DistrustRank: Spotting False News Domains

Conference Paper · May 2018

DOI: 10.1145/3201064.3201083

Vinicius Woloszyn
Universidade Federal do Rio Grande do Sul

DistrustRank: Spotting False News Domains.

Vinicius Woloszyn
Federal University of Rio Grande do Sul / L3S Research Center
Porto Alegre, Brazil
vwoloszyn@inf.ufrgs.br

Wolfgang Nejdl
L3S Research Center
Hannover, Germany
nejdl@L3S.de

ABSTRACT

In this paper we propose a semi-supervised learning strategy to automatically separate fake News from reliable News sources: DistrustRank. We first select a small set of unreliable News, manually evaluated and classified by experts on fact checking portals. Once this set is created, DistrustRank constructs a weighted graph where nodes represent websites, connected by edges based on a minimum similarity between a pair of websites. Next it computes the centrality using a biased PageRank, where a bias is applied to the selected set of seeds. As an output of the proposed model we obtain a trust (or distrust) rank that can be used in two ways: a) as a counter-bias to be applied when News about a specific subject is ranked, in order to discount possible boosts achieved by false claims; and b) to assist humans to identify sources that are likely to be source of fake News (or that are likely to be reputable), suggesting websites that should be examined more closely or to be avoided. In our experiments, DistrustRank outperforms the supervised approaches in either ranking and classification task.

CCS CONCEPTS

• **Information systems** → *World Wide Web*; *Content ranking*;

KEYWORDS

Credibility Analysis, Rumor Detection, Text Mining

ACM Reference Format:
Vinicius Woloszyn and Wolfgang Nejdl. 2018. DistrustRank: Spotting False News Domains.. In *WebSci '18: 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3201064.3201083

1 INTRODUCTION

Many people have access to News through different online information sources, ranging from search engines, digital forms of mainstream News channels to social network platforms. Compared with traditional media, information on the Web can be published quickly, but with few guarantees on the trustworthiness and quality. This issue can be found in different domains, such as fake reviews on collaborative review websites, manipulative statements about companies, celebrities, and politicians, among others [5, 9].

The task of assessing the believability of a claim is a thorny issue. Kumar's work [8] reports that even humans sometimes are not able to distinguish hoax from authentic ones, and that quite a few people could not differentiate satirical articles from the true News (e.g. www.nypost.com/2018/02/01/mom-teams-up-with-daughter-to-fight-girl-on-school-bus/). With the increasing number of hoaxes and rumors, fact-checking websites like *snopes.com*, *politifact.com*, *fullfact.org*, have become popular. These websites compile articles written by experts who manually investigate controversial claims to determine their veracity, providing shreds of evidence to for the verdict (e.g. true or false).

Many works have addressed the problem of false claims detection. Most of them rely on supervised algorithms such as classification and regression models [2, 7, 8, 13–15, 17]. However, the quality of results produced by supervised algorithms is dependent on the existence of a large, domain-dependent training data set. The task of creating a data set of News claims, besides being a manual process dependent on motivated annotators, fails to consider the most recent News. Despite the typical inferior performance, semi-supervised methods are an attractive alternative to avoid the labor-intense and error-prone task of manual annotation of training data sets.

In this paper, we propose DistrustRank, a novel semi-supervised algorithm that identifies unreliable News websites based only on the headline extracted from the News article's link. In the News Websites, the News article is generally shared using a long link which contains the News headline and acts as a good summary of the News article content. This choice is motivated by performance issues, since for a fast and scalable method the extraction of features for comparison cannot be time-consuming. Additionally, using only links instead of entire News article content is a good strategy to help the integration of DistrustRank with search engines since it does not need additional features. The use of links as the main feature is also a common strategy in other areas, such as Query Re-Ranking [1, 16].

DistrustRank constructs a weighted graph where nodes represent websites, connected by edges based on a minimum similarity between a pair of websites, and then compute the centrality using a biased PageRank, where a bias is applied to the selected set of seeds. In addition, DistrustRank takes into account fake websites

WebSci '18, May 27–30, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5563-6/18/05…$15.00
https://doi.org/10.1145/3201064.3201083

similarities, as a minimum similarity threshold is dynamically defined based on the characteristics of the set of false websites. The resulting graph is composed of several components, where each component represents websites with similar characteristics. Next, a search that begins at some particular node v will find the entire connected component containing v. Finally, the centrality index of the neighbors of v are used to compose the final distrust rank.

The output of the method presented in this paper is a trust (or distrust) rank that can be used in two ways:

(1) as a counter-bias to be applied when News about a specific subject is ranked, in order to discount possible boosts achieved by false websites;
(2) to assist people to identify sources that are likely to be fake (or reputable), suggesting which websites should be examined more closely or to be avoided.

Our experiments on websites indexed by Internet Archive[1] reveal that DistrustRank outperforms the chosen supervised baseline (Support Vector Machine) in terms of imitating the human experts judging about the credibility of the websites.

The remaining of this paper is organized as follows. Section 2 discusses previous works on fake News detection. Section 3 presents details of the DistrustRank algorithm. Section 4 describes the design of our experiments, and Section 5 discusses the results. Section 6 summarizes our conclusions and presents future research directions.

2 RELATED WORK

Several studies have addressed the task of assessing the credibility of a claim. For instance Popat et al. [13] proposed a new approach to identify the **credibility** of a claim in a text. For a certain claim, it retrieves the corresponding articles from News and/or social media and feeds those into a distantly supervised classifier for assessing their credibility. Experiments with claims from the website *snopes.com* and from popular cases of Wikipedia hoaxes demonstrate the viability of Popat et al proposed methods. Another example is TrustRank [6]. This work presents a semi-supervised approach to separate reputable good pages from spam. To discover good pages it relies on an observation that good pages seldom point to bad ones, i.e. people creating good pages have little reason to point to bad pages. Finally, it employs a biased PageRank using this empirical observation to discover other pages that are likely to be good.

Controversial subjects can also be indicative of dispute or debate involving different opinions about the same subject. Detect and alert users when they are reading a controversial web page is one way to make users aware of the information quality they are consuming. One example of **controversy** detection is [2] which relies on supervised k-nearest-neighbor classification that maps a webpage into a set of neighboring controversial articles extracted from Wikipedia. In this approach, a page adjacent to controversial pages is likely to be controversial itself. Another work in this sense is [12] which aims to generate contrastive summaries of different viewpoints in opinionated texts. It proposes a Comparative LexRank, that relies on random walk formulation to give a score to a sentence based on their difference to others sentences.

Factuality Assessment is another way to asses the information quality. Yu et al.'s work [20] aims to separate opinions from facts, at both the document and sentence level. It uses a Bayesian classifier for discriminating between documents with a preponderance of opinions, such as editorials from regular News stories. The main goal of this approach is to classify a document/sentence in factual or opinionated text from the perspective of the author. The evaluation of the proposed system reported promising results in both document and sentence levels. Other work on the same line is [14], which proposes a two-stage framework to extract opinionated sentences from News articles. In the first stage, a supervised learning model gives a score to each sentence based on the probability of the sentence to be opinionated. In the second stage, it uses these probabilities within the HITS schema to treat the opinionated sentences as Hubs, and the facts around these opinions are treated as the Authorities. The proposed method extracts opinions, grouping them with supporting facts as well as other supporting opinions.

There also some works that analyze how a piece of information flows over the internet. For instance, [3] presents an interesting analysis about how Twitter bots can send spam tweets, manipulate public opinion and use them for online fraud. It reports the discovery of the 'Star Wars' botnet on Twitter, which consists of more than 350,000 bots tweeting random quotations exclusively from Star Wars novels. It analyzes and reveals rich details on how the botnet is designed and gives insights on how to detect **virality** in Tweeter.

Other works analyze the writing style in order to detect a false claim. [7] reports that Fake News in most cases are more similar to satire than to real News, leading us to conclude that persuasion in the fake News is achieved through heuristics rather than the strength of arguments. It shows that the overall title structure and the use of proper nouns in titles are very significant in differentiating fake from real. It gives an idea that fake News is targeted for audiences who are not likely to read beyond titles and that they aim at creating mental associations between entities and claims. Decrease the **readability** of texts is also another way to overshadow false claims on the internet. Many automatic methods to evaluate the readability of texts have been proposed. For instance, Coh-Metrix [4], which is a computational tool that measures cohesion, discourse, and text difficulty.

Most of the works just cited rely on supervised learning strategies addressed to assess News articles using few different aspects, such as credibility, controversy, factuality and virality of information. Nonetheless, a common drawback of supervised learning approaches is that the quality of the results is heavily influenced by the availability of a large, domain-dependent annotated corpus to train the model. Unsupervised and semi-supervised learning techniques on the other hand, are attractive because they do not imply the cost of corpus annotation. In short, our method uses a semi-supervised strategy where only a small set of unreliable News

[1] https://web.archive.org/

websites is used to spot another bad News websites using a biased PageRank.

3 DISTRUSTRANK ALGORITHM

To spot unreliable News websites, without a large annotated corpus, we rely on an important empirical observation: fake News pages are similar to each other. This notion is fairly intuitive, while News websites approach a broad scope of subjects, unreliable pages are built to mislead people in specific areas, such as fake News about companies, politicians and celebrities. Additionally, some of the News websites analyzed share copies of the same unreliable News. Figure 1 shows the distribution of the similarity between fake and true News websites. Using a Wilcoxon statistical test [18] with a significance level of 0.05, we verified that the similarity between false News websites is statistically higher to true News websites.

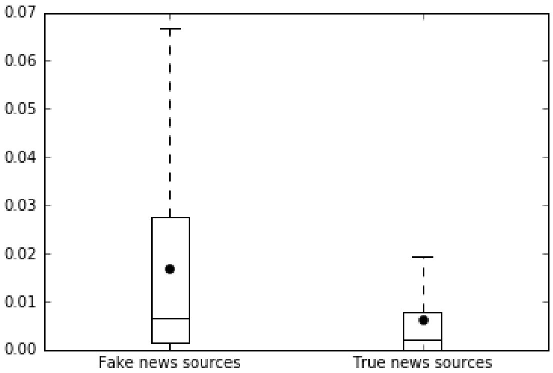

Figure 1: The distribution of the URL similarity between false and true News domains, where * represent the mean.

The intuition behind DistrustRank is that the credibility score of a website can be regarded as the problem of encountering websites which headlines do not differ much from fake websites headlines. To solve this problem, our approach relies on the concept of graph centrality to rank websites according to their estimated centrality.

We propose to represent the relationship between websites as a graph, in which the vertices represent the website, and the edges are defined in terms of the similarity between pairs of vertices. We define similarity as a function that measures the textual similarity of the headlines present in the URLs shared by News websites. Our hypothesis is that fake News websites have a high centrality index since they are similar to many other fake News websites. The biased centrality index produces a ranking of vertices' importance, which in our approach indicates the distrust of websites.

Let L be a set of websites, and $r \in L$ a tuple $\langle d, u \rangle$, where $r.d$ represents the domain of a website and $r.u$ a set of links for their News. DistrustRank builds a graph representation $G = (V, E)$, where $V = R$ and E is a set of edges that connects pairs $\langle u, v \rangle$ where $v, u \in V$, and uses biased PageRank to calculate centrality scores for each vertex. The main steps of the DistrustRank algorithm: (a) it builds a similarity graph G between pairs of News websites; (b) the graph is pruned (G') by removing all edges that do not meet a minimum similarity threshold, dynamically calculated based on the average similarity between URLs of fake domains; (c) a search that begins at some particular node v will find the entire connected component containing v; (d) using biased PageRank, the centrality scores are calculated and used to construct a ranking. The pseudocode of DistrustRank is displayed in Algorithm (1), where G and G' are represented as adjacency matrices W and W'. In the remaining of this section, we detail the similarity function, and the process to obtain the centrality index ranking.

Algorithm 1 - DistrustRank Algorithm (L, S , β): S

- Input: a set of websites L, a set of unreliable websites S and β is the base threshold.
- Output: ordered list O containing the their distrust score.

```
%building a similarity graph
for each $u, v \in L$ do
    $W[u, v] \leftarrow sim\_txt(u.u, v.u)$
end for
%pruning the graph based on mean similarity of S
$\overline{E} \leftarrow mean\_similarity(S)$
for each $u, v \in L$ do
    if $W[u, v] \geq \overline{E} * \beta$ then
        $W'[u, v] \leftarrow 1$
    else
        $W'[u, v] \leftarrow 0$
    end if
end for
%computing a biased centrality
$B \leftarrow BiasedPageRank(W', b)$
$N \leftarrow \{\}$
%finding components that contain S
for each $s \in S$ do
    $Q \leftarrow \{s\}$
    while there is an edge $(u, v)$ where $u \in Q$ and $v \notin Q$ do
        $Q \leftarrow Q \cup \{v\}$
    end while
    $N \leftarrow N \cup Q \cap s$
end for
%reordering $N$ according to their centrality
$O \leftarrow sort\_by\_centrality(N, B)$
Return $O$
```

3.1 Similarity between websites

News websites usually provide a long link to their News articles which contains the headline of the News, and this link is a good summary of the News article content. For instance, Table 1 gives two examples of long links to News articles and their headlines. DistrustRank only takes into consideration the terms (i.e. words) extracted from the long links, represented as unigrams weighted by Term *Frequency-Inverse Document Frequency* (TF-IDF) in order to compute the similarity of pairs of websites. This choice is motivated by performance issues, since for a fast and scalable method, we must be able to handle big graphs and the extraction of features for comparison cannot be time-consuming. Crucially, to use only the links instead of the full articles content is a good strategy. In this way, DistrustRank can easily be integrated to search engines, as it does not need additional features.

Therefore, we define the similarity between websites as the cosine similarity of News headlines, represented by their respective TF-IDF vectors, as detailed in Equation 1.

$$f(u, v) = sim\_txt(u, v) \tag{1}$$

where $sim\_txt \in [0, 1]$ represents the cosine similarity between the *TF-IDF* vectors of two websites u and v.

Table 1: Reliable News' URLs, their headlines and Extracted Terms

URL	News Headline	Terms extracted from URL
www.nydailyNews.com/new-york/education/bronx-teacher-sparks-outrage-cruel-slavery-lesson-article-1.3793930	Bronx teacher sparks outrage for using black students in cruel slavery lesson	[new-york, education, bronx, teacher, sparks, outrage, cruel, slavery, lesson]
www.nypost.com/2018/02/01/mom-teams-up-with-daughter-to-fight-girl-on-school-bus/	Mom teams up with daughter to fight girl on school bus	[mom, teams, up, with, daughter, to, fight, girl, on, school, bus]

3.2 Similarity Threshold (β)

Since centrality in our approach is highly dependent on significant similarity, we can disregard websites links which the similarity scores are below a minimum threshold. However, setting an appropriate threshold is a tricky problem [19]. While a high threshold may mistakenly consider as similar websites that have very little in common, conversely, a low threshold may disregard important links between websites.

Using Equation 2, we prune the graph based on a minimum similarity between websites. The result is a weighted graph represented by the adjacency matrix W', where $W'(u, v)$ assumes 1 if an edge that connects u and v exists, and 0 otherwise. To tune our results, we employ a base threshold β that varies according to the mean similarity of false News websites.

$$W'(u, v) = \begin{cases} 1, & f(u, v) \geq \overline{E} * \beta \\ 0, & otherwise \end{cases} \tag{2}$$

In Equation 2, $f(u, v)$ is the similarity score according to Equation 1; $\overline{E}$ is the mean similarity of the News website dataset, and β is the base threshold.

3.3 Biased Centrality

While a regular version of PageRank algorithm computes a static score to each website, a biased version of PageRank [6] can increase artificially the score of some specific websites. A vector of scores is employed to assign a non-zero static bias to a special set of websites. Then the biased PageRank spreads the bias during the iterations to the pages they point to. The matrix equation of Biased PageRank is:

$$r = \alpha * T * r + (1 - \alpha) * b \tag{3}$$

where b is the bias vector of non-negative entries summing up to one, r is the final centrality score, T is the transaction matrix and α a decay factor for bias.

DistrustRank employs a bias to the selected set of seeds (false News websites) which will be spread to their neighborhoods (similar websites). The intuition behind this approach is that we can reduce the 'distrust' score as we move further and further away from the bad seed websites.

Once the centrality scores are computed, we perform the breadth-first search (BFS) on a network graph, starting at some particular node $v \in Seeds$, and explore the neighbor nodes first, before moving to the next level neighbors. The centrality index of the neighbors of v is used to compose the final rank.

4 EXPERIMENT DESIGN

In this section, we detail the experimental setting used to evaluate DistrustRank. We describe the dataset used, the methods employed for comparison and the metric applied for evaluation, as well as details about DistrustRank parameterization.

4.1 Datasets

In order to evaluate our approach, we created two different data sets containing reliable and unreliable News extracted from true News websites and prominent fake News websites, as follows:

- **Reliable**: we extracted the most popular News websites from 10 different categories indexed by SimilarWeb[2]. SimilarWeb provides a ranking of the top world News websites in different categories. The categories used in this set are *Automotive, Celebrities and Entertainment, Sports, News and Media, Newspapers, Business, College and University, Weather, Technology, Magazines and E-Zines.* From each of these categories of News we used the 100 first most popular websites.
- **Unreliable**: The unreliable News websites were extracted from the Wikipedia's list of prominent fake News[3]. The total of websites in this list is 47, which represents nearly 5% of the total of reliable News sources used in this experiment.

For all websites in both data sets previously listed, we used Internet Archive in order to extract the links to their News articles. Figure 2 depicts the distribution of the URLs collected in this task. However, not all of these websites were employed in the evaluation process, since we are just interested in reliable News that could

[2] https://www.similarweb.com/top-websites/category/News-and-media

[3] https://en.wikipedia.org/wiki/List_of_fake_News_websites

provide a fair evaluation. For that, we performed a pre-selection of the Reliable News according to the following aspects:

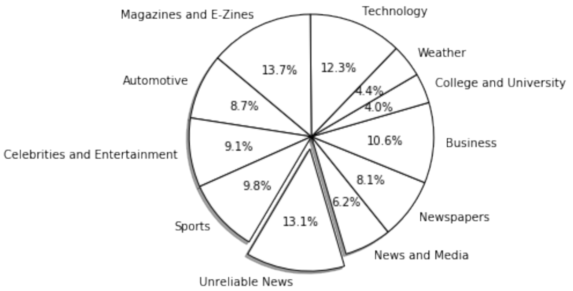

Figure 2: Distribution of collected URLs per category of News, where the categories were extracted from similarweb.com

(1) We only used reliable News articles that are similar to our unreliable News data set. This choice is motivated to make sure that our approach is able to identify fake News in a set of similar News, since it increases the difficulty of the task and makes it comparable to a real-world problem. Therefore, we compared the intersection of the two sets using the Jaccard similarity coefficient[11]. Some News categories, namely *Weather*, *College and University* and *Automotive*, did not achieve a minimum similarity (> 0.4) and therefore were not used in our final data set. Figure 5 shows the similarity between Reliable News categories and our Unreliable News data set.

(2) We only used URLs where the extracted headline contains more than 3 words recognized by an English Dictionary [4]. We only considered headlines extracted from the long links, because less than 3 words links do not provide enough information to provide a right classification. Figure 4, shows the distribution of URLs per website.

(3) we only used News Articles that were published after 2010. This ensures an evaluation that uses a broad scope of News, increasing the diversity of the vocabulary, therefore making the problem harder. Figure 3 shows the distribution of the News over the years used in this work.

Table 5 provides some statistics about the final data set employed in this work. From the initial 1000 Reliable News websites collected, we ended-up with 502 in accordance with our requirements previously described.

4.2 Validation

We adopted k-fold cross-validation, where the unreliable sample is randomly partitioned into k equal size subsamples. For each fold, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times, where at the end of the process all instances in the unreliable set are used for both training and validation, and each observation is used for validation exactly once. The full reliable data set, which contains unlabeled websites, is used to construct the Graph in all the folds. Finally, the mean precision is computed by using the precision of the k results from each fold, producing a single estimation.

Table 2: Summary of the reliable and Unreliable News websites used in this work.

	Domains	URLs (News)	Terms	URL/Terms
Unreliable	47	37320	158501	4.24
Reliable	502	396422	1281794	3.23

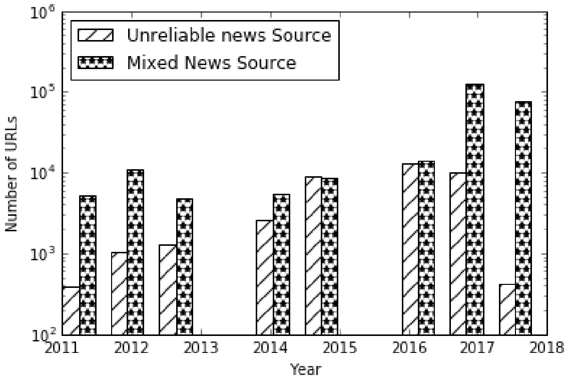

Figure 3: Year's distribution of collected News, ranging from 2010 to 2018.

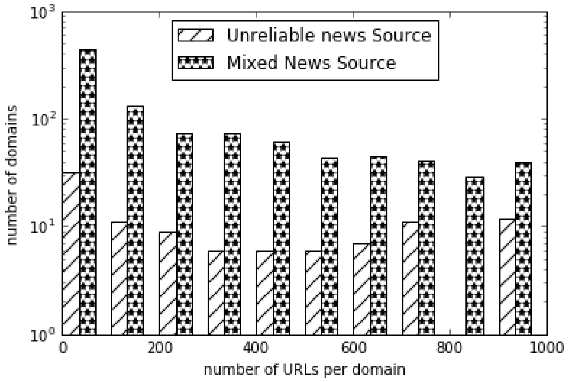

Figure 4: Distribution of URL's number collected per domain.

4.3 Defining a Similarity Threshold (β)

The parameter β has influenced the results obtained. In order to get more accurate rank, we estimated the best parameter using

[4] https://www.abisource.com/projects/enchant/

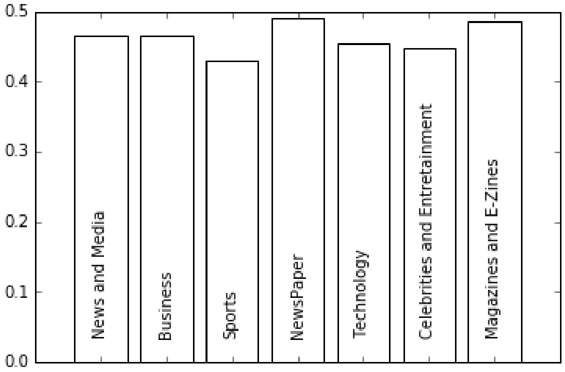

Figure 5: Jaccard Similarity between News categories and fake News that achieve the minimum similarity (>0.4).

a numerical optimization method. We used Newton-Conjugate-Gradient, which is employed to minimize functions of multiple variables, where the best β found that minimizes the Mean Squared Error in our dataset is $\beta = 0.849$.

4.4 Metrics

In order to evaluate our proposed work in a classification task we adopted the standard information metrics, such as precision, recall and f1. For the assessment of the ranking task, we used precision@k. The metrics employed ca be briefly described as follows:

- *Precision:* the fraction of the websites classified as fake that are really fake News. $Precision = \frac{tp}{tp+fp}$
- *Recall* is the fraction of the fake websites that were successfully identified. $Recall = \frac{tp}{tp+fn}$
- *F-1* corresponds to the harmonic mean between precision and recall. $f1 = 2 * \frac{precision*recall}{precision+recall}$
- *Precision@k* corresponds to the precision using the k-firsts elements of the rank.

where tp is the number of positive instances correctly classified as positive, tn number of negative instances correctly classified as negative, fp negative instances wrongly classified as positive, and fn is the number of positive instances wrongly classified as negative. We defined positive instances as fake News websites and negative instance as reliable News websites.

4.5 Baseline

To measure the gap between our method and a supervised one, we compared our results with the ones using Support Vector Machine, referred to as *SVM*. We employed a linear kernel, recommended for text classification, and which generally uses TF-IDF vectors with a lot of features.

5 RESULTS AND DISCUSSION

In this section, we present the results and discuss the evaluation of our proposed approach in two different tasks: Ranking of websites and Binary classification.

5.1 Ranking Task Assessment

To perform a comparison between the ranking of websites generated by DistrustRank and SVM, we used Precision@10, i.e. we evaluate the precision of the models using the top-10 firsts elements of the rank. We adopted Precision@10 because it usually corresponds to the number of relevant results on the first page on a search engine (e.g. google.com). Additionally, in order to better understand the behavior of the models in a small training data set, we vary from 0 to 10 the quantity of websites. It is important to note that for the training step, each model received exactly the same seed set (that was randomly selected from the training set).

Figure 6 shows that Distrustrank yields better results for all quantity of seeds analyzed, which are excellent results for a semi-supervised model. While DistrustRank needs only 7 seeds to achieve a precision of 100% (i.e. all the top 10 websites ranked are truly fake), SVM needs 9 seeds to obtain the same precision. Additionally, all the results obtained by our approach using different quantities of seeds showed to be superior to the baseline, where the difference ranges from 10 to 40 percentage points (pp). Using a Wilcoxon statistical test [18] with a significance level of 0.05, we verified that DistrustRank results are statistically superior in this task.

As a matter of fact, the good performance of Distrustrank in this task is expected. Supervised learning strategies generally need a large training data set to yield models with higher predictive power that can generalize well to a new data set. DistrustRank however, is designed considering the empirical observation, that fake News websites are similar to each other. The use of this domain knowledge in our model, trough a biased graph centrality, allows a better performance in small data sets.

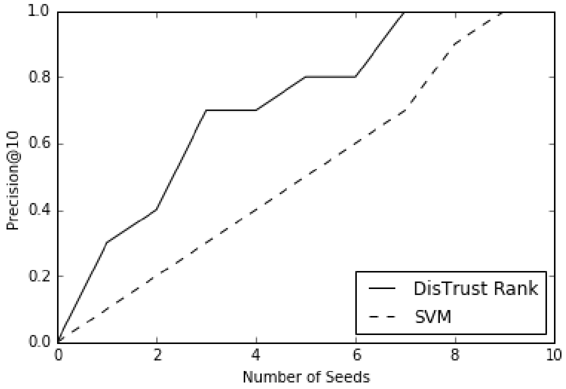

Figure 6: Number of seeds used to train the model.

5.2 Classification Task Assessment

This task consists in predicting the class of a website (e.g. Reliable or Unreliable News website). In this experiment, the positive class represents the Unreliable News websites and, the Negative class represents Reliable News websites. We used 2-fold cross-validation, where we randomly shuffle the data set into two sets d_0 and d_1 with equal size. We then train on d_0 and validate on d_1, followed by training on d_1 and validating on d_0. This choice is motivated by the lack of positives instances of fake News websites.

DistrustRank was originally designed to rank websites, however, in order to provide a proper evaluation against SVM, we adapted the ranking to act as a classifier. In a classification task, we are able to compare our approach using complex metrics, such as ratios of false positive and negatives, true positives and negative, as well as precision, recall, and f-1. To transform a ranking into a classification, we used the first k-top websites of DistrustRank's rank as positive class and the rest of the rank as negative. Obviously, setting an optimal value of k without a priori knowledge of the distribution of fake news websites is a tricky problem. Nonetheless, for evaluation purposes, we use k=47, since we know a priori that this is the number of fake websites in our data set.

Table 3 and 4 show that DistrustRank presented a lower error rate in both positive and negative classes, where the differences range from 38 to 47 pp. Additionally, we also analyzed the performance of the models using standard Information Retrieval metrics. Table 5 shows that DistrustRank outperformed the SVM model in Precision, Recall and f1, where differences are 16.96, 14.89 and 15.89 pp, respectively.

Table 3: Confusion Matrix of DistrustRank.

	Predicted Positive	Predicted Negative
Actual Positive	**36**	11
Actual Negative	9	**493**

Table 4: Confusion Matrix of SVM.

	Predicted Positive	Predicted Negative
Actual Positive	**29**	18
Actual Negative	17	**485**

Table 5: Summary of Results

	Precision	Recall	F1
DistrustRank	**0.8**	**0.7659**	**0.7825**
SVM	0.6304	0.6170	0.6236

The SVM model presented a similar error distribution among positive and negative. This was expected, since for the learning step we used an equal quantity of positive and negative instances, and that it generally leverages in a learning of an equal distribution of the classes. However, even using a larger amount of data for the training, it still presents lower precision and recalls when compared to our approach. In our experiments, we observed that the vocabulary employed by fake News is similar to the one used in reliable News. This textual similarity explains the worst results of the supervised learning model. On the other hand, DistrustRank presented better results using the same amount of data to the training step, due to its semi-supervised strategy.

6 CONCLUSION AND FUTURE WORK

In this paper we have put forward a novel semi-supervised approach to spot fake News website: DistrustRank. From a small set of fake News website, it creates a graph where vertices correspond to websites and edges to the similarity between the News that they share. Next it applies a biased Pagerank to identify other fake websites. The similarity is defined in therms of cosine difference of the TF-IDF vectors of words extracted from the News links, which usually contains the headline.

Our evaluation showed that DistrustRank can effectively identify a significant number of unreliable News (fake News) websites with less data to the training step. In a search engine, DistrustRank can be used either to filter the pages retrieved to the user, or in combination with other metrics to rank search results. The main contributions of this work are the following:

(1) a new semi-supervised method to identify Unreliable News Websites, i.e. it does not depend on a large annotated training set;
(2) formulation of a similarity function that is computational inexpensive since it only relies on links to represent the similarity between websites;
(3) a better performance in the tasks of ranking and classification, using only a small set of unreliable News websites;
(4) creation of pre-selected data set, containing the News category, date and similarity content; this final data set contains News websites, long links to the News and their headlines.

As future work, we would like to consider different ways to measure the similarity between websites. One possible way is using Word Embedding [10]. It provides a vector representation that allows words with similar meaning to have a similar representation. For instance, this representation could be applied to News links that contain different terms but the same semantic meaning: e.g. *killer* and *murderer*. Another research direction would be to employ different features, such as the time of each News as a decay parameter to measure the similarity between nodes.

ACKNOWLEDGMENTS

This work was partially funded by the European Research Council under ALEXANDRIA (ERC 339233) and CAPES, a Brazilian government institution for scientific development.

REFERENCES

[1] Eda Baykan, Monika Henzinger, and Ingmar Weber. 2013. A comprehensive study of techniques for URL-based web page language classification. *ACM Transactions on the Web (TWEB)* 7, 1 (2013), 3.
[2] Shiri Dori-Hacohen and James Allan. 2013. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management.* ACM, 1845–1848.

[3] Juan Echeverria and Shi Zhou. 2017. Discovery, Retrieval, and Analysis of the'Star Wars' Botnet in Twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 1–8.

[4] Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher* 40, 5 (2011), 223–234.

[5] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 729–736.

[6] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 576–587.

[7] Benjamin D Horne and Sibel Adali. 2017. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398* (2017).

[8] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 591–602.

[9] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. 2012. Truth finding on the deep web: Is the problem solved?. In *Proceedings of the VLDB Endowment*, Vol. 6. VLDB Endowment, 97–108.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[11] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 1.

[12] Michael J Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 66–76.

[13] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2173–2178.

[14] Pujari Rajkumar, Swara Desai, Niloy Ganguly, and Pawan Goyal. 2014. A novel two-stage framework for extracting opinionated sentences from news articles. In *Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing*. 25–33.

[15] Shafiza Mohd Shariff, Xiuzhen Zhang, and Mark Sanderson. 2017. On the credibility perception of news on Twitter: Readers, topics and features. *Computers in Human Behavior* 75 (2017), 785–796.

[16] Tarcisio Souza, Elena Demidova, Thomas Risse, Helge Holzmann, Gerhard Gossen, and Julian Szymanski. 2015. Semantic URL Analytics to support efficient annotation of large scale web archives. In *Semanitic Keyword-based Search on Structured Data Sources*. Springer, 153–166.

[17] Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating Deep Linguistic Features in Factuality Prediction over Unified Datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 352–357.

[18] Frank Wilcoxon, SK Katti, and Roberta A Wilcox. 1970. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. *Selected tables in mathematical statistics* 1 (1970), 171–259.

[19] Vinicius Woloszyn, Henrique DP dos Santos, Leandro Krug Wives, and Karin Becker. 2017. Mrr: an unsupervised algorithm to rank reviews by relevance. In *Proceedings of the International Conference on Web Intelligence*. ACM, 877–883.

[20] Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 129–136.