Five Shades of Untruth: Finer-Grained Classification of Fake News

Liqiang Wang[1,2], Yafang Wang[2,(✉)], Gerard de Melo[3], Gerhard Weikum[1]
[1]*Max Planck Institute for Informatics, Saarbrücken, Germany*
[2]*Department of Computer Science, Shandong University, Jinan, China*
[3]*Department of Computer Science, Rutgers University, New Brunswick, USA*
wanglq1989@gmail.com, yafang.wang@sdu.edu.cn, gdm@demelo.org, weikum@mpi-inf.mpg.de

***Abstract*—Prior work on algorithmic truth assessment on unreliable content, has mostly pursued binary classifiers – factual vs. fake – and disregarded the finer shades of untruth. On the other hand, manual analysis of questionable content has proposed a more fine-grained classification: distinguishing between hoaxes, irony and propaganda, or the six-way rating by the PolitiFact community. In this paper, we present a principled approach to capture these finer shades in automatically assessing and classifying news articles and claims. We systematically explore a variety of signals from both news and social media, and give an analysis of the underlying features.**

***Index Terms*—fake news, unreliable content, social media, fine-grained classification**

I. Introduction

A recent large-scale study of content spreading in Twitter has shown that fake news is disseminated substantially faster, farther, deeper and more broadly than reliable content on comparable topics [1]. The ability of fake news and doubtful claims to outpace serious reporting and verified facts gives them an undue advantage in influencing public opinions. This big societal problem has motivated researchers to develop largely automated methods for assessing the truth of news and statements, leading to tools for fact checking, credibility assessment and trust analysis (see, e.g., [2]–[5]). These methods are based on a variety of powerful data mining and machine learning techniques. However, this prior work only focuses on binary classification: factual or fake.

Fake news typically comes with a specific intent, such as for business profit or political purposes. Rashkin et al. distinguishes *satire*, *hoax*, *propaganda* and *trusted news* by a new taxonomy SHPT [4]. The most reputed online community for manual checking and assessing claims, PolitiFact, uses a six-way "Truth-O-Meter" rating system with labels *true*, *mostly true*, *half true*, *mostly false*, *false* and *pants-on-fire*. To reflect such finer-grained classifications, we devised a taxonomic hierarchy in Fig. 1. It captures both the SHPT scheme and the PolitiFact ratings by mapping their labels into our tree, and this leads to five major categories of fakeness: *factual*, *propaganda*, *hoax* and *irony*, as well as two refinements of propaganda into *incomplete* context and *manipulative* statements.

Based on this new taxonomy, we develop a hierarchical classifier that labels doubtful news or statements with one of our five "shades of untruth". In contrast to prior work, we tap on kinds of signals from social media for both classification and analysis. The paper's salient contributions are: 1) We propose a new taxonomy to incorporate both fake news intents and claim ratings. 2) We present a method for fine-grained classification of questionable news and statements, harnessing features from social-media contents.

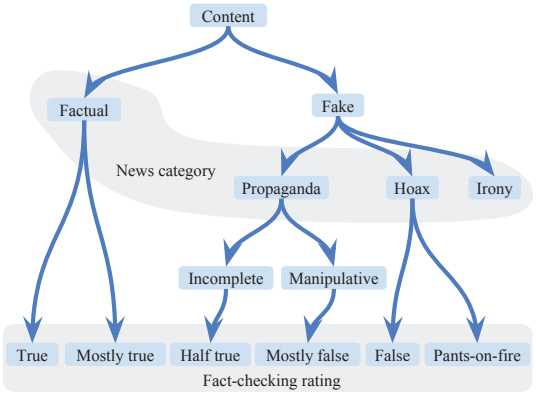

Fig. 1. Classification hierarchy of fake content.

II. Dataset

SHPT. In Table I, we list the sources for different kinds of news. For trusted news, we sampled articles from BBC News provided by the STICS service [6]. 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, then decompose it into keywords, finally connect them with logical "AND" operator to query Twitter. To avoid noisy results, only headlines with no less than 5 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 as a news sharing. We thus crawl the comments appearing in the conversation thread for each news sharing.

TABLE I
SHPT Datasets Statistics

Type	Source	Docs	Shares	Comments	Date
Satire	The Onion	5,000	1,800,295	578,433	Aug. 2013 ∼ Mar. 2018
Hoax	American News	5,000	109,228	14,371	Feb. 2016 ∼ Mar. 2018
Propaganda	Natural News	5,000	230,352	15,315	May. 2017 ∼ Mar. 2018
Trusted	BBC News	5,000	2,124,903	596,940	Aug. 2016 ∼ Mar. 2018

PolitiFact. For each assessed statement, the PolitiFact site provides an article explaining the pertinent background and

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

details. 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 given in Table II.

TABLE II
POLITIFACT DATASETS 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	Date	Jan. 2014 $\sim$ Mar. 2018.

III. METHOD AND EXPERIMENT

A. Feature Categories

Named Entities. While different news domains differ in the 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 [7], 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 Lexicon. 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 [8].

Subjectivity Lexicon. 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 [9] for subjectivity cues in our experiment.

B. Feature Computation

For a feature type f and a corresponding lexicon L^f, we have a $\left|L^f\right|$-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:

$$\begin{gathered}\mathbf{v}^f_{d,i} = \mathrm{tfidf}(d,w) \quad w = i^{th} \text{ word} \in \ L^f \\ f \in \{\mathrm{allWords}, -\mathrm{Entities}, \mathrm{Entities}, \mathrm{Sent.}, \mathrm{Subj.}\}\end{gathered} \tag{1}$$

The features include: all tokens, words excluding entities ("–Entities"), entities only, sentiment lexicon and subjectivity lexicon. 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. Then we rely on a logistic regression (LR) model with one-versus-rest strategy for multi-class classification. For the experiment we use 5-fold cross validation, L2 regularization and a Newton-type solver. The tolerance e as the termination criterion is set as 0.0001.

C. Classification Performance Analysis

Table III gives the obtained accuracies of the classifiers on both the SHPT and PolitiFact datasets. For the latter, we consider both the 4-way and 6-way target classification scheme (cf. Table II). We evaluate the different feature set variants discussed earlier. The results can be summarized as follows:

1) The combination of content (articles or statements) and social media information can improve the prediction quality, especially on the PolitiFact dataset, which indicates the effectiveness of the tweet comments feature.
2) It is substantially more difficult to classify the individual statements in 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, the other features can give acceptable performance and reduce the feature dimension significantly at the same time.
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.

TABLE III
CLASSIFICATION ACCURACY ON SHPT AND POLITIFACT DATASETS (6 CLASS AND 4 CLASS LABELING).

Dataset	Input	All Tokens	–Entities	Entities	Sent.	Subj.
SHPT-4 class	Headline	**0.791**	0.739	0.440	0.686	0.504
	Articles	**0.975**	0.966	0.857	0.942	0.847
	Tweets	**0.601**	0.592	0.493	0.534	0.501
	Both	**0.981**	0.975	0.881	0.954	0.871
Politi-6 class	Statements	**0.274**	0.269	0.238	0.257	0.214
	Tweets	**0.257**	0.256	0.215	0.249	0.236
	Both	0.306	**0.311**	0.251	0.290	0.249
Politi-4 class	Statements	**0.420**	0.413	0.372	0.408	0.303
	Tweets	**0.339**	**0.339**	0.286	0.332	0.320
	Both	**0.458**	0.450	0.397	0.434	0.367

IV. ACKNOWLEDGEMENTS

The authors wish to acknowledge the support provided by the National Natural Science Foundation of China (61503217) and China Scholarship Council (201606220187). Gerard de Melo's research is funded in part by ARO grant W911NF-17-C-0098 (DARPA SocialSim).

REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
[2] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *SIGKDD Explorations*, vol. 17, no. 2, pp. 1–16, 2015.
[3] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social media," in *WWW*, 2017, pp. 1003–1012.
[4] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *EMNLP*, 2017, pp. 2931–2937.
[5] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," in *ACL*, 2017, pp. 422–426.
[6] J. Hoffart, D. Milchevski, and G. Weikum, "STICS: searching with strings, things, and cats," in *SIGIR*, 2014, pp. 1247–1248.
[7] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *ACL*, 2005, pp. 363–370.
[8] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.
[9] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *EMNLP*, 2005, pp. 347–354.