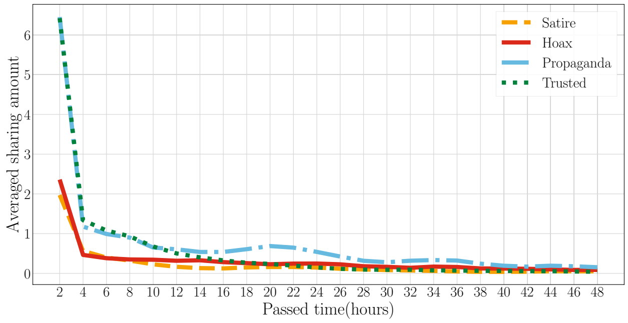
kinds of articles, people often focus on the broader topics, drifting away from the core topic of the original article. As we have analyzed in Sect. 5.3, "propaganda" news tends to refer to other news articles. One of its goals is to steer public attention to a target subject. As a consequence, tweet comments drift more easily to other topics. For "hoax" news, readers focus more on discussions of the content reliability. Thus, more evidence or background information from other related topics is introduced into the discussion.

5.6 News and statement sharing analysis

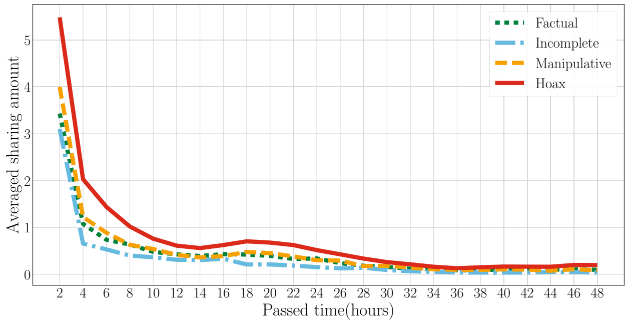
Sharing amount analysis News sharing on social networks is an important aspect reflecting on the popularity and influence of news. As mentioned, we obtained news sharing data by searching for headlines on Twitter. Through these data, we can analyze the sharing of news from multiple facets.

Some basic statistics on PolitiFact are shown in Fig. 8. While one might expect that the obviously true and false statements would have less influence, it turned out that mixed true statements have a lower sharing level. This may be because false and ridiculous statements spark more interest and attention and thus have a greater chance of being shared. True statements being checked are usually hard to believe; however, they are factual. This contrast attracts more user interest, which entails more sharing. Another fact is that there are more mixed true statements. The sharing of SHPT articles is analyzed in Table 1. We observe that satire and trusted news are more popular.

Timeline analysis of sharing After the initial sharing, the volume of sharing changes with time. Different kinds of news exhibit different patterns on the timelines in Fig. 9a and b. On the SHPT dataset, the popularity of the website influences the volume of sharing of each article. To reduce this influence, we use the ratio of the total news sharing volume as a normalizing factor. We find that there is a rapid drop within about half a day. After about 20 h, there is a small burst for propaganda news in SHPT and hoax statements in PolitiFact, suggesting that these kinds of content have a longer life cycle. In addition, the trusted and propaganda news articles show higher levels of sharing, while for statements, the hoaxes are more widely shared.



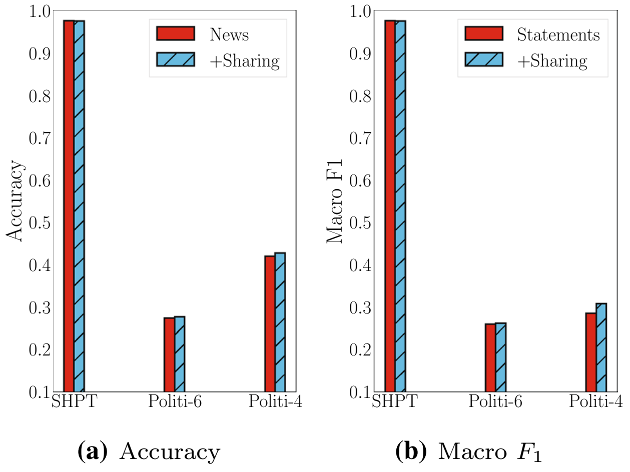
(a) SHPT



(b) PolitiFact 4 class

Fig. 9 Sharing volume over time for SHPT and PolitiFact

Feature effectiveness analysis As the above analysis indicates, different kinds of unreliable content exhibit different patterns for the sharing feature on Twitter. Hence, we validate the effectiveness of this feature on the classification tasks. For each news article or statement, we compute the sharing volume on Twitter. This feature is normalized by min–max normalization. Then, we combine it with the original news article or statement content feature, which is labeled as "+Sharing" in Fig. 10. From the experimental results in Fig. 10, we observe that on the SHPT dataset, the sharing feature cannot improve the classification effectiveness. The reason is that the accuracy is already fairly high based on the news content, leaving little room for improvement. On "PolitiFact-4 class" and "PolitiFact-6 class", the sharing feature improves the effectiveness, but the gain is not substantial. Considering that the feature only contains the sharing amount, the small gain is reasonable, suggesting further research to explore more elaborate features about sharing dynamics. We conclude that the sharing feature has potential for analyzing the dissemination of unreliable content.

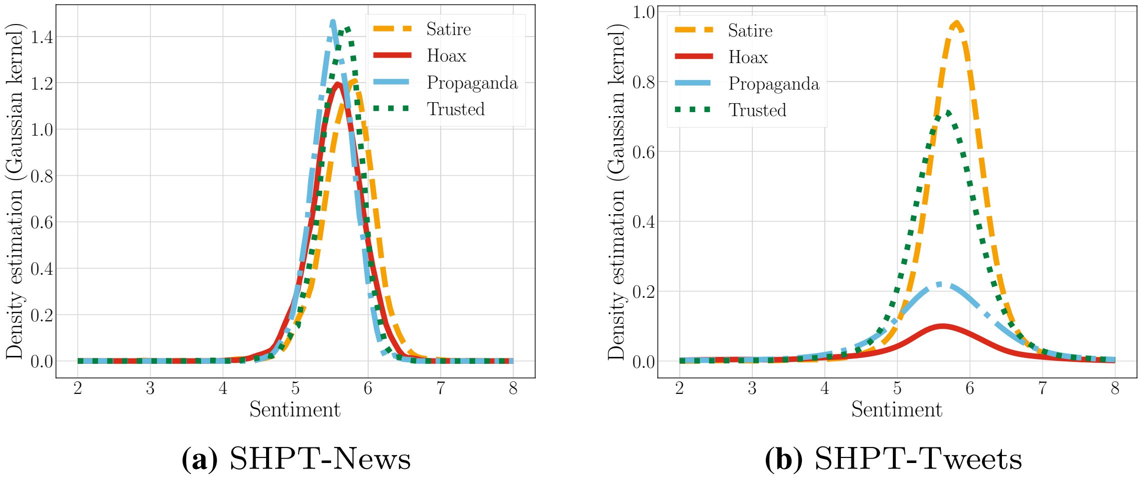


(a) Accuracy **(b)** Macro F_1

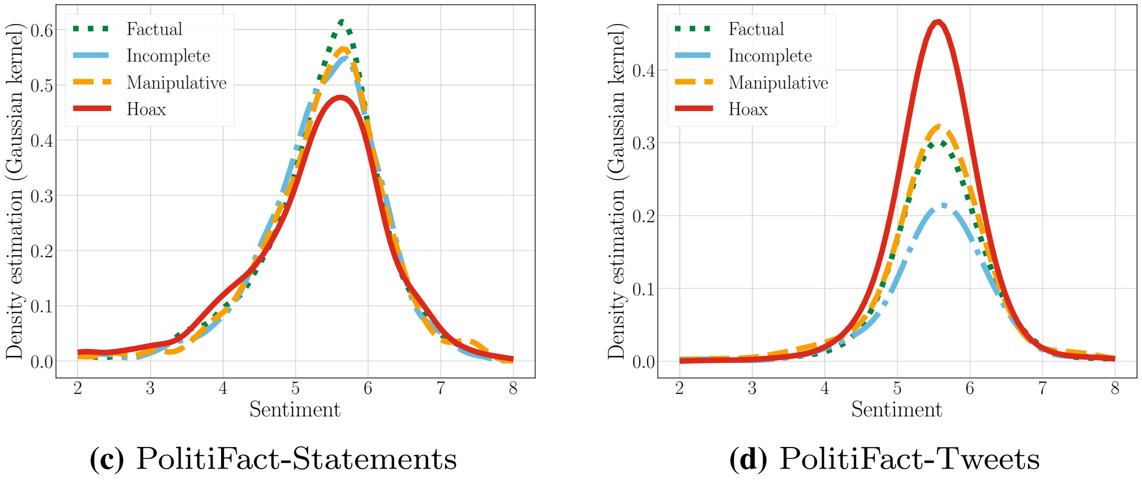
Fig. 10 Classification accuracy and Macro F_1 comparison with and without sharing feature

5.7 Sentiment analysis

The sentiment distribution results are shown in Fig. 11. We find that the sentiment score distributions of both claims and social aspects appear to follow a normal distribution. The sentiment score on its own is not a sufficiently strong signal for classification. However, we still can observe some interesting trends. On SHPT, satire and trusted news are more positive than hoax and propaganda news. In the social commentary, the tweet reactions to satirical news are more positive. The tweet reactions to hoax and propaganda news are more dispersed with respect to the sentiment scores, which means that the reactions to such news vary a lot from positive to negative. By comparing Fig. 11(a) and (c), we see that the sentiment of PolitiFact statements shows a wider distribution than SHPT news. This indicates that statements evoke a wider range of sentiments. The reason for this may be that statements are usually shorter, so the sentiment is crisper. From Fig. 11(d), we observe that the readership on Twitter

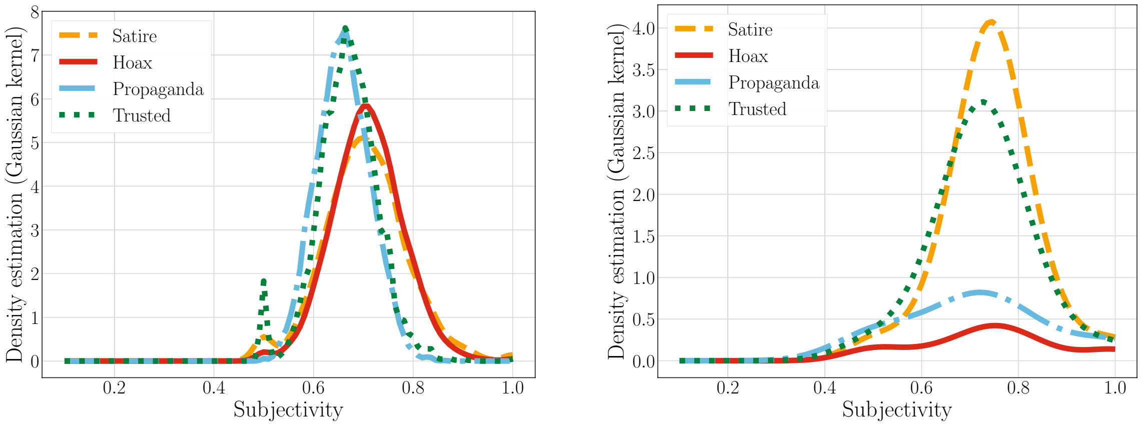


(a) SHPT-News **(b)** SHPT-Tweets

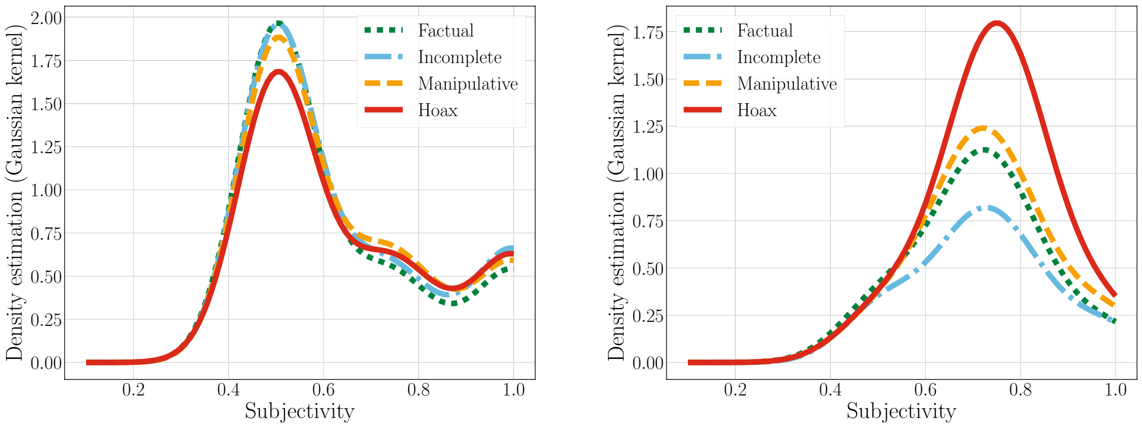


(c) PolitiFact-Statements **(d)** PolitiFact-Tweets

Fig. 11 Sentiment score distribution estimation with Gaussian kernel. Sentiment scores range from 1 to 10



(a) SHPT-News **(b)** SHPT-Tweets



(c) PolitiFact-Statements **(d)** PolitiFact-Tweets

Fig. 12 Subjectivity score distribution estimation with Gaussian kernel. Subjectivity scores range from 0 to 1

responds similarly to different kinds of statements on the sentiment dimension and only the comment diversity differs.

5.8 Subjectivity analysis

The subjectivity distribution results on both datasets are shown in Fig. 12. We find that the subjectivity distribution of SHPT news and PolitiFact statements is different. There are more extremely subjective statements on PolitiFact, which indicates that the spoken statements are more subjective than the written articles. From Fig. 12(a), we find that hoax and satire news are more subjective. The small burst in the plot for trusted news indicates that some trusted news mostly uses objective language. This is in line with our expectation that unreliable content is more subjective and inaccurate. Regarding the tweet reactions, we notice that satirical and hoax news and hoax statements receive more subjective comments. This suggests that on subjective content, the tweet comments also become more subjective.

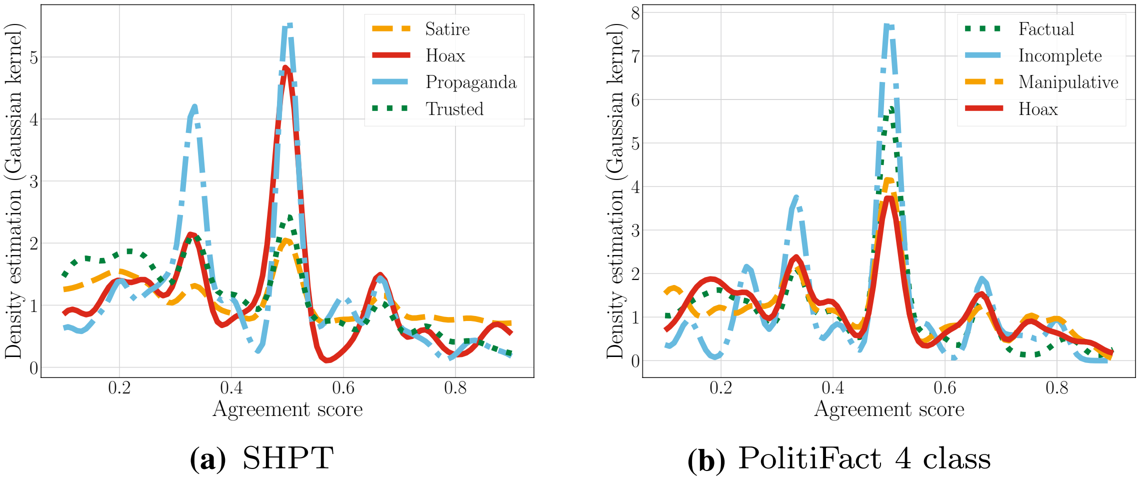
5.9 Stance analysis

The average agreement score and the score distribution are shown in Table 8 and Fig. 13, respectively. The results meet our expectations in that the propaganda and truth-mixed content have more potential impact. Especially, for the PolitiFact statements, incomplete and manipulative statements gain much higher support on social media. On the SHPT dataset,

Table 8 Average agreement score of different types of unreliable content

SHPT	Score	PolitiFact	Score
Trusted	0.384	Factual	0.403
Propaganda	0.427	Incomplete	**0.442**
Satire	**0.445**	Manipulative	0.424
Hoax	0.438	Hoax	0.418

The highest score on each dataset is bolded



(a) SHPT **(b)** PolitiFact 4 class

Fig. 13 Agreement score distribution on SHPT and PolitiFact

the satirical news obtains the highest support ratio. We conjecture that this is because satirical news are not intended to deceive, and readers can assess the reliability of the news; so they tend to post comments with an entertaining tone, in agreement with each other. On both datasets, trusted and factual content yields comparably lower support, because factual content reports on genuine events. Thus, user comments often agree in their stances on factual news.

5.10 Five shades of untruth

Finally, we can summarize our findings with respect to the five shades of untruth, incorporating both news articles and fact-checked statements. In this regard, we have provided a mapping between the news categories and fact-checking ratings in a hierarchical taxonomy. Overall, through the above analytics we can arrive at the following findings with respect to similarities and differences between the two.

1. Factual content still plays an important role, especially on the news. Trusted news articles can lead to prominent sharing and extensive discussion with small topic drift on social media. Such articles are more objective than others and factual statements as well tend to use more accurate words.
2. News articles and statements adopt different ways of achieving propaganda goals. News articles mention target objects directly in the context and headlines, which leads the readers to the targeted topic. The larger topic drift on social media also indicates the use of this trick for propaganda. The creators may take actions seeking to make the propaganda articles available during longer periods of tie to extend the life cycle. Statements on PolitiFact usually achieve this goal via incomplete or manipulative information. By decreasing the deception intent, this kind of statement may succeed in steering the reader's attention away from possible reliability judgments toward the topic of the statements, which is indicated by the comment words and reader stance analysis.
3. Hoax news is more subjective and negative with more attractive headlines and also attracts more diverse comments. Hoax statements are more active and attractive with a longer life cycle and more sharing. Twitter users often are able to recognize the intent of hoax news and statements. Thus, hoax content proves less effective in convincing potential readers. However, it can still lead a hot topic to garner more attention and discussion.
4. For ironical content, satirical news is more popular and more widely endorsed on social media. Both the writing style and tweet reactions are more funny and subjective. As a consequence, the topic drift between the satirical news and tweets is smaller. The headline style also distinguishes itself from that of other kinds of news. While we do not have an irony mapping to the PolitiFact rating system, the experimental results show that the "false" and "pants-on-fire" statements partially overlap with ironical content.

6 Use case study: speaker profiling

Our proposed taxonomy enables users to drill down into deeper levels of misinformation. In this section, we discuss a use case where our model yields such refined insights.

In Sect. 5.2, we have demonstrated the effectiveness of the speaker's credibility feature for content classification. Moreover, the credibility feature itself is beneficial for profiling speakers with regard to different rhetorical styles, intents and ways of conveying biased perspectives. Figure 14 shows an example for this kind of speaker profiles. It compares

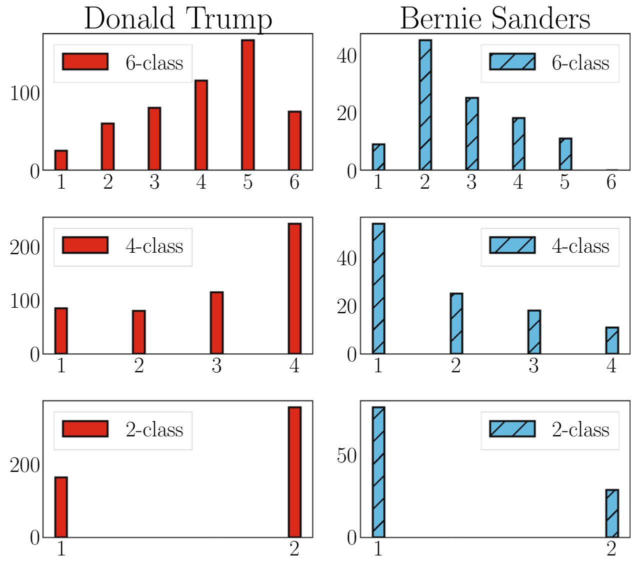


Fig. 14 Comparison of credibility profiles between two speakers, Donald Trump and Bernie Sanders. The *x*-axis is the taxonomy label, and the *y*-axis is the frequency of statements for the respective label. “6-class” includes the 6 ratings from “True” to “Pants-on-fire” corresponding to “1” to “6”. “4-class” and “2-class” are according to Table 2

the credibility labels between the PolitiFact statements by Donald Trump and Bernie Sanders, under three different taxonomies: 6-way, 4-way and the binary 2-way (merely, fact vs. fake). Clearly, the finer-grained taxonomies reveal more interesting observations on the speaking habits of these two politicians—going beyond the usual black-or-white picture. Trump’s statements have a high fraction of propaganda: statements with distorted and misleading context all the way to being manipulative, but they are not necessarily complete fake (i.e., “pants-on-fire”). Sanders has a fair share in this middle ground as well, but the majority of his statements are in the mostly true category—still often with incomplete context and hence not in the fully-true bin. These finer shades cannot be revealed by any binary fact-or-fake classifier.

We conduct the following experiment to explore the performance of our methods for this kind of speaker credibility profiling.

1. We use the PolitiFact dataset and define the credibility profile **c** of speaker s as in Sect. 4.1: $\mathbf{c}_i^s$ is the number of statements from speaker s belonging to the ith category of a taxonomy.
2. We sample 30 speakers as test cases, where each speaker has between 14 to 38 statements, 651 statements in total. These are withheld from the training data for learning classifiers for 6-way, 4-way and 2-way taxonomies. Because of its superior performance in the earlier experiments, we use the deep learning classifier with all words as features combining statements and tweets.
3. The trained classifiers are applied to the test data, and the resulting labels are used to predict the credibility profile $\mathbf{o}^s$ for each speaker. As a measure of the prediction quality, we use the mean squared error (MSE) between $\mathbf{o}^s$ and the ground-truth profile $\mathbf{c}^s$ and compute the macro-averaged MSE over all 30 test speakers. We also report the macro-averaged F1 scores of the classifiers.

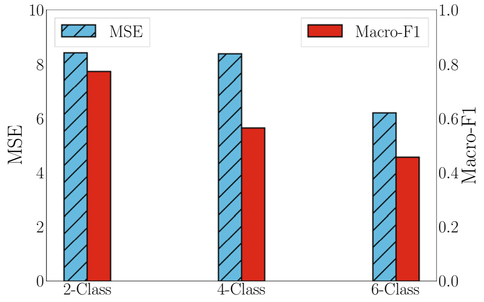


Fig. 15 Performance of credibility profiling under different taxonomies

We show the results of this experiment in Fig. 15. It is obvious that the F1 scores decrease as the number of classes in the taxonomy increases. This is natural and unavoidable, as multi-class learning is inherently more difficult than binary learning. It is remarkable, though, that the MSE in predicting an entire speaker profile decreases with more classes. In other words, fine-grained classifier is better suited for predicting profiles than a simple binary model (at least for this dataset). This underlines the practical benefits of refined taxonomies.

7 Conclusion and outlook

This paper introduces a taxonomic hierarchy to integrate a news categorization scheme and a fact-checking rating system. We devise different kinds of multi-class classifiers over an expressive range of features, including linguistic cues as well as user credibility and news dissemination in social media. In our experiments, deep learning outperforms logistic regression. However, the latter provides better interpretability and more easily supports assessing the impact of features. The presented feature analysis studies linguistic aspects, sentiment and subjectivity cues, the credibility history of users, the stance polarity in tweets and the sharing and spreading of news on Twitter.

As a word of caution, we would like to emphasize that all automated classifiers are merely proxies to assess the credibility of news and claims. The machine learning predictions should not be over-interpreted as “truth finding”.

In fact, the motivation of our work has been to move away from the black-or-white picture of binary predictions to a more refined and informative analysis of different shades of misinformation.

There are many opportunities for future work. Reactions on social media could play a more important role in the analysis and understanding of the intents behind misinformation, going beyond our dissemination features. Another important topic for future research is coping with adversarial behavior: can the authors of fake news and manipulative claims outsmart state-of-the-art classifiers by adapting to its features? How can such attacks in turn be countered by a learning-based tool? We plan to investigate these issues in our ongoing endeavor on understanding and combating misinformation.

Acknowledgements The authors wish to acknowledge the support provided by the National Natural Science Foundation of China (61503217) and China Scholarship Council (2016062-20187). Gerard de Melo's research is in part supported by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under Contract No. W911NF-17-C-0098. Gerhard Weikum's work is partly supported by the ERC Synergy Grant 610150 (imPACT). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Berghel H (2017a) Alt-news and post-truths in the "fake news" era. Computer 50(4):110–114

Berghel H (2017b) Lies, damn lies, and fake news. Computer 50(2):80–85

Bourgonje P, Schneider JM, Rehm G (2017) From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In: EMNLP workshop: natural language processing meets journalism, pp 84–89

Campan A, Cuzzocrea A, Truta TM (2017) Fighting fake news spread in online social networks: actual trends and future research directions. In: IEEE international conference on big data, IEEE

Conroy NJ, Rubin VL, Chen Y (2015) Automatic deception detection: methods for finding fake news. JASIST 51(1):1–4

Dai AM, Olah C, Le QV (2015) Document embedding with paragraph vectors. arXiv preprint arXiv:150707998

Del Vicario M, Quattrociocchi W, Scala A, Zollo F (2018) Polarization and fake news: early warning of potential misinformation targets. arXiv preprint arXiv:180201400

DiFranzo D, Gloria-Garcia K (2017) Filter bubbles and fake news. XRDS: Crossroads. ACM Mag Stud 23(3):32–35

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9(Aug):1871–1874

Farajtabar M, Yang J, Ye X, Xu H, Trivedi R, Khalil E, Li S, Song L, Zha H (2017) Fake news mitigation via point process based intervention. arXiv preprint arXiv:170307823

Finkel JR, Grenager T, Manning C (2005) Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL, pp 363–370

Fourney A, Racz MZ, Ranade G, Mobius M, Horvitz E (2017) Geographic and temporal trends in fake news consumption during the 2016 US presidential election. In: CIKM, ACM, pp 2071–2074

Hoffart J, Milchevski D, Weikum G (2014) STICS: searching with strings, things, and cats. In: SIGIR, pp 1247–1248

Jang SM, Kim JK (2018) Third person effects of fake news: fake news regulation and media literacy interventions. Comput Hum Behav 80:295–302

Jin Z, Cao J, Zhang Y, Luo J (2016) News verification by exploiting conflicting social viewpoints in microblogs. AAAI

Kim J, Tabibian B, Oh A, Schölkopf B, Gomez-Rodriguez M (2017) Leveraging the crowd to detect and reduce the spread of fake news and misinformation. arXiv preprint arXiv:171109918

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning, pp 1188–1196

Li Y, Gao J, Meng C, Li Q, Su L, Zhao B, Fan W, Han J (2015) A survey on truth discovery. SIGKDD Explor 17(2):1–16

Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Popat K, Mukherjee S, Strötgen J, Weikum G (2017) Where the truth lies: explaining the credibility of emerging claims on the web and social media. In: WWW, pp 1003–1012

Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B (2017) A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:170205638

Rashkin H, Choi E, Jang JY, Volkova S, Choi Y (2017) Truth of varying shades: analyzing language in fake news and political fact-checking. In: EMNLP, pp 2931–2937

Rath B, Gao W, Ma J, Srivastava J (2017) From retweet to believability: Utilizing trust to identify rumor spreaders on twitter. In: ASONAM, ACM, pp 179–186

Riedel B, Augenstein I, Spithourakis GP, Riedel S (2017) A simple but tough-to-beat baseline for the fake news challenge stance detection task. arXiv preprint arXiv:170703264

Rony MMU, Hassan N, Yousuf M (2017) Diving deep into clickbaits: who use them to what extents in which topics with what effects? In: ASONAM, ACM, pp 232–239

Rubin VL, Chen Y, Conroy NJ (2015) Deception detection for news: three types of fakes. JAIST 52(1):1–4

Ruchansky N, Seo S, Liu Y (2017) CSI: a hybrid deep model for fake news detection. In: CIKM, ACM, pp 797–806

Shu K, Sliva A, Wang S, Tang J, Liu H (2017a) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newslett 19(1):22–36

Shu K, Wang S, Liu H (2017b) Exploiting tri-relationship for fake news detection. arXiv preprint arXiv:171207709

Singhania S, Fernandez N, Rao S (2017) 3HAN: a deep neural network for fake news detection. In: ICONIP, Springer, pp 572–581

Spivey MJ (2017) Fake news and false corroboration: interactivity in rumor networks. In: COGSCI, pp 3229–3234

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151

Wang L, Wang Y, de Melo G, Weikum G (2018) Five shades of untruth: finer-grained classification of fake news. In: ASONAM, IEEE, pp 593–594

Wang WY (2017) "Liar, liar pants on fire": a new benchmark dataset for fake news detection. ACL. https://doi.org/10.18653/v1/P17-2067

Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 english lemmas. Behav Res Methods 45(4):1191–1207

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: EMNLP, pp 347–354

Wu L, Liu H (2018) Tracing fake-news footprints: characterizing social media messages by how they propagate. In: ICWSDM, ACM, pp 637–645

Yin W, Kann K, Yu M, Schütze H (2017) Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:170201923

Yu PD, Tan CW, Fu HL (2017) Rumor source detection in finite graphs with boundary effects by message-passing algorithms. In: ASONAM, ACM, pp 86–90

Zhou C, Sun C, Liu Z, Lau F (2015) A C-LSTM neural network for text classification. arXiv preprint arXiv:151108630