Department

Supervised Learning for Fake News Detection

Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto
Universidade Federal de Minas Gerais

Editor: Erik Cambria, Nanyang Technological University, Singapore

***Abstract*—A large body of recent works has focused on understanding and detecting fake news stories that are disseminated on social media. To accomplish this goal, these works explore several types of features extracted from news stories, including source and posts from social media. In addition to exploring the main features proposed in the literature for fake news detection, we present a new set of features and measure the prediction performance of current approaches and features for automatic detection of fake news. Our results reveal interesting findings on the usefulness and importance of features for detecting false news. Finally, we discuss how fake news detection approaches can be used in the practice, highlighting challenges and opportunities.**

Social media systems have been dramatically changing the way news is produced, disseminated, and consumed, opening unforeseen opportunities, but also creating complex challenges. A key problem today is that social media has become a place for campaigns of misinformation that affect the credibility of the entire news ecosystem.

A unique characteristic of news on social media is that anyone can register as a news publisher without any upfront cost (e.g., anyone can create a Facebook page claiming to be a newspaper or news media organization). Consequently, not only traditional news, corporations are increasingly migrating to social media (https://www.comscore.com/Insights/Blog/Traditional-News-Publishers-Take-Non-Traditional-Path-to-Digital-Growth). Along with this transition, not surprisingly, there are growing concerns about fake news publishers posting "fake" news stories, and often disseminating them widely using "fake" followers.[1] As the extensive spread of fake news can have a serious negative impact on individuals and society, the lack of scalable fact checking strategies is especially worrisome.

Not surprisingly, recent research efforts are devoted not only to better comprehend this phenomenon[1] but also to automatize the detection of fake news.[2,3,4] While a fully automated approach for the fake news problem can be quite

Digital Object Identifier 10.1109/MIS.2019.2899143
Date of current version 3 May 2019.

Published by the IEEE Computer Society

controversial and is still open for debate, a pertinent research question is: *What is the prediction performance of current approaches and features for automatic detection of fake news?*

Most of the existing efforts in this space are concurrent work, which identify recurrent patterns on fake news after they are already disseminated, or propose new features for training classifiers, based on ideas that have not been tested in combination. Thus, it is difficult to gauge the potential that supervised models trained from features proposed in recent studies have for detecting fake news. This paper briefly surveys existing studies on this topic, identifying the main features proposed for this task. We implement these features and test the effectiveness of a variety of supervised learning classifiers when distinguishing fake from real stories on a large, recently released and fully labeled dataset. Finally, we discuss how supervised learning models can be used to assist fact-checkers in evaluating digital content and reaching warranted conclusions.

FEATURES FOR FAKE NEWS DETECTION

Most of the existing efforts to detect fake news propose features that leverage information present in a specific dataset. In contrast, we use a recently released dataset that allows us to implement most of the proposed features explored in previous works.[5] It consists of 2282 BuzzFeed news articles related to the 2016 U.S. election labeled by journalists and enriched with comments associated with the news stories as well as shares and reactions from Facebook users.

In this paper, we discarded stories labeled as "non factual content" (12%), and merged those labeled as "mostly false" (4%) and "mixture of true and false" (11%) into a single class, henceforth referred as "fake news." The remaining stories correspond to the "true" portion (73%). The rationale is that stories that mix true and false facts may represent attempts to mislead readers. Thus, we focus our analysis on understanding how features can be used to discriminate true and fake news.

On a coarse-grained level, features for fake news detection can be roughly categorized as follows: 1) features extracted from news content (e.g., language processing techniques); 2) features extracted from news source (e.g., reliability and trustworthiness); and 3) features extracted from environment (e.g., social network structure). Next, we briefly survey previous efforts, describing existing features and how we implemented them.

Textual Features consist of the information extracted from the news text, including the text body, the headline, and the text message used by the news source. For news articles embedded in images and videos, we applied image processing techniques for extracting the text shown on them. In total, we evaluated 141 textual features. Features were grouped in sets, which are described next.

1) *Language Features (Syntax):* Sentence-level features, including bag-of-words approaches, "n-grams" and part-of-speech (POS tagging) were explored in previous efforts as features for fake news detection.[2,6] Here, we implemented 31 features from this set including number of words and syllables per sentence as well as tags of word categories (such as noun, verb, adjective). In addition, to evaluate writers' style as potential indicators of text quality, we also implemented features based on text readability.
2) *Lexical Features:* Typical lexical features include character and word-level signals,[7,6] such as amount of unique words and their frequency in the text. We implemented linguistic features, including number of words, first-person pronouns, demonstrative pronouns, verbs, hashtags, all punctuations counts, etc.
3) *Psycholinguistic Features:* Linguistic Inquiry and Word Count (LIWC)[8] is a dictionary-based text mining software whose output has been explored in many classification tasks, including fake news detection.[4] We use its latest version (2015) to extract 44 features that capture additional signals of persuasive and biased language.
4) *Semantic Features:* There are features that capture the semantic aspects of a text[2,3] are useful to infer patterns of meaning from data.[9] As part of this set of features, we consider the toxicity score obtained from Google's API (https://www.perspectiveapi.com/#/). The API uses machine learning models to quantify the extent to which a text (or comment, for instance) can be perceived as "toxic." We did

not consider strategies for topic extraction since the dataset used in this paper was built based on news articles about the same topic or category (i.e., politics).

5) *Subjectivity:* Using TextBlob's API (http://textblob.readthedocs.io/en/dev/), we compute subjectivity and sentiment scores of a text as explored in previous efforts.[4]

News Source Features consist of information about the publisher of the news article. To extract these features, we first parsed all news URLs and extracted the domain information. When the URL was unavailable, we associated the official URL of news outlet with news article. Therefore, we extract eight (eight) indicators of political bias, credibility and source trustworthiness, and use them as detailed next. Moreover, in this category, we introduce a new set composed of five features, called domain localization (see below).

1) *Bias:* The correlation between political polarization and spread of misinformation was explored in previous studies.[10] In this paper, we use the political biases of news outlets from the BuzzFeed dataset as a feature.
2) *Credibility and Trustworthiness:* In this feature set, we introduce seven new features to capture aspects of credibility (or popularity) and trustworthiness of domains. We collect, using Facebook's API (https://developers.facebook.com), user engagement metrics of Facebook pages that published news articles (i.e., "page talking about" count and "page fan" count). Then, we use the Alexa's API to get the relative position of news domain on the Alexa Ranking (https://www.alexa.com). Furthermore, using this same API, we collect Alexa's top 500 newspapers. Based on the intuition that some unreliable domains may try to disguise themselves using domains similar to those of well-known newspapers, we define the dissimilarity between domains from the Alexa ranking and news domains in our dataset (measured by the minimum edit distance) as features. Finally, we use indicators of low credibility of domains compiled[11] as features.
3) *Domain Location:* Ever since creating fake news became a profitable job, some cities have become famous because of residents who create and disseminate fake news (https://www.bbc.com/news/magazine-38168281). In order to exploit the information that domain location could carry, a pipeline was built to take each news website URL and extract new features, such as IP, latitude, longitude, city, and country. First, for each domain, the corresponding IP was extracted using the trace route tool. Then, the ipstack API was used to retrieve the location features. Although localization information (i.e., IP) has been previously used in works on bots or spam detection, to the best of our knowledge, there are no works that leverage these data in the context of fake news detection.

Environment Features consist of statistics of user engagement and temporal patterns from social media (i.e., Facebook). These features have been extensively used in previous efforts,[12] especially to better understand the phenomenon of fake news.[13] Next, we detail the 21 features from this category.

1) *Engagement:* We consider number of likes, shares, and comments from Facebook users. Moreover, we compute the number of comments within intervals from publication time (900, 1800, 2700, 3600, 7200, 14400, 28 800, 57 600 and 86 400 s), summing up to 12 features.
2) *Temporal Patterns:* Finally, to capture temporal patterns from user commenting activities, we compute the rate at which comments are posted for the same time windows defined before.

CLASSIFICATION RESULTS

We evaluate the discriminative power of the previous features using several classic and state-of-the-art classifiers, including *k*-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forests (RF), Support Vector Machine with RBF kernel (SVM), and XGBoost (XGB). Given that we used hand-crafted features, there was no need to include a neural network model in the comparison since it would only associate weights with the features, rather than find new ones.

Table 1. Results obtained for different classifiers w.r.t AUC and F1 score.

Classifier	AUC	F1
KNN	0.80±0.009	0.75±0.008
NB	0.72±0.009	0.75±0.001
RF	**0.85±0.007**	**0.81±0.008**
SVM	0.79±0.030	0.76±0.019
XGB	**0.86±0.006**	**0.81±0.011**

RF and XGB performed best.

We measure the effectiveness of each classifier w.r.t. the area under the ROC curve (AUC) and the Macro F1 score. In this case, the resulting AUC is the probability that a model will rank a randomly chosen fake news higher (more false) than a randomly chosen news article. The AUC is especially relevant for fake news detection since the decision threshold can be used to control the tradeoff between true and false positive rates. The F1 score combines precision and recall per class in a single metric and the Macro F1 score provides the overall performance of the classifier.

We compute 95% confidence intervals for the mean AUC and F1 by performing a fivefold split between training and test set, repeated ten times with different shuffled versions of the original dataset (a total of 50 runs). Table 1 shows the empirical results obtained from the fitted models using all features previously described.

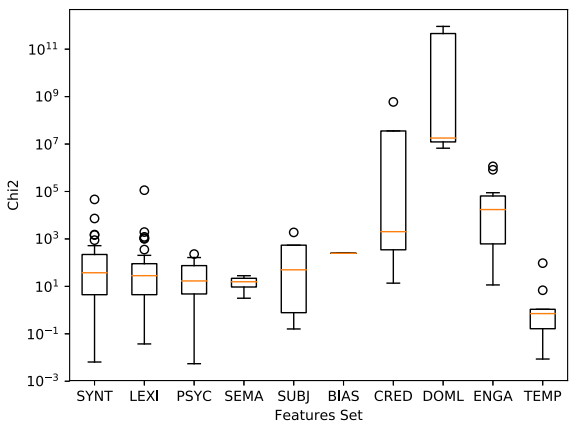

Figure 1. ROC curve for the XGboost classifier. For BuzzFace, it is possible to correctly classify almost all of fake news with only 40% of false positive rate.

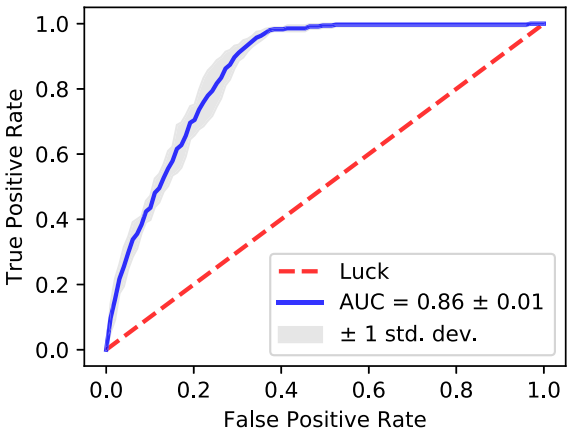

Figure 2. Chi-Square feature importance.

For each classifier, we learn a model from a set of previously labeled (i.e., preclassified) data, and then use it to classify new (unseen) news articles into "fake" or "not fake." The best results were obtained by RF and XGB classifiers, statistically tied with 0.85 ($\pm$ 0.007) and 0.86 ($\pm$ 0.006) for AUC, respectively.

Moreover, inspecting the ROC curve for XGB (see Fig. 1), we observe that it is possible to choose a threshold so as to correctly classify almost all of fake news (true positive rate $\approx$ 1), while misclassifying 40% of the true news (false positive rate $\approx$ 0.4). This can be useful, especially in assisting fact checkers to identify stories that are worth investigating. Finally, we assessed the relative power of the selected attributes in discriminating each class from the other by ranking features from each set based on X^2 (Chi Squared). Fig. 2 shows the results. Although all feature sets have some discriminatory power, there are some of them (e.g., credibility and localization of news sources, and news engagement) that can be more useful to improve the performance of models for fake news detection.

FAKE NEWS DETECTION IN PRACTICE

Fact checking is a damage control strategy that is both essential and not scalable. It might be hard to take out the human component out of the picture any time soon, especially if these news regard sensitive subjects such as politics. In the case of social networks and search engines, predictions made by models for fake news detection could be used internally to limit

the audience of news stories likely to be fake. This is why automatic labeling of news stories raises so many questions about fairness and algorithm transparency, suggesting that it is likely that the final call will still depend on an expert at the end point for a long time.

On the bright side, automatic fake news detection could be used by fact checkers as an auxiliary tool for identifying content that is more likely to be fake. Our results show that the prediction performance of proposed features combined with existing classifiers has a useful degree of discriminative power for detecting fake news. Our best classification results can correctly detect nearly all fake news in our data, while misclassifying about 40% of true news, which is already sufficient to help fact checkers. In this context, providing explanations that supported the algorithm's output is crucial. For example, a certain story was considered false because it was posted by new newspaper hosted in the same IP address than a known blacklisted fake news source. Additionally, this kind of approach requires a continual pipeline where more stories get labeled each day and are, in turn, fed back to the models. Rather than verifying only the most suspicious stories, an active learning solution can be put in place, so that the model can also indicate which stories should be investigated in order to improve its prediction performance. More importantly, fake news is a relatively recent problem and the cost to label large datasets is still very high. In the future, larger volumes of labeled data will enable us to explore other techniques such as deep learning and push the boundaries of prediction performance.

ACKNOWLEDGMENTS

This work was supported in part by Google, CAPES, MASWeb (Grant FAPEMIG/PRONEX APQ-01400-14), CNPq, and Fapemig.

REFERENCES

1. D. M. J. Lazer *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
2. N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in *Proc. Annu. Meeting Assoc. Inf. Sci. Technol.*, 2015, pp. 1–4.
3. W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 422–426.
4. S. Volkova, K. Shaffer, J. Jang Yea, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 647–653.
5. G. Santia and J. Williams, "BuzzFace: A news veracity dataset with facebook user commentary and egos," in *Proc. 12th Int. AAAI Conf. Web Soc. Media*, 2018, pp. 531–540.
6. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.
7. C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 675–684.
8. J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," Mahway: Lawrence Erlbaum Associates, vol. 71, 2001.
9. E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov./Dec. 2017.
10. F. N. Ribeiro, L. Henrique, F. Benevenuto, A. Chakraborty, J. Kulshrestha, M. Babaei, and K. P. Gummadi, "Media bias monitor: Quantifying biases of social media news outlets at large-scale.," in *Proc. of the Twelfth International AAAI Conference on Web and Social Media*, 2018, pp 290–299.
11. C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," 2017, *arXiv:1707.07592*.
12. M. Ebrahimi, A. H. Yazdavar, and A. Sheth, "Challenges of sentiment analysis for dynamic events," *IEEE Intell. Syst.*, vol. 32, no. 5, pp. 70–75, Sep./Oct. 2017.
13. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

Julio C. S. Reis is currently working toward the PhD degree in computer science at the Universidade Federal de Minas Gerais, Brazil. Contact him at julio.reis@dcc.ufmg.br.

André Correia is currently working toward the B.Sc. degree in information systems at the Universidade Federal de Minas Gerais, Brazil. His main interest is applied machine learning. Contact him at andrecorreia.dcc@gmail.com.

Fabrício Murai is an assistant professor in the Computer Science Department, Universidade Federal de Minas Gerais, Brazil. His research lies in the application of mathematical modeling, statistics and machine learning to informational and social networks. Contact him at murai@dcc.ufmg.br.

Adriano Veloso is an associate professor of Computer Science at the Universidade Federal de Minas Gerais, Brazil. His interests are in machine learning and natural language processing. Contact him at adrianov@dcc.ufmg.br.

Fabrício Benevenuto is an associate professor in the Computer Science Department, Universidade Federal de Minas Gerais, Brazil. His research lies in topics related to social computing, computational journalism, and sentiment analysis. He is the corresponding author. Contact him at fabricio@dcc.ufmg.br.