Explainable Machine Learning for Fake News Detection

Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, Fabrício Benevenuto
Universidade Federal de Minas Gerais, Computer Science Department
Belo Horizonte, Minas Gerais, Brazil
{julio.reis,andre.correia,murai,adrianov,fabricio}@dcc.ufmg.br

ABSTRACT

Recently, there have been many research efforts aiming to understand fake news phenomena and to identify typical patterns and features of fake news. Yet, the real discriminating power of these features is still unknown: some are more general, but others perform well only with specific data. In this work, we conduct a highly exploratory investigation that produced hundreds of thousands of models from a large and diverse set of features. These models are unbiased in the sense that their features are randomly chosen from the pool of available features. While the vast majority of models are ineffective, we were able to produce a number of models that yield highly accurate decisions, thus effectively separating fake news from actual stories. Specifically, we focused our analysis on models that rank a randomly chosen fake news story higher than a randomly chosen fact with more than 0.85 probability. For these models we found a strong link between features and model predictions, showing that some features are clearly tailored for detecting certain types of fake news, thus evidencing that different combinations of features cover a specific region of the fake news space. Finally, we present an explanation of factors contributing to model decisions, thus promoting civic reasoning by complementing our ability to evaluate digital content and reach warranted conclusions.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Applied computing** → *Sociology*.

KEYWORDS

Fake News; Civic Reasoning; Features; Social Media

1 INTRODUCTION

More than a decade after their emergence, social media systems are used by over a third of the world's population [13]. These systems have significantly changed the way users interact and communicate online, spawning a whole new wave of applications and reshaping existing information ecosystems. In particular, social media systems have been dramatically changing the way news is produced, disseminated, and consumed in our society.

These changes, however, started an actual information war in the few last years, favoring misinformation campaigns, reducing the credibility of news outlets in these environments [35], and potentially affecting news readers opinions on critical matters for our society. Misinformation, spin, lies and deceit have of course been around forever, but the emergence of fake news has quickly evolved into a worldwide phenomenon, and while there are efforts attempting to better comprehend this phenomenon [14, 20], it is not surprising that most existing efforts are devoted to detecting fake news [9, 37, 39, 41]. Typically, most of these efforts reduce the problem to a classification task, in which news stories are labeled as fact/fake and supervised learning is then used to separate fact from fake with a model learned from the data. Fake news detection gained traction and attention, especially in assisting fact checkers to identify stories that are worth investigating [17, 28].

Despite the undeniable importance of the existing efforts in this direction, they are mostly concurrent work, which propose complementary solutions and features to train a classifier, providing hints and insights that are rarely or never tested together. Little is known about the discriminating power of features proposed in the literature, either individually or when combined with others. Some may be adequate for pinpointing specific types of fake news, while others are more general but not sufficiently discriminating. Moreover, while explaining the decisions made by the proposed models is central to understand the structure of fake content, this discussion is often left aside. In this work, we address all of these issues.

In particular, we want to provide answers to the following questions. *How hard is the detection task? Do we really need all these features, or should we focus on a smaller set of more representative features? Is there a trade-off between feature discriminating power and robustness to pattern variations? Is there a clear link between features and the type of fake news they can detect?*

To answer these questions, we first conduct a systematic survey, identifying existing features for fake news detection and proposing new ones. This results in almost 200 features to consider. To implement and evaluate these features, we used a public dataset recently released by BuzzFeed that was enhanced with Facebook commentaries on labeled news stories [32]. Since the considered features may have a variety of complex nonlinear interactions, we employ a classification algorithm with significant flexibility. Specifically, we chose a fast and effective learning algorithm called extreme gradient boosting machines, or simply XGB [7]. Finally, we performed an unbiased search for XGB models, so that each model is composed of a set of randomly chosen features. We enumerated roughly 300K models, enabling us to perform a unique macro-to-micro investigation of the considered features.

ACM WebSci '19, June 30–July 03, 2019, Boston, MA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326027

Findings: Our analysis unveil the real impact of a sleigh of features for fake news detection. Particularly, our results show that:

- Our unbiased model exploration reveals how hard is fake news detection, as only 2.2% of the models achieve a detection performance higher than 0.85 in terms of the area under the ROC curve (or simply, AUC);
- We found that among the best models, some features appear up to five times more often than others;
- We distinguish a small set of features that are not only highly effective but also contribute the most to increasing the robustness of the models;
- We place models in a high dimensional space, so that models that output similar decisions are placed close to each other. We then cluster the model space, and a centroid analysis reveals that prototype models are very distinct from each other. Our cluster analysis by AUC reinforce these results. For centroid prototypes, we present an explanation of factors contributing to their decisions. Our findings suggest that models within different groups separate fake from real content based on very different underlying reasons.

Additionally, our effort provides other valuable contributions as we survey a large number of recent and related works and we attempt to implement all previously explored features to detect fake news. We also proposed novel features which showed to be useful for the best models we generated. On the other hand, we emphasize that this paper is not about proposing the best combination of features or the best XGB model, but about investigating features' informativeness and simple models that can be generated from them, as well as using these models to explain predictions made for news stories.

The rest of the paper is organized as follows. Section 2 presents the background, including important definitions, related works and an overview of the main features for fake news detection proposed in the literature. In Section 3, we describe our experimental methodology proposed for this work. We then present and discuss our results and their implications in Section 4. Finally, we present in Section 5 concluding remarks and directions for future work.

2 BACKGROUND AND RELATED WORK

We begin by providing important definitions used in our work, then we describe our effort to survey existing works that propose features for detecting fake news.

2.1 Definitions

Fake news is a topic that still lacks a clear or universally accepted definition. In this work we adopt the definition of fake news and fake news detection used in previous works [2, 35].

Definition 2.1. **(Fake News)** *"is a news article that is intentionally and verifiable false".*

Definition 2.2. **(Fake News Detection.)** Given an unlabeled piece of news $a \in \mathcal{A}$, a model for fake news detection assigns a score $S(a) \in [0, 1]$ indicating the extent to which a is believed to be fake. For instance, if $S(a') > S(a)$, a' is more likely to be fake than a according to the model. A threshold τ can be defined such that the prediction function $F : \mathcal{A} \rightarrow \{\text{fake}, \text{not fake}\}$ is

$$F(a) = \begin{cases} \text{fake} & \text{if } S(a) > \tau, \\ \text{not fake} & \text{otherwise.} \end{cases}$$

2.2 Features for Fake News Detection

Broadly speaking there are two kinds of efforts to tackle the fake news problem. The first kind aims at better comprehending the phenomenon [20, 40]. Particularly, Vosoughi *et al.* [40] shows that fake news tends to spread faster than the real news. Lazer *et al.* [20] call for an interdisciplinary task force to approach this complex problem. The second kind of existing efforts comprises those that propose solutions to the problem or provide insights on how to detect fake news, i.e. discussing typical patterns that can be used as features. For instance, Pérez-Rosas *et al.* [25] conduct a set of learning experiments to build accurate fake news detectors using linguistic features. Similarly, Volkova *et al.* [39] build linguistic models to classify suspicious and trusted news. Overall, most of the existing efforts in this space are concurrent work which use specific data and feature sets to train classifiers without providing clear guidelines on which features are useful to detect and explain fake news.

The literature is quite broad if we consider efforts related to information credibility, rumor detection, and news spread. Next, we conduct a systematic survey on these efforts aiming to identify the features proposed by them. Table 1 presents a summary of this survey along with some of techniques used to extract those features. At a high-level, we can categorize features explored in previous works as follows: (i) features extracted from news content (e.g. language processing techniques) [15, 39, 42, 43]; (ii) features from source (e.g. reliability and trustworthiness) [21]; and finally (iii) features extracted from the environment, which usually involves signals extracted from the social network repercussion and spread [8].

Overall, our work provides contributions that encompass the two kinds of efforts previously described, since (i) we provide a better understanding of the fake news phenomena by explaining how these features are used in the decisions taken by computation models designed to detect fake news, and (ii) we evaluate the use of machine learning with different combinations of features. Our survey on existing features is also an important contribution on its own.

3 METHODOLOGY

In this section, we describe the dataset used in this work as well as implementation details for a large set of features for fake news detection. In addition to features from previous works, we propose a novel set of features for fake news detection that includes features that measure text quality. Finally, we describe our experimental setup and present our framework for quantifying the informativeness of features for fake news detection.

3.1 Dataset

Most of the existing efforts to detect fake news are limited by the data they use. Ideally, to implement all features from previous efforts we would need a dataset that contains for each news story labeled by specialists, their textual content, information about their sources, and about the dissemination of these news, particularly

Extracted from...	Feature Set	Techniques mostly used	References
News Content	Language Structures (Syntax)	Sentence-level features, such bag-of-words approaches, "n-grams", part-of-speech (POS tagging)	[9, 19, 27, 31, 35, 42]
	Lexical Features	Character level and word-level features, such as number of words, characters per word, hashtags, similarity between words, etc	[1, 5, 6, 15, 18, 27, 29, 35, 42, 43]
	Moral Foundation Cues	Moral foundation features	[39]
	Images and Videos	Indicators of manipulation and image distributions	[16]
	Psycholinguistic Cues	Additional signals of persuasive language such as anger, sadness, etc and indicators of biased language	[15, 19, 31, 39, 40]
	Semantic Structure	Word embeddings, "n-grams" extensions, topic models (e.g. latent Dirichlet allocation (LDA)), contextual informations	[5, 8, 9, 12, 31, 41–43]
	Subjectivity Cues	Subjectivity score, sentiment analysis, opinion lexicons	[27, 31, 39]
News Source	Bias Cues	Indicators of bias (e.g. politics), polarization	[29]
	Credibility and Trustworthiness	Estimation of user' perception of source credibility	[6, 34, 35]
Environment (Social Media)	Engagement	Number of page views, likes (on Facebook), retweets (on Twitter), etc	[6, 11, 12, 15, 18, 33, 35, 37, 40]
	Network Structure	Friendship network, complex network metrics	[6, 9, 12, 15, 18, 27, 33–35, 38–40]
	Temporal Patterns and Novelty	Time-series, propagation, novelty metrics	[6, 11, 12, 19, 33–35, 38, 40]
	User' Information	Users' profiles and characteristics across individual level and group level (e.g. their friends and followers)	[6, 15, 19, 27, 29, 33–35, 37, 38, 40]

Table 1: Overview of features for fake news detection presented in previous work.

in social media systems. We use a recently created dataset, named BuzzFace [32], with almost all of these characteristics. It consists of 2,282 news articles related to the 2016 U.S. election labeled by BuzzFeed journalists [36]. The BuzzFace dataset consists of an enriched version of the one created by BuzzFeed, with over 1.6 million comments associated to the news stories as well shares and reactions from Facebook users.

The news stories in the dataset are labeled into four categories: mostly true, accounting to 73% of all news articles, mostly false (4%), mixture of true and false (11%) and non-factual (12%). For simplicity, we discarded the non-factual content and merged the mostly false with the mixture of true and false into one single class, referred as "fake news" (349 out of 2,018 stories). The rationale is that stories that mix true and false facts may represent attempts to mislead readers. Thus, we focus our analysis into understanding how features are able to distinguish two classes, true and fake news.

Note: A typical pre-processing step is to separate factual from non-factual content. This task is easier than classifying factual data as fake or true since it is not necessary to check the veracity of the information using external sources. For illustration purposes, we conduct a small experiment to evaluate the accuracy of XGB [7] when discriminating factual and non-factual news using the features that will be described in Section 3.2. Our simple classifier performed very well, yielding 0.882±0.024 of AUC. It is possible to achieve even higher performance levels by choosing features better tailored for this task. For this reason, this work assumes that non-factual data was already removed and only factual data is used as input. The alternative approach is to consider a multi-label classification problem, but this has the potential to increase the number of instances that need to be verified by an expert.

3.2 Our Implementation of Features for Fake News Detection

Next, we briefly describe how we implemented or adapted the features summarized in Table 1. In total we considered 172 features for fake news detection.

3.2.1 News Content. We consider as news content not only the news story but also its headline and any message that was posted by a news source when releasing it in online social networks. For news articles embedded in images and videos, we applied image processing techniques for extracting text shown on them. In total, we evaluated 141 textual features. The main feature sets are described next.

Language Structures (SYNT, for syntax). We implemented 31 sentence-level features, including number of words and syllables per sentence. Features also include indicators of the 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 [10].

Lexical Features (LEXI). We implemented 59 linguistic features, including number of words, first-person pronouns, demonstrative pronouns, verbs, hashtags, all punctuations counts, etc.

Psycholinguistic Cues (PSYC). Linguistic Inquiry and Word Count (LIWC) [24] is a dictionary-based text mining software. We use its latest version (2015) to extract 44 features that capture additional signals of persuasive and biased language.

Semantic Structure (SEMA). We implemented semantic features, including the toxicity score obtained from Google's API[1]. 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 work was built based on news articles about the same topic or category (i.e. politics).

Subjectivity Cues (SUBJ). Using TextBlob's API[2], we compute subjectivity and sentiment scores of a text.

3.2.2 News Source. To extract features from news source, we first parsed all news URLs and extracted the domain information. When the URL was unavailable, we associated the official URL of news outlet to news article. Therefore, we extract 8 (eight) indicators of political bias, credibility and source trustworthiness, and use them as detailed next. Moreover, in this category, we introduce a new

[1] https://www.perspectiveapi.com/#/
[2] http://textblob.readthedocs.io/en/dev/

set composed by 5 (five) features, called domain localization (see below).

Bias Cues (BIAS). We use the political biases of news outlets from BuzzFeed dataset as a feature.

Credibility and Trustworthiness (CRED). In this feature set, we introduce 7 (seven) new features to capture aspects of credibility (or popularity) and trustworthiness of domains. We collect, using Facebook's API[3], 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 API to get the relative position of news domain on the Alexa Ranking [4]. Furthermore, using this same API, we collect Alexa's top 500 newspapers. Based on the hypothesis 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 non-zero edit distance) as features. Last, we use indicators of low credibility of domains compiled in [34] as features.

Domain Location (DOML). Ever since creating fake news became a profitable job, some cities have become famous because of residents who create and disseminate fake news [5]. In order to exploit the information that Domain localization 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 traceroute Linux command. Then the ipstack API is used to retrieve the location features. Although localization information (i.e. IP) has previously been used in works that exploit bots or spam detection [26], to the best of our knowledge there are no works that explore this data in fake news detection context.

3.2.3 Environment (Social Media). As indicators of user engagement and temporal patterns, we use information from Facebook. Next, we detail the 21 features from this category.

Engagement (ENGA). We use 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, 28800, 57600 and 86400 seconds), summing up to 12 features.

Temporal Patterns (TEMP). To capture temporal patterns from user commenting activities, we compute the rate at which comments are posted for the same time windows defined before.

3.2.4 Novel and Disregarded Features. Despite our efforts to include all the features described before, a few of them could not be included for various reasons. First, BuzzFace does not contain information related to network structure (i.e. Facebook connections). Additionally, some features, such as those extracted from images and videos were used in related problems [39], but are out of the scope of this work as our dataset contains mostly textual data.

More importantly, 19 of the previously described features are novel. In particular, we proposed all features related to domain, including IP, latitude, longitude, city, county, and domain credibility. We also proposed other features such as toxicity and readability to assess the writing style of news stories. Later on we show that some of these features were proven valuable for fake news detection.

[3]https://developers.facebook.com
[4]https://www.alexa.com
[5]https://www.bbc.com/news/magazine-38168281

3.3 Unbiased Model Generation

The exact approach to assess the real impact of features for fake news detection would require the exhaustive enumeration of all possible combinations of features, so that one model is obtained for each combination in the power set. Obviously, inspecting all possible subsets of features is computationally prohibitive. Instead, we sample the model space by randomly selecting the features that compose a model. More precisely, we begin by enumerating all possible 1-feature and 2-feature models (172 and 14,706 models, respectively). Next, we take each of the 2-feature models and include one new feature chosen uniformly at random, so as to build 3-feature models. This step is repeated until we reach models composed of 20 features (a total of 294,292 models). In each step we ensure that each feature is included the same number of times and that no feature appears twice within the same model. This compensates for the smaller number of few-feature models by keeping the number of models constant regardless of the number of features.

3.3.1 Classification Algorithm. The features we consider may have a variety of complex nonlinear interactions. Capturing these interactions requires a classification algorithm with significant flexibility. For this reason, we chose a learning algorithm called gradient boosting machines. The main idea of gradient boosting machines is to combine multiple models into a stronger one. More specifically, models are iteratively trained so that each model is trained on the errors of the previous models, thus giving more importance to the difficult cases. At each iteration, the errors are computed and a model is fitted to these errors. Finally, the contribution of each base model to the final one is found by minimizing the overall error of the final model. Fitting the base models is computationally challenging so we used a recent, high performance implementation of gradient boosting machines, called XGBoost (or simply, XGB) [7].

3.3.2 Evaluation. In order to evaluate how accurate the learned models are, we employ the standard area under the ROC curve (AUC [4]), which takes into account the sensitivity-specificity trade-off. Basically, the AUC is an estimate of the probability that a model will rank a randomly chosen fake news case higher than a randomly chosen fact case. The AUC is robust to class imbalance and considers all possible classification thresholds.

For each model, we performed a 5-fold cross-validation. The dataset is partitioned into five partitions, out of which four are used as training data, and the remaining one is used as the validation-set. The process is then repeated five times with each of the sets used exactly once as the validation-set, thus producing five results. Hence, the reported AUC values are averaged over the five runs. Further, we employ the mean absolute deviation (or simply, MAD) in order to get a sense of how spread out the AUC values are through the five validation sets. Therefore, for each model we have an estimate of its predictive accuracy and variability.

3.4 Feature Importance and Shapley Additive Explanations

Effective models perform decisions that are usually hard to explain. However, understanding why a model has made a specific decision is paramount in any fake news detection application scenario, as it provides insight into the reasons why the content was considered to be fake, tooling fact-checkers with the facts that contributed most to the decision.

The typical approach for explaining the decisions of a model is based on calculating the impact (or importance) each feature has on the decision. Feature importance can be defined as the increase in the model prediction error after feature values are permuted, since this operation breaks the relationship between the feature and the outcome. Therefore, a feature is important if permuting its values increases the model error, because the model relied on the feature for the correct decision. On the other hand, a feature is not important if permuting its values keeps the model error unchanged, because the model ignored the feature for the decision.

Often, however, features interact with each other in many different and complex ways in order to perform accurate decisions. Thus, the feature importance is also given as a function of the interplay between the features. In this case, Shapley values [22] can be used to find a fair division scheme that defines how the total importance should be distributed among the features. In fact, Shapley values are theoretically optimal and are the unique consistent and locally accurate attribution values. Unfortunately, Shapley values can be challenging to compute, and thus we focus on explaining only the top-most effective models.

4 RESULTS

In this section we describe the results of the experiments designed to answer our research questions. In Section 4.1, we investigate the predictive accuracy and variability of the features. In Section 4.2, we focus only on the best performing models in order to evaluate models in terms of effectiveness and variability. Then, in Section 4.3, we cluster the model space according to the features present in each model, and we construct an investigation to understand the role of features in the model decisions. Finally, we attempt to explain the decisions made by some prototype models in Section 4.4.

4.1 Features: Accuracy and Variability

We quantify the predictive accuracy of a feature by considering all models in which the feature was included. More specifically, the predictive accuracy of a feature is given as the average AUC value of all models in which the feature was included. Similarly, the variability of a feature is given as the average MAD value of all models in which the feature was included. Figure 1 shows how features are distributed in terms of predictive accuracy and variability. Clearly, there is a small number of features for which the predictive accuracy is significantly higher. Specifically, around 5% of the considered features are included into models in which the average AUC values are higher than 0.85. The majority of the features are associated with significantly lower average AUC values. The same trend is observed when we investigate the distribution of features in terms of variability. Around 3% of the considered features are associated with relatively low variability.

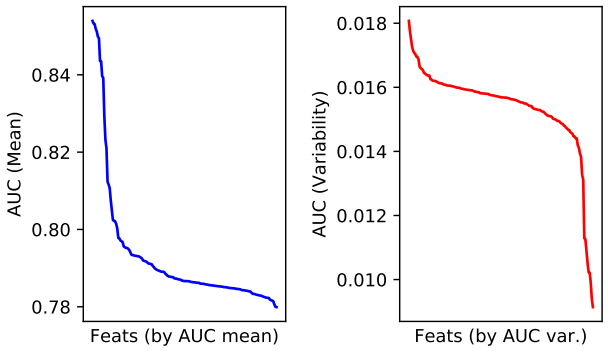


Figure 1: Distribution of features. Left – Predictive accuracy. Right – Variability.

4.2 Top 10 % Models: Accuracy and Variability

Now we investigate whether relatively simple models (composed by up to 20 features) can perform consistently well across the dataset. In order to do so, we take the top 10% models w.r.t. AUC. Among the best performing models, we are interested in those that exhibit low variability.

Figure 2 shows a scatter plot of the top 10% models w.r.t AUC, each represented by a dot. Each dot diameter is proportional to the ratio between the respective model's AUC mean and variability. Cartesian coordinates of each dot center are obtained from the vector of the probabilities assigned by the model to each fake news case in the validation set.

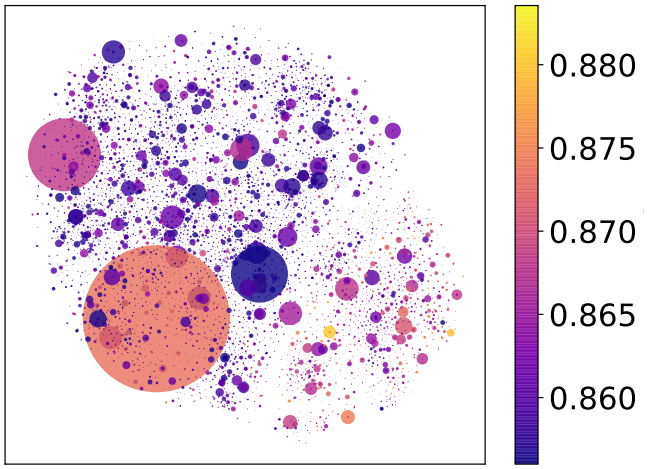


Figure 2: Each point represents a model. Color indicates AUC value. Diameter indicates accuracy consistency across folds (i.e., larger diameter implies lower variability). Probabilities associated to fake news stories by each model are taken as a vector to compute its 2D t-SNE [23] representation (i.e., for defining the model's position in the plot).

First, we note that the mean AUC is in the range $[0.855, 0.885]$ and, therefore, the different diameters are mostly due to AUC's variability. We observe the presence of very few models with excellent

performance on average (yellow dots, AUC > 0.88), but with high variability. On the other hand, there are many models with lower variability, but with lower average AUC values (medium-sized purple and blue models, AUC < 0.865). Finally, there are two models with the good trade-off between performance and variability (pinkish dots, AUC ≈ 0.87). These cases will be discussed in the next sections.

To better understand the relationship between features and model performance, we take the best performing models and compute the prevalence of features. Also, to understand the relationship between them and AUC variability, we take, from the top performing models, the 10% with the highest and 10% with the lowest variability and compute the prevalence of features in these sets.

4.2.1 Accuracy. Considering models with the highest AUC values, features extracted from the environment (i.e. from social media), are more frequent (e.g. the number of shares (40%) and reactions count (29%)). Moreover, features that capture information regarding the location and credibility of domains (e.g. ranking position of domain from Social Alexa (39%), page reactions from users on social media (29%), IP of domain (28%), etc.), are very frequent in this group of models. Finally, features that capture political bias of news outlet (i.e. mainstream, left-leaning, and right-leaning (37%)) are quite prevalent in the best models. On the other hand, character level, word-level and sentence-level features (e.g. count) are less frequent in best models (7% of models on average).

4.2.2 Variability. While features from the social media (e.g. share count (13%)), IP of domain (14%) and semantic structure (e.g. toxicity (13%)) are very frequent in models with low variability, features from user engagement (from social media) (e.g. number of comments on first 7200 seconds (12%), political bias (12%) and Facebook page (12%)), occurred more often in models with high variability. Sentence level-features and Psycholinguistic features are very frequent both in models with high and low variability.

In sum, we conclude that there are many combinations of features that yield models with high performance and low variability. In the next section, we investigate whether these models are redundant (i.e. identify similar sets of fake news) or complementary.

4.3 Clustering the Model Space

In order to understand whether the top 10% models cover different regions of the space of fake news, we cluster them from binary vector representations that indicate which features are present in each model. To cluster these models, we use the standard K-Means algorithm based on Euclidean distances [3]. To find the optimum value of K, we use the Silhouette Score [30], which measures, on average, how tightly grouped all the members in different clusters are, and select the value of K, for which the Silhouette Score is the highest. In this work, we use $K = 6$. The sizes of the resulting clusters vary from 3,769 to 5,921 (mean 4,908 and std. dev. 703).

Once again, we embed the models in a 2D space based on the probabilities assigned to fake news cases and color code the models according to the cluster they belong to. Hence cohesive clusters indicate that models within the same cluster are better at identifying specific types of fake news. If this is the case, models that belong to the same cluster (i.e. share similar features) are expected to be close to each other in the embedding, indicating that they assign similar scores to the fake news in the test set. In fact, this is what we observe in Figure 3. Next, we analyze which types of features best describe each cluster.

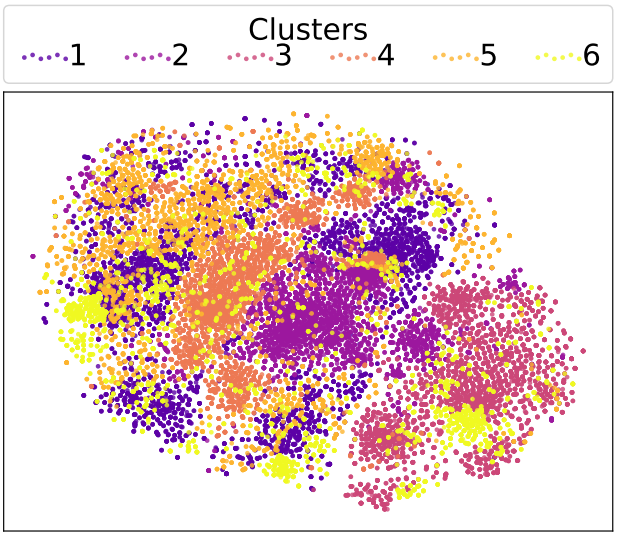


Figure 3: t-SNE representations [23] of models based on the scores associated to each fake news in the validation set. Colors indicate clusters found from binary vectors indicating which features were used in each model. Proximity between models from the same cluster suggest correlation between features used and fake news correctly detected.

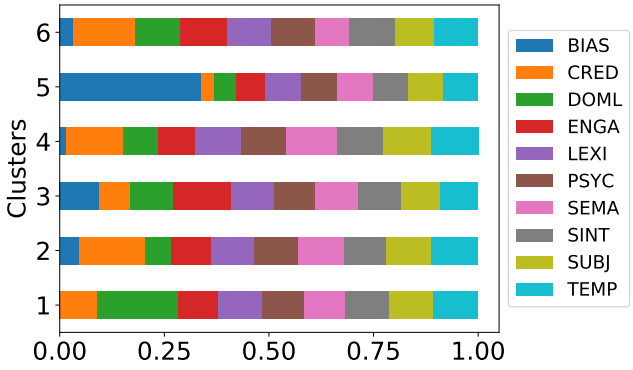
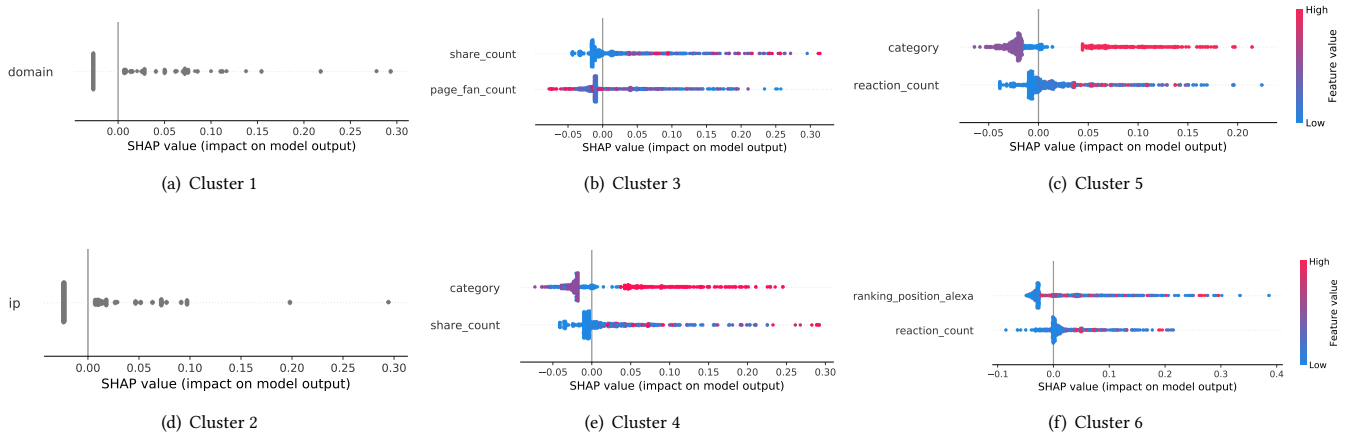


Figure 4: Descriptions of clusters in terms of feature sets, represented as segments. Segment lengths are normalized $R_{i,t}$ ratios and indicate how much more/less often features of type t appear in cluster i than in a random null model.

4.3.1 Describing clusters in terms of types of features. When we focus on the analysis of the top 10% performing models, features no longer appear with the same frequency. In addition, clusters include different numbers of models, each of which can include any number of features. In order to compare the frequency of specific types of



(a) Cluster 1 (b) Cluster 3 (c) Cluster 5

(d) Cluster 2 (e) Cluster 4 (f) Cluster 6

Figure 5: SHAP summaries for the closest models to each cluster centroid. These violin plots show the impact of each feature on model output (positive values on x-axis mean increased chance of being fake). Features are color-coded according to its values (highest: red, lowest: blue), except for Domain and IP, which do not have a meaningful scale. For instance, for Cluster 5's centroid, high values of `reaction_count` are associated to positive SHAP values.

features across clusters, we define a (random) null model. This allows us to determine how much more (less) often than expected a given feature type appears in a cluster.

Let $C_{i,t}$ be the number of times features of type t appear in models from cluster i for $t \in$ BIAS, CRED, ..., TEMP and $i = 1, \ldots, 6$. Multiple features of the same type are counted multiple times. Also, let $C_i = \sum_t C_{i,t}$ be the total number of features in cluster i. Denote by $N_t = \sum_i C_{i,t}$ the number of times features of type t appear among the top 10% performing models. The expectation of $C_{i,t}$ if features were assigned to clusters completely at random is $C_i N_t / \sum_t N_t$. Therefore, the ratio $R_{i,t} = C_{i,t} / (C_i N_t / \sum_t N_t)$ measures how much more (less) often features of type t appear in cluster i than in a random null model.

Figure 4 shows ratios $R_{i,t}$ normalized for each cluster i (i.e. divided by $\sum_t R_{i,t}$). Normalized ratios allow us to identify which types of features better describe each cluster. We note that clusters comprise combinations of features types in different proportions. All clusters use features from all features types, except for cluster 1, which does not include BIAS features. These features are more frequent in cluster 5 and less so in cluster 4. CRED features are very prevalent in clusters 2, 4 and 6, but less used by models in cluster 5. Finally, DOML features are very prevalent in cluster 1. Therefore, these observations combined with Figure 3 corroborate the hypothesis that models generated from different combinations of features are able to correctly identify different fake news groups.

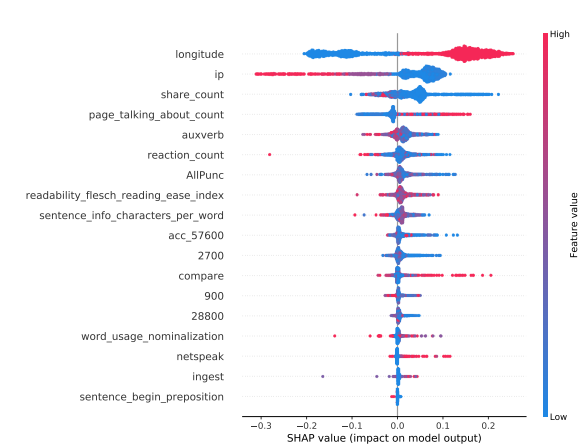
4.4 Explaining Model Decisions

In this section, we use SHAP [22] to explain why news are classified as fake or not by representative models of each cluster. SHAP is short for SHapley Additive exPlanations. It is a unified approach to interpreting model predictions. As such, SHAP assigns a "force" or importance value – positive or negative – to each feature in a particular prediction [22]. The output value (prediction) consists of the sum of the base value (average prediction over the validation set) and these forces (closer to 1.0 means more likely to be fake). In addition, SHAP allows us (i) to summarize the importance of a feature, and (ii) to associate low/high feature values to an increase/decrease in output values, through color-coded violin plots built from all predictions.

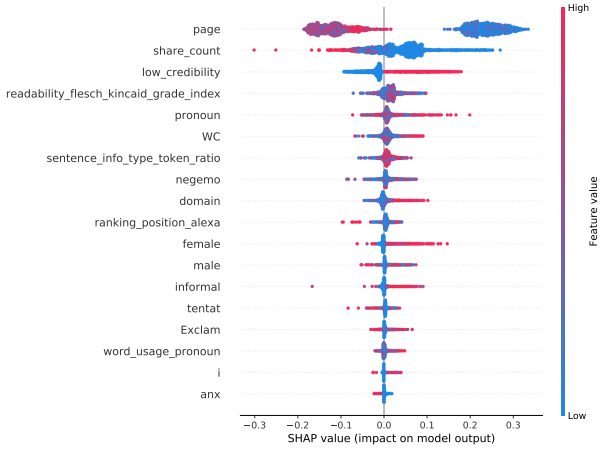
Representative models of each cluster were selected according to the following criteria: (1) by centroid proximity, where we select the closest model to the cluster centroid (Figure 5); and (2) by AUC, where we select in each cluster the model with the best performance w.r.t AUC (Figure 6).

Figure 5 shows violin plots of SHAP values for features used by each of the selected centroid models. Interestingly, we note that the closest models to the centroids have either one or two features, all of which come from feature sets DOML, BIAS, ENGA and CRED.

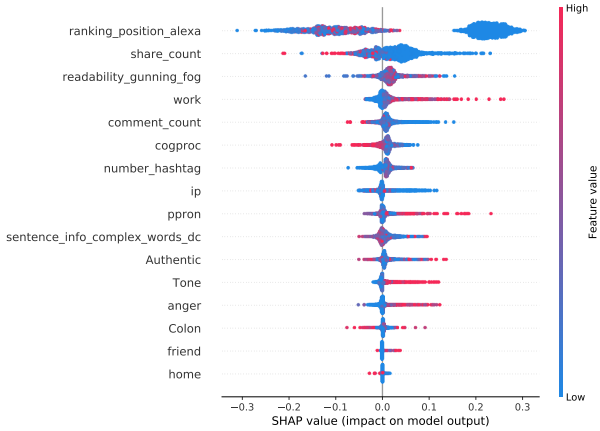
The representative models of clusters 1 and 2 have a single feature, domain and ip, respectively. In either case, we remove the feature value color-coding since low/high feature values are not meaningful in these cases. These plots are very similar, as there is a close mapping from ips to domains. We found that the three domains that have a large negative impact (i.e. less likely fake) on the output value are politico.com, abcnews.go.com and cnn.com. For models within cluster 3, high page fan counts have large impact – positive or negative – over predictions. Extremely high values though, are almost always associated with negative impact, since popular pages are less likely to share fake news. Large numbers of shares, however, tend to increase output values. This is consistent with recent research that shows that fake news are more likely to be shared [20, 40]. Models within cluster 4 also include number of shares as a feature and can be interpreted in the same way. Representative models of cluster 4 and 5 include categories of political bias as a feature, which takes on three values: -1 for left-leaning, 0 for mainstream and 1 for right-leaning. As expected, category has a negative impact on output for mainstream news (purple dots), but



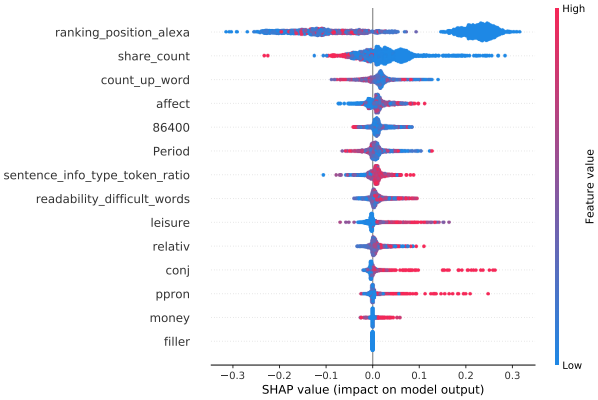
(a) Cluster 1



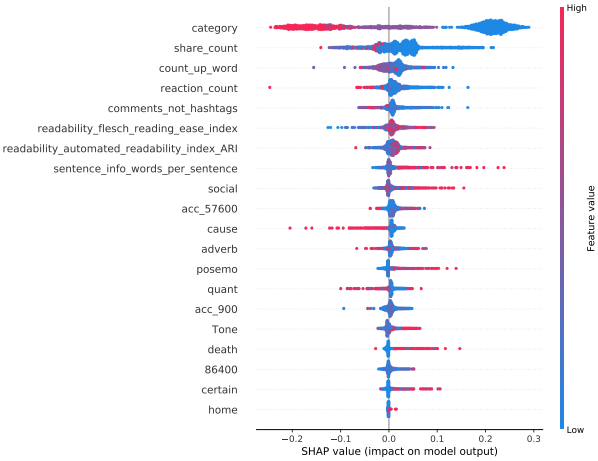
(b) Cluster 2



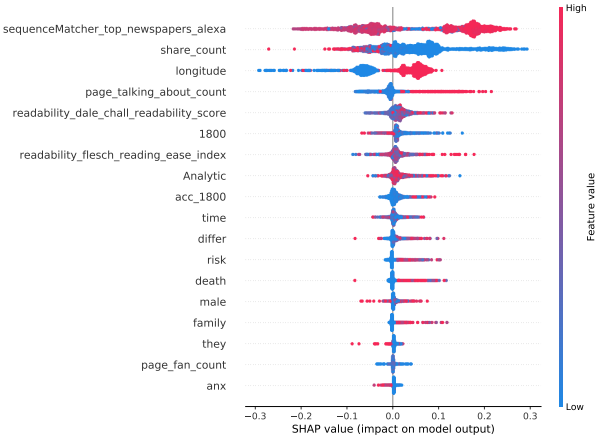
(c) Cluster 3



(d) Cluster 4

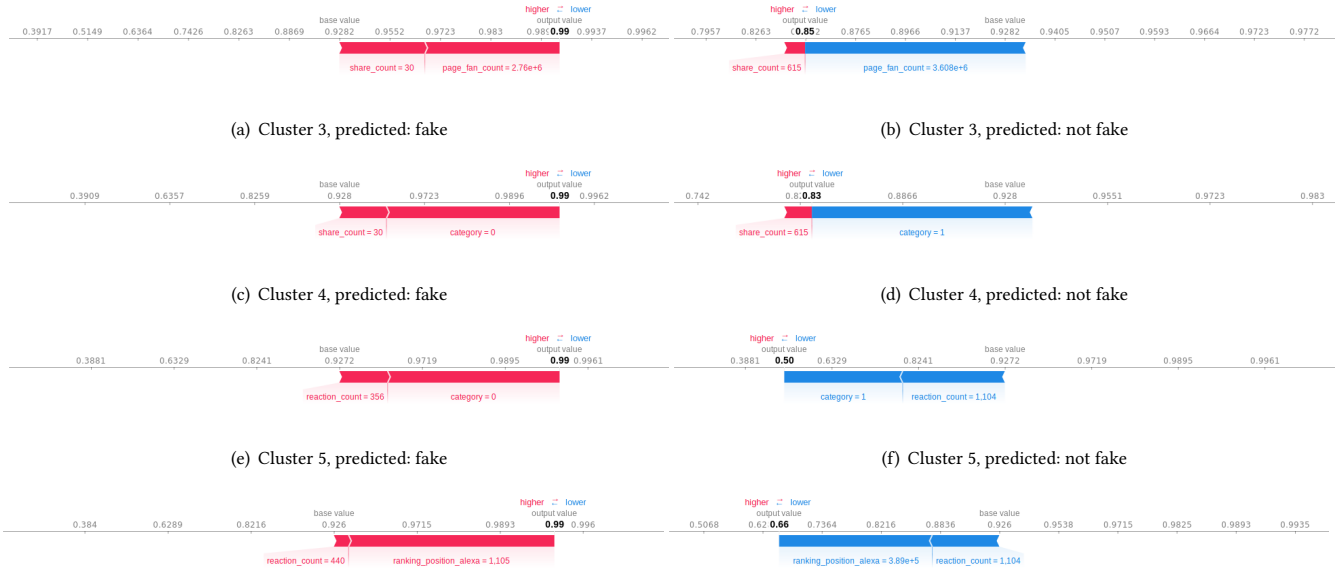


(e) Cluster 5



(f) Cluster 6

Figure 6: SHAP summaries for the highest AUC model in each cluster.



(a) Cluster 3, predicted: fake

(b) Cluster 3, predicted: not fake

(c) Cluster 4, predicted: fake

(d) Cluster 4, predicted: not fake

(e) Cluster 5, predicted: fake

(f) Cluster 5, predicted: not fake

(g) Cluster 6, predicted: fake

(h) Cluster 6, predicted: not fake

Figure 7: SHAP results on test cases for representative centroid models of each cluster. Base value (0.926) is the score of an instance with average feature values. Feature values that increase (decrease) prediction score are shown below red (blue) bars whose lengths depict the corresponding amount. Output value = base value + length of red bars − length of blue bars.

increases the prediction value if the source exhibits political bias. Models representing cluster 5 and 6 share number of reactions as a feature. In the same way as number of shares, higher values of this feature increase the chances that a piece of news is classified as fake. Last, the model representing cluster 6 also includes the ranking position retrieved from Alexa as a feature. As expected, very low values (top of the ranking) tend to have a large negative impact on the output.

In Figure 6 we present the SHAP results for the top performing models w.r.t AUC. Differently from the models closest to the centroids that have one or two features, the clusters in Figure 6 have much more features. Cluster 1 uses localization and Domain features, longitude and ip, which are features proposed in the present work. Clusters 2 and 3 rank engagement features nearly as the most influential features. For Clusters 3 and 4 Alexa's ranking position appears as one of the most important features, similar to the Centroid clusters. Cluster 5 and Cluster 6 are a mix of localization, Domain and Engagement features, which are top-of-the-rank features on the others clusters. Psycholinguistic cues are shown to be relevant in all clusters, once they appear in every model. Similar findings were obtained when analyzing the lowest variability models in each cluster, where most of them contain Psycholinguistic cues, Domain and Engagement Features.

Last, focusing on centroid models, we include, for each of the representative models, examples of news stories that are scored higher (lower) than average, indicating it is more (less) likely to be fake. Figure 7 shows SHAP results on different news stories, which explain the role that each feature had on the decision. For ethical reasons, we omit the results corresponding to Clusters 1 and 2, which respectively indicate domains and IPs less/more likely to publish fake news. Figures 7 (a,b) shows that very high page fan counts decrease the output value, while the effect of share counts may depend on the former feature. In Figures 7 (c,d) we observe that news from mainstream media (category 0) tend to have lower output values, whereas news from politically biased sources (categories -1 and 1) often receive a positive bump in their outputs. In Figures 7 (e,f) we observe other examples of category's impact and that the number of reactions has a similar behavior as the number of shares. Finally, in Figures 7 (g,h) we note that being at the bottom (top) of the Alexa's ranking has a negative (positive) impact on the output value.

5 CONCLUDING DISCUSSION

In this work we provide many contributions that are relevant to the field. First, we survey a large number of recent and related works as an attempt to implement all potential features to detect fake news. We proposed novel features, such as those related to the source domain, which appear within the best models up to five times more often than other features. Second, our framework reveals how hard is to detect fake news, as only a small fraction of the models (only 2.2%) achieve a detection performance higher than 0.85 in terms of AUC. We hope our effort can become a baseline for other solutions to the same problem.

Finally, our findings suggest that certain types of fake news tend to be identified by models with specific combinations of features. As a consequence, different models separate fake stories from real ones based on very different reasoning. This shows the complexity of the problem and allow us to understand how hard it is for a

single solution to tackle all forms of fake news stories. As future work we plan to categorize the fake news stories as a strategy to construct effective and robust ensembles of classifiers. For instance, in this work we showed the different models of clusters that are made of random combinations of features. This indicates that ensemble techniques that combine models from different clusters are a promising avenue for investigation.

ACKNOWLEDGMENTS

This work was partially supported by the project FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, process number APQ-01400-1, as well as grants from Google, CNPq, CAPES, and Fapemig.

REFERENCES

[1] Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using N-gram analysis and machine learning techniques. In *Int'l Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC)*.

[2] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.

[3] David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proc. of the Annual ACM-SIAM symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics.

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.

[5] Sreyasee Das Bhattacharjee, Ashit Talukder, and Bala Venkatram Balantrapu. 2017. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *Proc. of the Int'l Conference on Big Data (Big Data)*. IEEE.

[6] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proc. of the Int'l Conference on World Wide Web (WWW)*.

[7] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proc. of the Int'l Conference on Knowledge Discovery and Data Mining (KDD)*.

[8] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PLOS ONE* 10, 6 (2015).

[9] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *Proc. of the Annual Meeting of the (ASIS&T)*.

[10] Daniel H Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2017. A general multiview framework for assessing the quality of collaboratively created content on web 2.0. *Journal of the Association for Information Science and Technology* 68, 2 (2017), 286–308.

[11] Samantha Finn, Panagiotis Takis Metaxas, Eni Mustafaraj, Megan OâĂŹKeefe, Lindsay Tang, Susan Tang, and Laura Zeng. 2014. TRAILS: A system for monitoring the propagation of rumors on twitter. In *Proc. of the Computation + Journalism Conference (C+J)*.

[12] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor Cascades. In *Proc. of the Int'l AAAI Conference on Weblogs and Social (ICWSM)*.

[13] Kevin Gallagher. 2017. The Social Media Demographics Report: Differences in age, gender, and income at the top platforms. *http://www.businessinsider.com/the-social-media-demographics-report-2017-8, Business Insider* (2017).

[14] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B Everett, and others. 2018. Fake News vs Satire: A Dataset and Analysis. In *Proc. of the Int'l Conference on Web Science (WebScience)*.

[15] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *Proc. of the Int'l Conference on Social Informatics (SocInfo)*.

[16] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia* 19, 3 (2017), 598–608.

[17] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proc. of the Int'l Conference on Web Search and Data Mining (WSDM)*.

[18] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proc. of the WWW Companion*.

[19] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PLOS ONE* 12, 1 (2017).

[20] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, and others. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[21] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. On the discovery of evolving truth. In *Proc. of the Int'l Conference on Knowledge Discovery and Data Mining (KDD)*.

[22] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proc. of the Neural Information Processing Systems (NIPS)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc.

[23] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[24] J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* (2001).

[25] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *Proc. of the Int'l Conference on Computational Linguistics* (2017).

[26] Anirudh Ramachandran and Nick Feamster. 2006. Understanding the network-level behavior of spammers. In *Proc. of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)*.

[27] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and tracking political abuse in social media. In *Proc. of the Int'l AAAI Conference on Weblogs and Social (ICWSM)*.

[28] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems* 34, 2 (2019).

[29] Manoel H. Ribeiro, Pedro H. C. Guerra, Wagner Meira Jr., and VirgÃŋlio Almeida. 2017. "Everything I Disagree With is# FakeNews": Correlating Political Polarization and Spread of Misinformation. In *Proc. of Data Science + Journalism Workshop*.

[30] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[31] Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proc. of the Workshop on Computational Approaches to Deception Detection (NAACL-HLT)*.

[32] Giovanni Santia and Jake Williams. 2018. BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos. In *Proc. of the Int'l AAAI Conference on Weblogs and Social (ICWSM)*.

[33] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proc. of the WWW Companion*.

[34] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 4787.

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

[36] C. Silverman, L. Strapagiel, H. Shaban, E. Hall, , and J. Singer-Vine. 2016. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. *https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis, Buzzfeed* (2016).

[37] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. In *Proc. of the Workshop on Data Science for Social Good (SoGood)*.

[38] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake News Detection in Social Networks via Crowd Signals. In *Proc. of the WWW Companion*.

[39] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proc. of the Annual Meeting of the ACL*.

[40] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[41] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proc. of the Annual Meeting of the ACL*.

[42] Wei Wei and Xiaojun Wan. 2017. Learning to identify ambiguous and misleading news headlines. In *Proc. of the Int'l Joint Conference on AI (IJCAI)*.

[43] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proc. of the WWW Companion*.