Real or Not?

Identifying Untrustworthy News Websites using Third-Party Partnerships

Abstract

Untrustworthy content such as fake news and clickbait have become a pervasive problem on the Internet, causing significant socio-political problems around the world. Identifying untrustworthy content is a crucial step in countering them. The current best-practices for identification involve content analysis and arduous fact-checking of the content. To complement content analysis, we propose examining websites' third-parties to identify their trustworthiness. Websites utilize third-parties, also known as their digital supply chains, to create and present content and help the website function. Third-parties are an important indication of a website's business model. Similar websites exhibit similarities in the third-parties they use. Using this perspective, we use machine learning and heuristic methods to discern similarities and dissimilarities in third-party usage, which we use to predict trustworthiness of websites. We demonstrate the effectiveness and robustness of our approach in predicting trustworthiness of websites from a database of News, Fake News, and Clickbait websites. Our approach can be easily and cost-effectively implemented to reinforce current identification methods.

Keywords: Website third-parties, untrustworthy websites, prediction, machine learning, heuristics

"Tell me with whom you walk and I will tell you who you are." - Spanish proverb.

The rise of untrustworthy websites, such as fake news and clickbait, has become an important and pervasive problem for society. These websites negatively impact the experience of Internet users and ultimately the proper functioning of democratic societies [Akpan 2016]. Fake news websites publish fake content in support of an agenda, campaign, or other goals, irrespective of the truth. Clickbait websites provide enticing content, often untrue or inaccurate, that focuses on the curiosity gap of visitors to get them to click on their links. The goal of such websites is either monetization, relying on getting Internet traffic to their website, or advancing of an agenda, usually political in nature. Both fake news and clickbait websites are known to be responsible for the spread of misinformation, albeit with different objectives[1].

While there have been efforts in detecting and flagging fake news and clickbait content, it has proven difficult to accurately and cost-effectively do so [Shellenbarger 2016; Leetaru 2016]. The presence of fake news and clickbait is especially problematic in social networks, where users can share links to untrustworthy websites. Fake news and clickbait are known to go "viral" and rapidly propagate through social networks, given that they adhere to what the public deems interesting. In response, social network sites such as Facebook and Twitter have taken measures to combat fake news. The social network industry primarily analyzes the content of posts to detect untrustworthy news. [Pogue 2017] reports that Facebook uses a range of measures, such as reporting tools and external fact-checking services to detect fake news. These methods rely on humans to identify each article or post, which can be time-consuming and costly.

[1] https://www.engadget.com/2016/11/21/clickbait-fake-news-and-the-power-of-feeling/

Both academia and industry have attempted to detect and flag fake news and clickbait content on the Internet. Chen, Conroy and Rubin, use content cues such as semantics, as well as non-text cues, such as image analysis and user behavior, to recognize clickbait content [Chen et al. 2015]. They use a variety of machine learning methods to perform analysis on these cues. Chakraborty et al. develop a linguistic classification method to classify clickbait and non-clickbait headlines [Chakraborty et al. 2016]. Anand, Chakraborty and Park use a neural network on word representation in headlines to detect clickbait content [Anand et al. 2017]. Abbasi et al. pursue a slightly different approach where the authors use statistical learning theory to detect fake websites (different from fake news websites), using cues such as web page text, source code, URLs, images, and linkages [Abbasi et al. 2010]. These studies focus on the content of the websites. Existing methods do not distinguish real news websites from fake news and clickbait websites, but rather evaluate trustworthiness at the article level. Instead of content analysis, we propose an approach that uses the third-party partnership data, also known as the digital supply chain, to identify trustworthiness of websites. This approach can complement content-based methods to more accurately identify trustworthiness of websites.

Websites are composed of more than just content; they employ an infrastructure network of third-parties as their partners for operation. Websites use third-parties for a variety of purposes, such as increased functionality, performance improvement, and most commonly for advertising. Third-parties are the digital supply chain of a website, and therefore, an indication of the business model of the website [Gopal et al. 2018]. In this paper, we argue that third-parties, as the digital supply chain, can provide important information to help distinguish trustworthy (real news) and untrustworthy (fake news and clickbait) websites. Based on this, we set out to provide answers to the following research question: *Given the differences and similarities in*

third-party usage between websites, can we use third-party partners to identify whether a website is trustworthy? We use a framework based on business partner selection to posit that third-parties reveal important information about the type of websites and, more importantly, their trustworthiness. We then use a design science approach to propose machine learning and heuristic methods that can effectively and efficiently identify the trustworthiness of news websites. Our results on a dataset of trustworthy and untrustworthy websites show the power of our proposed methods in identifying website trustworthiness at the source (website) level.

1. Framework

Websites on the Internet extensively use third-parties, which form a significant part of their infrastructure and partner network, and consequently, their business model. Third-parties provide components on a website that are not managed by the website, and are utilized for purposes such as functionality (e.g. video streaming) and performance (e.g. content delivery), and most predominantly, targeting and advertising, which enable websites to monetize their visitor traffic. These third-parties form the digital supply chain of a website. While third-parties may not be visible to viewers on the websites, browsing tools can observe these third-parties. Studies exist on how third-party usage can be an indication of website behavior. For example, Gopal et al. study the extent of third-party usage among popular websites and find that websites in different industries have varying levels of third-party usage [Gopal et al. 2018].

We expect the third-party usage by websites to reflect their business model. Third-parties are the network partners providing an infrastructure upon which the website operates. Therefore, third-parties used by the website indicate its core business and value propositions. Due to the underlying requirements, some third-parties are commonly shared by customers in one industry sector and not others [Gopal et al. 2018]. Therefore, we can identify common third-parties used

in the news industry and, based on presence or absence of these third-parties, predict a website's trustworthiness. On another hand, fake news and clickbait websites need to maintain an appearance of legitimacy by mimicking the operation of real news websites, including third-party connections. However, this mimicking behavior can be difficult for two reasons – budget limitations and mutual selection between business partners – as some reputed third-parties may refuse to work with untrustworthy websites and it is hard to form new partnerships. Thus, we expect that the third-parties of fake news and clickbait websites would not fully replicate those of trustworthy websites. Next, we discuss the relationship between website business model and third-party structure in more detail.

Each industry has its own unique strategic needs. A firm's corporate strategy is greatly influenced by the particular needs of its industry, including customer preferences, available revenue streams, available supply chain partners, regulations, and environmental and technological constraints. A rational firm uses supply chain partners that best satisfy its unique strategic needs. While each firm's strategic needs may be unique, common requirements particular to each industry exist, that lead to the development of specialized supply chains catering to these unique requirements. It takes time and effort for firms to form trust in their providers and create supply chains, therefore, these supply chains are difficult and costly to change. These supply chain networks are due to the self-selection and embeddedness that is needed to cater to those requirements. Some examples include fine-grained information, for example in case of payment transfers [Uzzi 1993; Uzzi 1996]. In other words, while fake news and clickbait websites can disguise their content to look like a real news website, the third-parties they use reveal their untrustworthiness.

This line of reasoning undergirds our study and leads us to assume that supply chain partnerships reflect websites' business model. This idea has been extensively studied in the literature. One important stream in social network research, the embeddedness perspective, supports the notion that the general pattern of economic activities of a firm in continuous social relations could be explained by the characteristics of being part of a larger social structure [Choi and Kim 2008]. Polanyi and MacIver were perhaps the first to identify the social structures of modern markets [Polanyi and MacIver 1944], and Uzzi and Choi and Kim later charted this path [Choi and Kim 2008; Uzzi 1997]. Subsequent authors such as Schumpeter, Granovetter, and Uzzi examine the impact of inter-firm interconnectivity (i.e., the interfirm networks) on the economic action [Uzzi 1997; Schumpeter 2010; Granovetter 1985]. Choi and Kim state "... the concept of embeddedness refers to the contextualization of economic activity in ongoing patterns of social relations and captures the contingent nature of an economic actor's activities by the virtue of being embedded in a larger social structure" [Choi and Kim 2008]. We extend this logic to posit that third-parties are part of a website's larger social structure, reflecting the nature of the website's activities. According to Bowersox, Closs, and Stank, firms doing business with particular supply chain partners do so to have competitive advantages in their markets [Bowersox et al. 2003]. Thus, the operational objectives and methods of a website would, to an extent, reflect their supply chain of third-parties. We refer to the third-party supply chains as digital supply chains, where content and services are supplied from upstream third-parties to downstream websites.

The relationship between business model and supply chain partners applies to the untrustworthy and dishonest companies as well. [Sullivan et al. 2007] establish a link between illegal/unethical acts on interfirm networks, where such acts drive higher quality firms to leave

the network and be replaced by lower quality firms. The authors consider the impact of such incidents on the performance and growth of companies. [Chandler et al. 2013] study the impact of firm status (reputation) on the quality of interfirm partnerships and find that status has a positive impact on the quality of firms within a firms' network, and a negative impact on the diversity of the firms. In this paper, while we do not consider the effect over time of unethical and illegal actions that untrustworthy websites maneuver, we observe that the main purpose of the websites has a significant impact on their interfirm network in the form of third-parties, which we use to identify trustworthiness of websites.

While we discuss the third parties as a supply chain network, our work differs in approach from traditional social network analysis, which has been applied in areas such as fraud detection [Šubelj et al. 2013; Gopal et al. 2012]; impacts of unethical behavior on organizational networks [Sullivan et al 2007; Zuber 2015]; and study of criminal networks [Xu and Chen 2005; Hutchins and Benham-Hutchins 2010]. As a general approach, these works look at social network structure metrics such as cliques, centrality, and connectedness. Our approach focuses on the analysis of choice and composition of third-parties with an immediate connection to the website in question to identify its trustworthiness.

Our methodology is independent of the mechanisms that drive third-party selection. For example, one may consider the possibility that the differences between third-parties are due to the popularity of websites, or the quality of the services provided by third-parties. While popularity and quality may be correlated with trustworthiness, this does not impede our core analysis objective of identifying whether a website is trustworthy. Moreover, while third-parties are chosen by websites, third parties can also choose whether to provide services or content to any given website. For example, some third-parties such as Google and Facebook do not

advertise on websites that operate beyond the norms of society[2]. Several complicated and interesting underlying factors exist that drive the digital supply chains. While a detailed analysis of these factors is beyond the scope of analysis, we conduct robustness analysis which reveals that our approach continues to perform well even after accounting for popularity differences.

We can apply our proposed methodology to a wide variety of website types, as the third-party relationships capture the underlying business model of websites and can be used for identification in a number of application settings. While it is interesting to investigate websites other than news using a similar methodology, due to the significance of identifying fake news and clickbait websites to the society, the focus of the current paper is on predicting the trustworthiness of such websites.

2. Data and Research Design

We collect data from 3 categories of websites: real news, fake news and clickbait. We use the list of the 500 top news websites provided by alexa.com as of January 2016 for real news websites. A similar dataset for real news websites is used in [Gopal et al. 2018]. For fake news and clickbait websites, we use "False, Misleading, Clickbait-y, and Satirical News Sources" list provided by Harvard University Library [Zimdars 2016; Harvard 2017]. For clickbait websites, we choose the websites that are categorized as clickbait, but not fake. Similarly, for fake news websites, we chose the websites which are categorized as fake but not clickbait. While we initially treat these as separate categories, later in the analysis we combine them into one category, noted as *untrustworthy* due to significant similarities between the two in their digital

[2] https://www.forbes.com/sites/zarkodimitrioski/2019/02/25/big-companies-freezing-their-advertising-on-youtube-because-of-controversial-comments/#186f04973982 and https://www.washingtonpost.com/business/technology/controversial-digital-ad-placement-leaves-tech-companies-scrambling/2013/08/21/3609306e-04d4-11e3-9259-e2aafe5a5f84_story.html

supply chains. At the time of data collection, there were 74 clickbait-only websites and 66 fake-only websites (and 7 websites categorized as both fake news and clickbait, which we removed from our sample). In classification problems such as ours, unbalanced data sets (where the number of observations from each class are not equal), cause issues in the analysis. Therefore, we decided to include the same number of websites from each category. Accordingly, we choose to collect data from 50 websites in each category. We randomly select 50 fake news and 50 clickbait websites to use in the analysis. In Section 3, we show that fake news and clickbait categories are similar in terms of their digital supply chains. Therefore, in our analysis, we consider both to be part of a broader category of untrustworthy websites.

For real news websites, we find the 464 unique websites from the Alexa 500 top news websites list. We discarded 32 websites that are repeated and 4 websites that are classified as news, but for which the primary purpose of the website is not news (google.com, copyright.gov, purdue.edu, and harvard.edu). We collected third-party data for the real news websites in December 2017 and January 2018. After the data collection process, we discarded 6 real news websites that do not engage any third-party. These are mostly static websites that present information with minimal interactive features. At the end of this filtration, we are left with 458 unique websites to analyze. To perform the analysis on a balanced dataset, we split these 458 websites to bins of size 50 in our analysis as detailed later in this section.

We use a script (using Mozilla Firefox 55.0 and Lightbeam add-on version 1.3.2) to visit the websites and collect the third-party data. The browsing is done in Firefox with default settings, i.e. without the use of Do Not Track or ad blocking. The websites are loaded one by one, and each page is visited for the page loading time plus a 5-second wait time to allow for the third-parties to load. If the page fails to load completely, we visit the page for 20 seconds. We

collect the data in batches and refer to completing this process for all websites in a batch as a single run. This process is repeated 51 times for each batch of websites, for 51 runs. Because cookies are set on the computer in the first run in which a website is visited, we omit the first run data from our analysis, and analyze the remaining 50 runs. Prior to the first run, we clear the cookies. For each run, we browse the list of websites in random order. The data used for analysis is the list of any third-parties used in any of the 50 runs. Using this process, for each website, we collected data on all third-parties used (third-party connection addresses), content type (e.g. text, picture, etc.), whether a cookie is used, along with some other attribute for all 50 runs. The attribute we use in this study for identification of trustworthiness is the third-parties used on each website[3].

We consider the difference in website lists for real news (Alexa) versus clickbait and fake news websites (Harvard University Library) and whether this impacts our findings. While we gather the lists of websites from different sources, we use the same methodology to collect the third-party data of each website. Therefore, we do not expect to see any effects in our gathered data due to the sources of the lists being different; we believe the two sources to be an accurate depiction of the trustworthiness and untrustworthiness of news websites.

Figure 1 summarizes our observations of third-party usage across the three website categories. Real news websites have the highest average number of third-parties used, followed by clickbait and fake news.

[3] The full data, including additional attributes, using this method can be found at Lightbeam's documentation: https://github.com/mozilla/lightbeam/blob/master/doc/data_format.v1.1.md

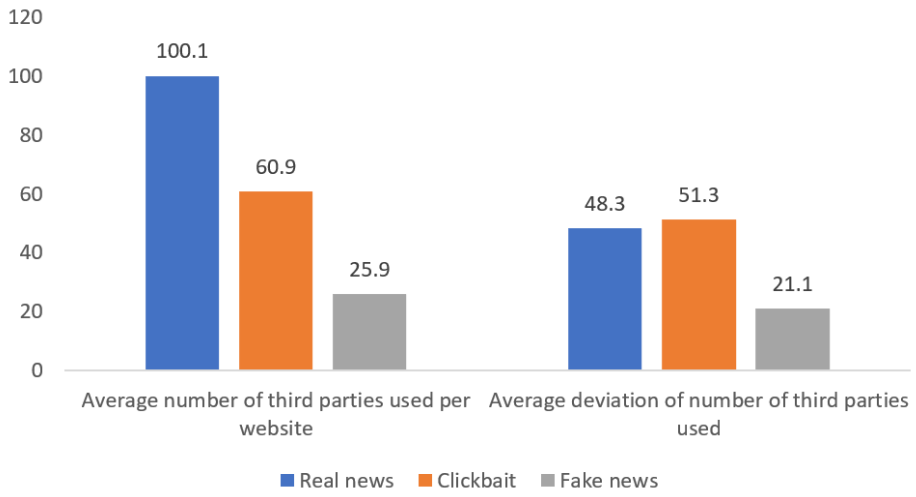

Figure 1: Third-party partners by website industry sector

As mentioned before, the number of real news websites (458) is much higher than fake news (50) and clickbait (50) websites. Out of the 458 real news websites, in each run, we randomly select 8 websites to be omitted from the training and used in testing only, giving us 9 sets of 50 real news websites to analyze in balanced data sets, i.e. 50 websites of each of real news, fake news, and clickbait[4]. We use a 10-fold cross-validation for our analysis to sufficiently account for any anomalies in the results due to sample bias. We create 9 folds of data with real news websites data. Each fold gets a turn in the training set as described above for the 10-fold cross-validation data set. With 50 real news, 50 fake news, and 50 clickbait websites, 9 different real news website sets get their turn as the 10-folds cross-validation data set. For each training fold for the real news set, we also perform extended validation on the remaining real news websites that serve as a holdout sample for each training set. For each of the 9 training sets with 50 real news websites, there are 450-50 = 400 excluded real news websites in the holdout sample. Thus, the category of every real news website is predicted 81 times while the 8 websites

[4] Moreover, in order to analyze the effect of website popularity on our analysis, we bin the real news websites according to rank. This analysis is provided in Robustness Check subsection.

that were randomly omitted from the training, will be predicted 90 times. The category of every fake news and clickbait website is predicted 9 times due to the relatively fewer number of identifiable websites available. For fake news and clickbait websites, no additional holdout sample exists, so we use the predictions made when those websites are members of the validation portion of the 10-folds cross-validation. With these preliminary predictions, we make a final prediction. If the number of predictions as real news is higher than the number of predictions as fake news and clickbait, then the website is classified as trustworthy. Otherwise, we classify it as untrustworthy. While our method enables us to categorize websites into three categories, due to the similarity of fake news and clickbait, we report our findings for trustworthy (real) and untrustworthy (fake news and clickbait) categories only.

3. Prediction Methods

We use and compare a combination of existing, proposed, and ensemble methods to predict the categories of the websites in our dataset. These methods are explained below.

3.1 Existing Methods

To compare the relative performance of our classification methods to what is currently available, we compare our results to 8 fake news detectors available on the internet as follows. B.S. Detector and Fake News Alert for Chrome are two recommended browser plugins by Harvard Library Guide [Harvard 2017]. Fakenewsai.com is a website to detect fake news websites. Open Mind[5], Fake News Detector (F)[6], Fake News Detector (H)[7], Fake News Blocker[8] and ZenMate

[5] https://openmind.press
[6] By DareDevelopers-giacomofava
[7] http://fakenewsdetector.org/
[8] By Floris de Bijl

Safe Search[9] are other fake news detection plugins selected from the Chrome web store. These existing methods categorize news at the article (page) level, while our analysis is at the source (website) level. We did not find other methods that identify trustworthiness at the source level. Nevertheless, we believe that this comparison exhibits the potential of our proposed methodologies and we discuss the implications later in this paper.

3.2 Proposed Methods

We propose one machine learning method (SVM) and 4 heuristics (CS5, STP, STP-CS5, and STP-SVM). We also present ensemble methods using proposed and existing methods. We develop several heuristic methods, which are based on our theoretical framework of the relationship between third-parties and websites. Some of the heuristics perform better than others in terms of certain measurements, and their performance varies in different scenarios, therefore, it is useful to have multiple methods at hand.

We first provide an analysis of the third-parties from a network science perspective, focusing on the common third-parties used by websites. Our methods are based on the idea that third-parties found in a certain category of websites are more similar to each other compared to third-parties from other categories. This perspective supports our use of supervised classification methods, where we create models based on a training set which can then be used for identifying websites from a test set. In a sense, there are three types of third-parties: 1) those that can be found mostly on trustworthy websites (e.g. adobe.com and ads-twitter.com), 2) those that can be found mostly on untrustworthy websites (e.g. 4dsply.com and ad-score.com), and 3) those that can be found on both (e.g. googleapis.com and google-analytics.com). The third-parties from the

[9] https://chrome.google.com/webstore/detail/zenmate-safesearch-and-fa/banafmipcbekalafkahkalgiodhliinf

first two categories can be useful indicators of the category of websites (trustworthy or untrustworthy). In practice, the third-parties are not so clear-cut and well-defined; therefore, we propose several heuristics that use similarities between third-parties used by categories of websites to identify the website categories.

To better understand the similarities among third-parties found on websites, we analyze the similarities both within and between categories of websites. For this purpose, we use the cosine similarity measure for each pair of websites. We define the third-party usage of each website i as a zero-one vector T^i of size N, where N is the total number of third-parties observed in all websites. For each website i if third-party $r \in \{1, \dots, N\}$ is used, the rth row in vector T^i will be 1 (or T_r^i=1), and if it is not used, the rth row in vector T^i will be 0 (or T_r^i=0). We define the cosine similarity measure between two websites i and j as:

$$s_{ij} = \frac{\sum_{r=1}^{N} T_r^i T_r^j}{\sqrt{\sum_{r=1}^{N} T_r^{i^2}} \sqrt{\sum_{r=1}^{N} T_r^{j^2}}}$$

In order to calculate the overall similarity measures, we calculate the average of cosine similarity among websites both within a single category and between websites of different categories. For a single category, the within-group average similarity for category k is calculated as:

$$SW^k = \frac{\sum_{i \in C_k} \sum_{j \in C_k, j \neq i} s_{ij}}{n^2 - n}$$

Where C_k is the list of websites in category k, and n is the number of websites in each category, that is $n = 50$. We calculate the between category average cosine similarity between two categories k and l as:

$$SB^{k,l} = \frac{\sum_{i \in C_k} \sum_{j \in C_l} s_{ij}}{n^2}$$

Table 1 provides the results for within and between category average cosine similarities. Using a t-test for comparing the similarities between each pair of websites, we find that both the within and between category comparisons of cosine similarities of real news websites are significantly different from both fake news and clickbait websites. On the other hand, we find that fake news and clickbait websites are not significantly different from each other in terms of third-party similarity. Therefore, we provide our results comparing the real news as trustworthy websites, versus untrustworthy websites including both fake news and clickbait websites. This analysis indicates that third-party similarities could be useful for differentiating real news websites from both fake news and clickbait websites, while providing evidence for our perspective that third-party usage depends on the business model of websites, and that websites with similar business models are similar in their third-party usage.

	News	Clickbait	Fake
News	0.23	0.19	0.17
Clickbait		0.30	0.30
Fake			0.30

Table 1. Average within and between category cosine similarity of third-party usage

One of our main goals in this paper is to provide methods for identifying trustworthy and untrustworthy websites. Thinking of this problem as a classification problem, the websites are our observations and third parties form our attributes. Due to the large number of attributes in our data (large number of third parties), use of traditional classification methods is inappropriate. We propose several classifiers that employ the extent of third-party usage across the different website categories for classification. We design basic methods that can demonstrate the effectiveness of using third-parties for classification without much computational complexity. Specifically, we propose two methods: 1) the cosine similarity classifier (CS5), that utilizes

similarities of third-party usage among known websites in each category to classify unknown websites, and 2) significant third-party classifier (STP), that considers the usage of significant third-parties, seen only in websites in a specific category to make classifications for an unknown website. We also develop a Support Vector Machines (SVM) classifier which is a powerful machine learning approach for classification in both linear and non-linear settings. As explained in Section 2, in our tests we use 10-fold cross-validation, where 10% of the data set is used as a test set and the remaining 90% as a training set, using each fold only once.

3.2.1 Cosine similarity classifier (CS5): Our cosine similarity classifier (CS5) is based on the cosine similarities of third-party usage among the test and training sets. In this method, to categorize a focal test-set website, the cosine similarity of this focal website with all other websites in the training set is calculated. Then, we consider the 5 training websites[10] with the highest cosine similarity to the focal test website and calculate the sum of the cosine similarities for training websites from each category. This gives us a similarity measure to each category from the top 5 similar websites[11]. We then choose the category with the highest total similarity, as the predicted category for that test website. We repeat this process for every test website to be categorized.

While rare, it is possible that this method cannot render a classification for some websites, where a tie occurs between the categories for the maximum sum of cosine similarities among the top 5 similar websites. In this case, we use a secondary method to classify those websites as follows: a test website is classified as the category that has the closest average

[10] We found 5 to be a good number of websites in terms of complexity and performance. Our method is robust when using a different number for the websites compared.
[11] We test other similarity measures in Section 4.1.2. Our results are robust to the similarity measure used.

number of third-parties used in the training set to that test website. Figure 2 provides a flowchart outline for the CS5 heuristic.

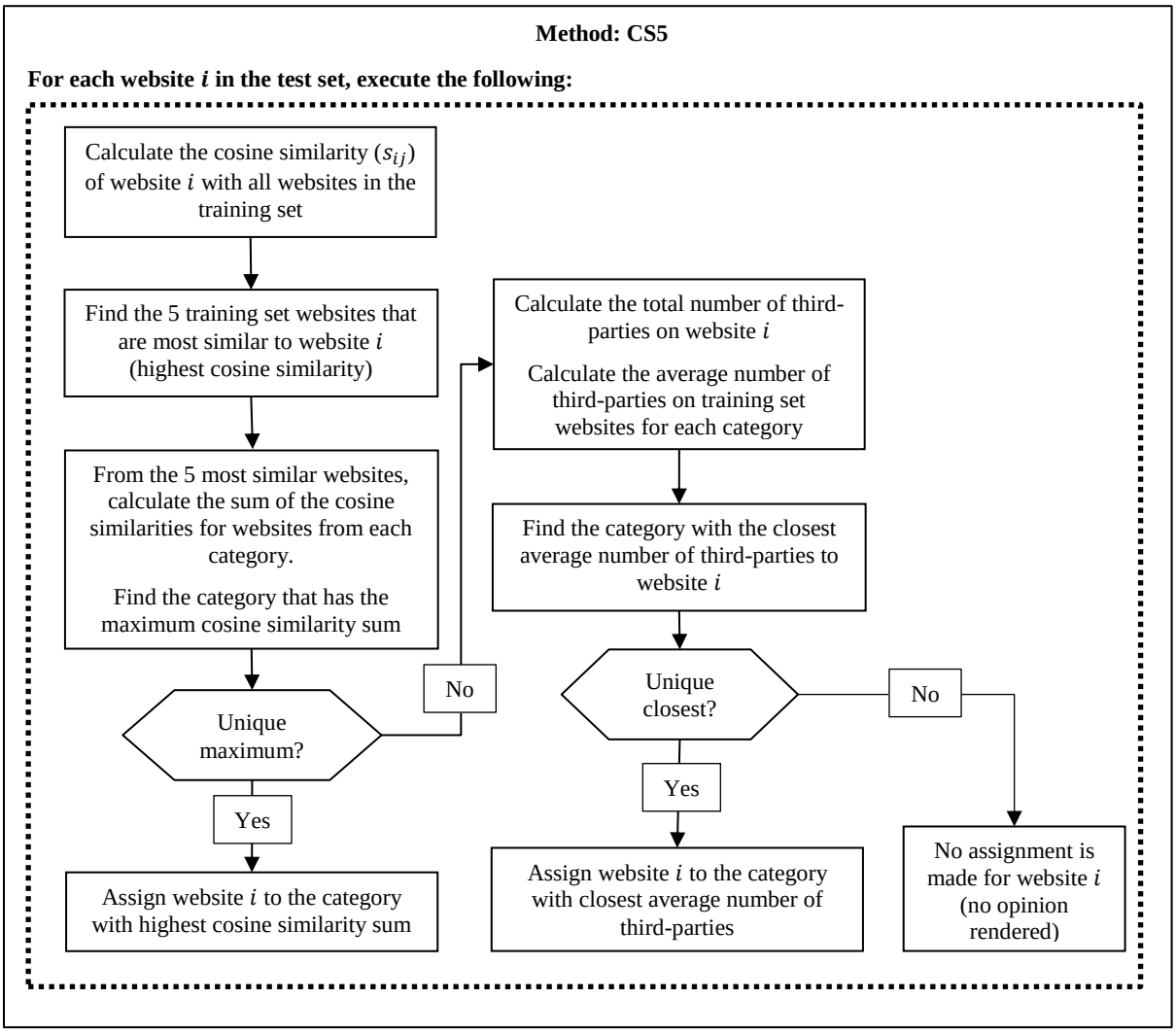

Figure 2. CS5 method flowchart

Pseudo-code for the CS5 method as shown in Figure 2 is presented below:

Begin CS5
Input: test set I (to be categorized) and training set websites $J = \{J^{News}, J^{Clickbait}, J^{Fake}\}$
For *websites* $i \in I$ *(all websites in test set)*
Calculate $s_{ij}, \forall j \in J^k$ *(cosine similarity of that website with all websites in training set)*
Find the 5 training websites with the highest cosine similarity to website i *, call this* $J^{k^{top5}}$
Calculate $Q^k = \sum_{j \in J^{k^{top5}}} s_{ij}$ *(sum of cosine similarities to test website i from each category k)*

If $\underset{k}{argmax}\{Q^k\}$ *has one unique solution (if there is one category* k *which has the highest sum of cosine similarities)*

Assign the test website i *to category* $kmaxQ = \underset{k}{argmax}\{Q^k\}$ *(assign website* i *to category* k *which has the highest sum of cosine similarities)*

Else

Calculate $V^i = \sum_{r=1}^{N} T_r^i$ *(total number of third-parties used by website i)*

Calculate $R^k = \frac{\sum_{i \in C_k} \sum_{r=1}^{N} T_r^i}{n_{Training}}$, *where* $n_{Training} = 45$ *is the number of training set websites in each category (*R^k *is the average number of third-parties used in training websites for each category k)*

Assign the test website i *to category* $kmaxR = \underset{k}{argmax}\{|V^i - R^k|\}$

End If

End For

End CS5

3.2.2. Significant third-party classifier (STP): The significant third-party classifier (STP) method is based on the set of significant third-parties. To obtain the significant third-parties, we performed a t-test for each third-party to see if any particular third-party in the training set is used significantly within a category (hypothesis test of whether or not the mean of usage for each third-party is zero in each category). Within these significant third-parties, we then find the set of third-parties that are used only by trustworthy (real news) training set websites, and the set of third-parties that are used only by untrustworthy (fake news and clickbait) training set websites. To categorize a test website, if it uses any third-parties from the trustworthy-only set and none from the untrustworthy-only set, we then categorize the website as trustworthy. In contrast, if the website uses third-parties from the untrustworthy-only set and none from the trustworthy-only set, we categorize it as untrustworthy. We find 115 significant third-parties that are used only by trustworthy websites (e.g. adobe.com and ads-twitter.com) and 7 significant third-parties, used only by untrustworthy websites (e.g. 4dsply.com and ad-score.com). We find another 151 third-parties that are used in both categories, which are not used by our method for identification. Note that this method, similar to CS5, does not categorize all websites. Figure 3 provides a flowchart outline of the STP method.

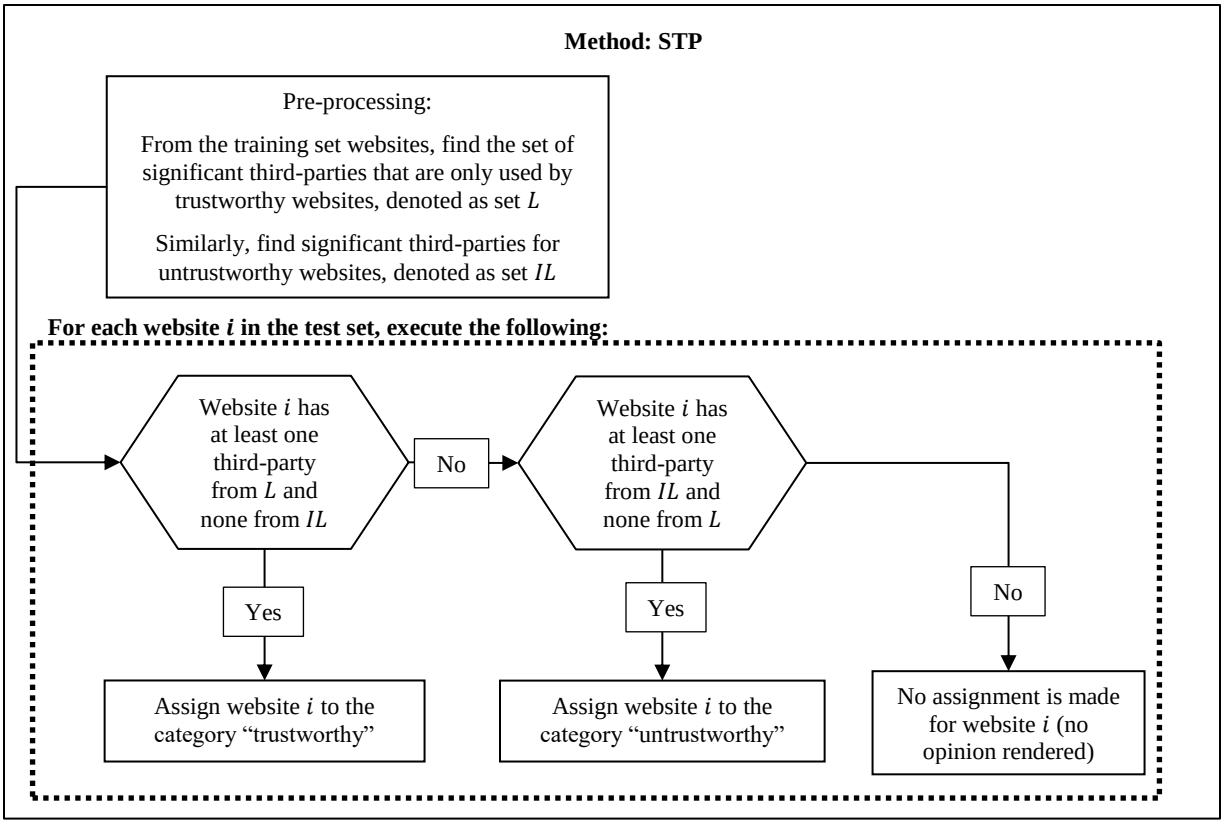

Figure 3. STP method flowchart

Pseudo-code for the STP method as shown in Figure 3 is presented below:

Begin STP
Input: test set I websites (to be categorized) and training set websites $J = \{J^{News}, J^{Clickbait}, J^{Fake}\}$
From websites in training set, find set of significant third-parties L only used by trustworthy websites, i.e. third-parties $l \in L$ *where* $\sum_{j \in J^{News}} T_l^j > 0$ *and* $\sum_{j \in J^{\{Clickbait,Fake\}}} T_l^j = 0$
From websites in training set, find the set of significant third-parties IL only used by untrustworthy websites, i.e. third-parties $l \in IL$ *where* $\sum_{j \in J^{News}} T_l^j = 0$ *and* $\sum_{j \in J^{\{Clickbait,Fake\}}} T_l^j > 0$
For *websites* $i \in I$ *(all websites in test set)*
  If *website i uses third-party(ies) from set L and none from set IL*
    Assign the test website i to category Trustworthy
  Else If *website i uses third-party(ies) from set IL and none from set L*
    Assign the test website i to category Untrustworthy
  End If
End For
End STP

3.2.3. SVM: We augment the previous heuristic methods with a linear Support Vector Machine (SVM) classification method. Linear SVM, a supervised learning method, builds a model from

the training data set by selecting the hyperplane that separates the classes with the highest accuracy and has the highest margin, where the margin is the distance of the nearest observation from each class to the hyperplane. By choosing the hyperplane with the highest margin among possible hyperplanes, the model increases the level of robustness. Before using the SVM, we perform additional data preparation using the k-means clustering technique. The use of k-means clustering enables us to reduce the dimensionality of the third-parties to *k*, while retaining the information needed for the SVM method to identify the websites as trustworthy or untrustworthy. We assign each third-party observed in the training data to one of the classes. We argue that the degree to which particular third-parties are utilized can be used to categorize websites. For example, particular third-parties may cater more to real news websites as opposed to fake news websites, and vice-versa. Figure 4 illustrates the k-means clustering, with k=4, of third-parties based on their ratios of real news, fake news, and clickbait websites. This figure clearly shows that there is a separation of third-party classes among the different website categories, based on the ratio of utilization by real news, fake news and clickbait websites. We conducted experiments with k=2, 3, 4, 5, 6 and 7, and find k=4 to have the highest prediction accuracy for categorization of websites based on third-party clustering.

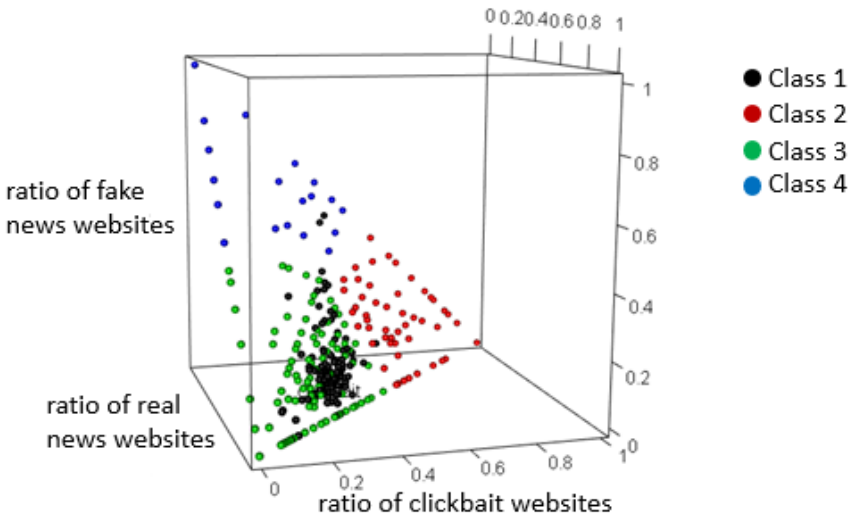

Figure 4. K-means clustering of third-parties

For the training data, we calculate the ratio of the third-parties that are seen in each cluster. To illustrate the relationship between the known website category and the 4 third-party clusters, consider Figure 5. We observe a separation between real news websites versus fake news and clickbait websites.

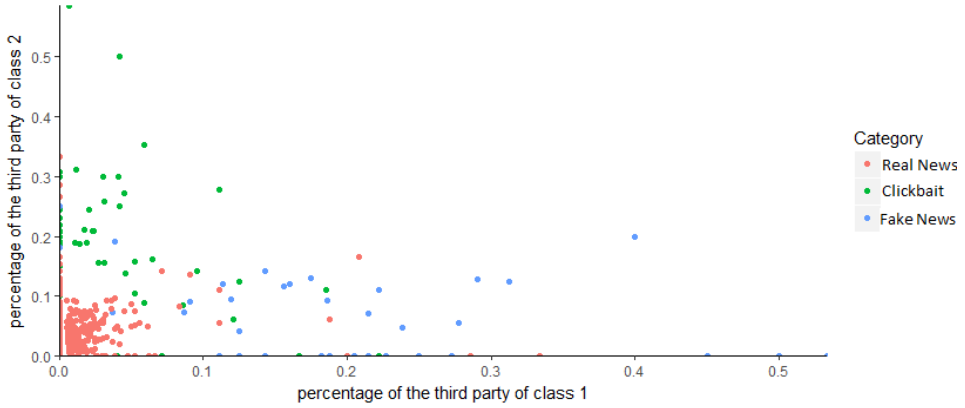

Figure 5.a. Illustration of websites based on third-party clusters 1 and 2

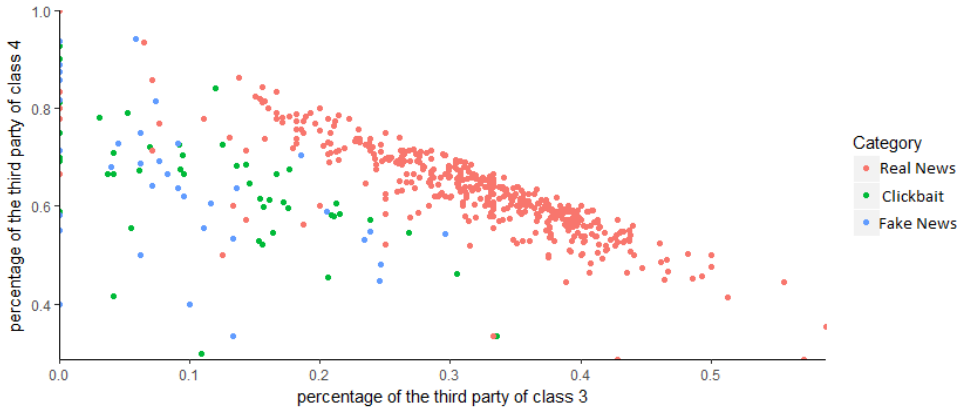

Figure 5.b. Illustration of websites based on third-party clusters 3 and 4

We then use the third-party cluster data from the *k*-means clustering as the input for the SVM model. Prediction output classifies a website as either trustworthy (real news) or untrustworthy (fake news and clickbait), and input values for each website are the percentages of the 4 third-party *k*-means clusters. As explained in Section 2, we use a 10-fold cross-validation to get the results.

3.3 Hybrid and Ensemble Classifiers

Other than the CS5, STP, and SVM methods, we also consider hybrid methods that are created by combining CS5 or SVM with STP. The STP-CS5 method categorizes websites first using the STP, and then uses CS5 to categorize the websites that are not classified using STP. STP-SVM works in a similar fashion. We use a threshold method to combine our proposed method with the existing methods from the literature to complement existing methods that work at the article-level using contents for identification. Using this threshold method, we first consider the prediction from one of our proposed methods. Due to our use of cross-validation, we have multiple predictions for a single website, which can be used to calculate a prediction confidence as the ratio of times the current prediction is made to the total number of predictions. If the

prediction confidence is higher than a specified threshold, we use the prediction from this proposed method. We find a threshold of 75% to be robust and give good results, though our findings are fairly consistent using other thresholds. If the prediction confidence is lower than 75%, we use the prediction from the existing method. Table 2 summarizes all methods used for classification.

Method		Acronym	Description
Existing	OpenMind	OM	From https://openmind.press
	ZenMate	ZM	From http://zenmate.com/
	Fakenewsai	FA	From http://Fakenewsai.com
Proposed	Significant Third Party	STP	First find the set of significant third parties. Then, within these significant third parties, find the set of third parties that are only used by websites in trustworthy training set websites, and the set of third parties that are only used by websites in untrustworthy training set websites. If the test website uses third parties from the trustworthy only set and none from the untrustworthy only set, then the website is categorized as trustworthy. In contrast, if the testwebsite uses third parties from the untrustworthy only set and none from the trustworthy only set, it is categorized as untrustworthy.
	Cosine Similarity Classifier	CS5	For each website in the test set, calculate its cosine similarity to websites in training set for each category. Take the top 5 websites with the highest similarity to the test website, and calculate the sum of the cosine similarities for training websites from each category to the test website. The category with the highest sum of cosine similarities in top 5 similar websites is chosen as the predicted category for the test website. For test websites that cannot be classified using this method, the test website is categorized as the category that has closest average number of third parties used in the training set to the test
	Support Vector Machine	SVM	Linear SVM, a supervised learning method, builds a model from the training data set by selecting the hyperplane that separates the classes with the highest accuracy and has the highest margin, where the margin is the distance of the nearest observation of each class to the hyperplane. By choosing the hyperplane with the highest margin among margins of possible hyperplanes, the model increases the level of robustness.
	STP-CS5		Use STP first. For websites that are not categorized using STP, use CS5.
	STP-SVM		Use STP first. For websites that are not categorized using STP, use TPF.
Ensemble	STP-SVM-OM		Use STP-SVM if its prediction confidence is higher than 75%, otherwise use OM.
	STP-SVM-ZM		Use STP-SVM if its prediction confidence is higher than 75%, otherwise use ZM.
	STP-SVM-FA		Use STP-SVM if its prediction confidence is higher than 75%, otherwise use FA.

Table 2. Summary of classification methods

4. Results

Table 3 shows the performance results for existing[12], proposed, and ensemble methods. Considering the classification terminology, in our classification, we take the "positive" outcome as untrustworthy, and the "negative" outcome as trustworthy. Other than true positive and true

[12] Here, we report the results only for the best-performing existing methods. The results for the remaining existing methods are provided in Appendix 1.

negative rates, we also provide the overall accuracy, the Matthews correlation coefficient, the F1 score, and Youden's J-statistic as overall performance measures for each method. Researchers often use these measurements to compare the performance of classification methods. The prediction rate indicates the ratio of the websites for which a given method was able to provide a prediction.

Method		True positive rate	True negative rate	Accuracy	F1 score	Matthews correlation coefficient	Youden's J-statistic	Prediction rate
Existing	OpenMind (OM)	0.635	1.000	0.943	0.777	0.771	0.635	0.998
	ZenMate (ZM)	0.884	0.956	0.945	0.835	0.804	0.840	1.000
	Fakenewsai (FA)	0.936	0.734	0.766	0.557	0.505	0.670	0.913
Proposed	STP	0.949	0.960	0.958	0.868	0.848	0.909	0.747
	CS5	0.977	0.820	0.845	0.667	0.632	0.797	1.000
	SVM	0.960	0.869	0.885	0.750	0.708	0.840	1.000
	STP-CS5	0.977	0.893	0.906	0.767	0.738	0.869	1.000
	STP-SVM	0.965	0.936	0.941	0.838	0.814	0.902	1.000
Ensemble	STP-SVM-OM (Threshold)	0.942	0.958	0.956	0.871	0.848	0.900	1.000
	STP-SVM-ZM (Threshold)	0.942	0.956	0.954	0.866	0.843	0.898	1.000
	STP-SVM-FA (Threshold)	0.976	0.947	0.952	0.865	0.844	0.924	0.996

Table 3. Performance of the existing, proposed, and ensemble methods

Considering these results, the best overall accuracy of our proposed methods is 0.958 for STP, with a true positive rate of 0.949 and a true negative rate of 0.960. This finding implies that STP is able to identify untrustworthy websites with high accuracy, while not mistaking trustworthy websites as untrustworthy. While the STP method is highly accurate, it is able to make a prediction for only 74.7% of the observations. The STP-SVM method builds on this and is able to make predictions for all websites with a true positive rate of 0.965, a true negative rate of 0.936, and an overall accuracy of 0.941. For the existing methods, ZenMate SafeSearch (ZM), Open Mind (OM), and Fakenewsai (FA) perform the best, but our proposed methods dominate them.

The ensemble methods drastically improve the performance over existing methods. Our proposed methods perform as well or better than ZM and FA with respect to all measurements. With OM, we improve on most measures, however, due to OM's bias towards predicting most websites as trustworthy, a tradeoff with true positive rate versus true negative rate exists. Looking at the Youden's J-statistic, a more balanced measure of accuracy in terms of the sizes of trustworthy and untrustworthy websites, the overall improvement is considerable.

4.1 Robustness Check

We conducted a number of robustness checks to assess the generalizability of our approach. We first analyze the impact of website popularity, and then alter our analysis using similarity measures other than the cosine similarity measure proposed in Section 3.

4.1.1. Website popularity: One possible limitation of our analysis may be that our prediction methods capture differences in popularity amongst news websites as opposed to trustworthiness, based on business model differences as we hypothesize. The popularity of websites, rather than their trustworthiness could be driving the differences in third-party sharing. If this were so, then our methods should be able to identify untrustworthy websites compared to only popular trustworthy websites. In order to address this issue, we perform additional analysis using only less popular real news websites within specific popularity rank ranges. We perform this analysis based on the rank order of the real news websites, focusing on the bottom half of the website list. For this analysis, starting with rank order 259, we test all 4 sets of smaller websites in increments of 50, performing a standard 10-fold cross-validation testing on 50 Real News, 50 Clickbait, and 50 Fake websites. The results are reported in Table 4.

Method		True positive rate	True negative rate	Accuracy	F1 score	Matthews correlation coefficient	Youden's J-statistic	Prediction rate
259-308	STP	0.944	0.952	0.947	0.958	0.889	0.897	0.380
	CS5	0.940	0.840	0.907	0.931	0.788	0.780	1.000
	SVM	0.920	0.880	0.907	0.929	0.792	0.800	1.000
	STP-CS5	0.940	0.840	0.907	0.931	0.788	0.780	1.000
	STP-SVM	0.930	0.960	0.940	0.954	0.871	0.890	1.000
309-358	STP	0.957	1.000	0.970	0.978	0.933	0.957	0.440
	CS5	0.950	0.880	0.927	0.945	0.834	0.830	1.000
	SVM	0.940	0.840	0.907	0.931	0.788	0.780	1.000
	STP-CS5	0.950	0.880	0.927	0.945	0.834	0.830	1.000
	STP-SVM	0.930	0.920	0.927	0.944	0.838	0.850	1.000
359-408	STP	0.979	0.929	0.967	0.979	0.907	0.907	0.407
	CS5	0.930	0.700	0.853	0.894	0.661	0.630	1.000
	SVM	0.920	0.860	0.900	0.925	0.776	0.780	1.000
	STP-CS5	0.930	0.700	0.853	0.894	0.661	0.630	1.000
	STP-SVM	0.940	0.900	0.927	0.945	0.836	0.840	1.000
409-458	STP	0.974	0.933	0.963	0.974	0.908	0.908	0.360
	CS5	0.940	0.840	0.907	0.931	0.788	0.780	1.000
	SVM	0.900	0.840	0.880	0.909	0.733	0.740	1.000
	STP-CS5	0.940	0.840	0.907	0.931	0.788	0.780	1.000
	STP-SVM	0.900	0.900	0.900	0.923	0.783	0.800	1.000

Table 4. Results using less-popular real news websites

We can see that the true positive rate is above 0.90 for all methods; regardless of the popularity of lower-ranked real news websites used for training, the identification of untrustworthy websites is strong. The true positive rate, true negative rate, and accuracy are all above 0.90 for all popularity ranges for the STP-SVM method with a prediction rate of 1. Thus, we can validate our method of using third-parties to identify website trustworthiness in using only less popular real news websites within specific relevant ranges. We offer this strong reinforcement of proof-of-concept, demonstrating that our proposed methodology identifies the trustworthiness of a website rather than its popularity.

Additionally, we analyze the global rank ordering for 513 of the websites in our sample[13] and partition them into sets of 50 across the classification of websites as untrustworthy or

[13] This analysis is done on the week ending September 27, 2019 and there is a survival bias compared to our original sample, with 65 untrustworthy and 448 trustworthy websites available for this measure.

trustworthy. In Table 5, we provide the true positive and negative rate of STP-SVM method with global rank-order distribution.

	GLOBAL RANK ORDER - STP-SVM MODEL			
	Untrustworthy websites	Trustworthy websites	True positive rate	True negative rate
1-50	2	48	1.000	1.000
51-100	1	49	1.000	0.918
101-150	3	47	1.000	0.894
151-200	5	45	0.800	0.956
201-250	N/A	50	N/A	0.940
251-300	4	46	1.000	0.935
301-350	6	44	0.833	0.886
351-400	3	47	0.667	0.957
401-450	3	47	1.000	0.957
451-500	27	23	1.000	0.870
501-513	11	2	1.000	0.500

Table 5. STP-SVM method results with global rank order distribution

From the results based on rank order, we see that not all trustworthy websites are highly ranked, and not all untrustworthy websites are ranked low. A good representation of untrustworthy websites exists in higher rankings, while trustworthy websites have a bigger presence throughout. The STP-SVM results are consistent and robust for most of the ranking ranges considering the true positive and the negative rate. Therefore, we can confidently reject the proposition that popularity is the only driver of differences in third-party usage by untrustworthy and trustworthy websites.

4.1.2. Similarity measures: In Section 3, we propose heuristics based on the cosine similarity (CS5 and STP-CS5). These methods assume that similarities in third-party usage among websites signal their underlying business models, which can be used for identification of website trustworthiness. In order to test the robustness of this approach, we propose methods using other

similarity measures available in the literature, including Jaccard index (*Jacc*), Sorensen-Dice coefficient (*SDC*), and inverse of Hamming distance (*Hamm*, we use inverse of the Hamming distance to get a measure of similarity rather than distance). In order to operationalize this, we change the similarity measure in the CS5 algorithm (Section 3.2.1) from S_{ij} for cosine similarity, to the following:

$$Jacc_{ij} = \frac{\sum_{r=1}^{N} T_r^i T_r^j}{\sum_{r=1}^{N} T_r^i + \sum_{r=1}^{N} T_r^j - \sum_{r=1}^{N} T_r^i T_r^j}$$

$$SDC_{ij} = \frac{2 \sum_{r=1}^{N} T_r^i T_r^j}{\left(\sum_{r=1}^{N} T_r^i\right)^2 + \left(\sum_{r=1}^{N} T_r^j\right)^2}$$

$$Hamm_{ij} = \frac{N}{\left(N - \sum_{r=1}^{N} T_r^i T_r^j\right)}$$

In all these measurements, the more common third-parties used between two websites, the higher the similarities among the two. Using these measurements, we recreate the results and compare them to the CS5 and STP-CS5 methods provided in Table 6.

Method	True positive rate	True negative rate	Accuracy	F1 score	Matthews correlation coefficient	Youden's J-statistic	Prediction rate
CS5	0.977	0.820	0.845	0.667	0.632	0.797	1.000
Jacc5	0.950	0.828	0.849	0.693	0.644	0.778	1.000
SDC5	0.960	0.734	0.774	0.604	0.545	0.694	1.000
Hamm5	0.550	0.989	0.910	0.688	0.667	0.539	1.000
STP-CS5	0.977	0.893	0.906	0.767	0.738	0.869	1.000
STP-Jacc5	0.950	0.895	0.905	0.782	0.743	0.845	1.000
STP-SDC5	0.960	0.860	0.878	0.738	0.696	0.820	1.000
STP-Hamm5	0.550	0.978	0.901	0.667	0.631	0.528	1.000

Table 6. Methods using alternative similarity measures

Results using Jacc5, SDC5, STP-Jacc5, and STP-SDC5 are consistent with results from CS5 and STP-CS5. We expect this result because the Jaccard index and Sorensen-Dice

coefficient behave similarly to CS5. However, we observe that Hamm5 and STP-Hamm5 have a classification bias towards trustworthy websites (high true negative rate, low true positive rate), implying that they do a poor job of identifying untrustworthy websites. The Hamming distance is not well-suited for use with large vectors such as ours, where a third-party vector can contain a few thousand items. Nevertheless, this method performs well with respect to overall accuracy and F1 score. Overall, we can see that our methodology does not rely heavily on use of a specific similarity measure, and that any construct that can provide a reasonable measure of third-party usage similarity, performs well.

5. Discussion

This paper presents a novel approach for identifying untrustworthy news websites, using only observable data derived from the back-end third-party partnerships or digital supply chains. This approach can reinforce the traditional methods for identification that focus primarily on the content of these websites. We find that there are significant similarities in websites' third-party usage within a category of websites, as well as dissimilarities in third-party usage across different categories. We can utilize these similarities and dissimilarities, driven by websites' business models, to design effective and efficient classifiers to identify website trustworthiness. We propose several basic, but effective, classification methods based on third-party usage. These methods predict the category of websites with high accuracy, and are at the same time simpler and more computationally efficient than existing content-based methods. Our approach can be easily implemented in practice and in real-time.

From a practical perspective, our approach complements existing content-based approaches and can be leveraged to achieve an overall superior performance. Existing approaches typically analyze the content of the news articles, whereas our approach performs on

another level, i.e. at the source level, examining the websites strictly on the third-parties they use. Content-based approaches can incorporate our analytical framework; for example, content can be scrutinized more closely, and a lower threshold can be set if it originates from a website that our approach classifies as untrustworthy. We show that integrating the two approaches improves existing methods. Fake news and clickbait websites may modify their content to overcome content-based detection approaches. However, we contend that such "cat and mouse" games are harder in our approach, as third-party relationships that drive the core business of these sites are more difficult to alter to avoid detection.

We additionally contribute to the classification methodology by identifying how to utilize k-means as a critical component of data preparation for the SVM website classifier to improve its performance. The effectiveness of the k-means clustering allows us to derive robust results from small datasets. Moreover, we propose several similarity-based methods for identification and show the effectiveness of these methods in our problem context. Additionally, aggregating the results of two heuristics allows us to create ensemble methods that outperform either of the methods alone.

In practice, our approach can provide real-time prediction of website trustworthiness. At least in two ways, this approach can be implemented. First, a browser extension can keep track of the third-parties that websites utilize in a database, and use this to predict the type of websites as users access them. Alternatively, a browser extension can visit the same website multiple times to get a better picture of the third-party usage by websites, and use this data to predict the website category. Our proposed approach outperform current content-based methods that require costly and time-consuming content monitoring in terms of ease of use and cost.

While we focus on the identification of untrustworthy news websites in this paper, our approach can be applied to other contexts. In our context, the third-party relationships capture the underlying business model of websites. The underlying framework of using digital supply chains for identification is impartial to the problem setting. Researchers can readily apply it to other classification problems where the supply chain structure is visible, given that the problem is sufficiently bounded, and the supply chain captures the underlying business model. For example, researchers can apply a similar method to identify untrustworthy e-commerce, health, or business websites.

This paper serves as a first attempt at utilizing information "beyond content" for detecting untrustworthy websites. Future researchers may address several limitations. First, we use a limited dataset to train the classifier model. Larger training datasets could help create better classification outcomes. Second, the methods proposed in this paper represent the first attempt at classification methods based on third-party usage. There is ample scope to develop more sophisticated models using additional attributes and more advanced analysis to achieve even better accuracies. Finally, we use only the third-party information that can be observed in the training data set, but third-party participants can also evolve over time. The proposed methods are designed to operate on websites that are completely unknown to the training set. Our methods are particularly adept with respect to this classic "relevant range" issue. A special case is where a website has no third-parties (we observed 6 such cases), and our method is unable to categorize these special case websites. Another special case is where we observe no third-party connections for a website contained in the testing set that are observed in the training set. Larger training sets and periodic retraining with updated training sets (to obtain new and updated third-party list) can combat this issue. We consider third-party digital supply chains to identify website

trustworthiness. Other potentially interesting underlying mechanisms exist that make this identification possible, and we have been indifferent to these mechanisms in our study. For example, the quality and type of third-parties may be correlated with the website business models. We believe that future research can delve into these nuances to provide further insights into the mechanisms by which these partnerships are developed.

References

Abbasi, A., Z. Zhang, D. Zimbra, H. Chen, J. F. Nunamaker Jr., Detecting fake websites: the contribution of statistical learning theory. *MIS Quarterly* (2010) 435-461.

Akpan, N. The very real consequences of fake news stories and why your brain can't ignore them (2016), (available at http://www.pbs.org/newshour/updates/real-consequences-fake-news-stories-brain-cant-ignore/).

Anand, A., T. Chakraborty, N. Park. We used neural networks to detect clickbaits: you won't believe what happened next! - *European Conference on Information Retrieval* (Cham, 2017), 541-547.

Bowersox, D., D. Closs, T. Stank, How to master cross-enterprise collaboration. *Supply Chain Management Review* ***7*** (4), 18–27 (2003).

Chakraborty, A., B. Paranjape, S. Kakarla, N. Ganguly. Stop clickbait: detecting and preventing clickbaits in online news media - *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference* (2016), 9-16.

Chandler, D., P.R. Haunschild, M. Rhee, C.M. Beckman, The effects of firm reputation and status on interorganizational network structure. *Strategic Organization, 11*(3), pp.217-244 (2013).

Chen, Y., N. J. Conroy, V. L. Rubin. Misleading online content: recognizing clickbait as false news - *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection* (2015), 15-19.

Choi, T. Y., Y. Kim, Structural embeddedness and supplier management: a network perspective. *Journal of Supply Chain Management* ***44***(4), 5-13 (2008).

Gopal, R., R. Patterson, E. Rolland and D. Zhdanov. "Social Network Meets Sherlock Holmes: Investigating the Missing Links of Fraud." *Computer Fraud and Security*, 2012(7), pp. 12-18. (2012)

Gopal, R., H. Hidaji, R.A. Patterson, E. Rolland, D. Zhdanov. How Much to share with third-parties? A website's dilemma and users' privacy concerns. *MIS Quarterly* ***42***(1), 143-164 (2018).

Granovetter, M. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* ***91***(3), 481-510 (1985).

Harvard library, Research guides: Fake news, misinformation, and propaganda (2017), (available at https://guides.library.harvard.edu/fake).

Hutchins, C., M. Benham-Hutchins. Hiding in plain sight: criminal network analysis. *Computational and Mathematical Organization Theory*, 16(1), pp.89-111 (2010)

Leetaru, K. Why stopping 'fake' news is so hard (2016), (available at https://www.forbes.com/sites/kalevleetaru/2016/11/30/why-stopping-fake-news-is-so-hard/).

Pogue, D. What Facebook is doing to combat fake news (2017), (available at www.scientificamerican.com/article/pogue-what-facebook-is-doing-to-combat-fake-news).

Polanyi, K., R. M. MacIver, *The great transformation* (Beacon Press, Boston, 1944), vol. 2, pp. 145.

Shellenbarger, S., Most students don't know when news is fake, Stanford study finds (2016), (available at www.wsj.com/articles/most-students-dont-know-when-news-is-fake-stanford-study-finds-1479752576).

Schumpeter, J. A., *Capitalism, socialism and democracy* (Routledge, London, 2010).

Šubelj, L., Š. Furlan, M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, 38(1), pp. 1039-1052 (2011)

Sullivan, B.N., P. Haunschild, and K. Page, Organizations non gratae? The impact of unethical corporate acts on interorganizational networks. *Organization Science* 18.1: 55-70 (2007).

Uzzi, B. D., "The dynamics of organizational networks: structural embeddedness and economic behavior," thesis, State University of New York at Stony Brook, Stony Brook, NY, (1993).

Uzzi, B. D., The source of consequences of embeddedness for the economic performance of organizations: the network effect. *American Sociological Review* ***61***(4), 674-698 (1996).

Uzzi, B. D.,Social structure and competition in interfirm networks: the paradox of embeddedness. *Administrative Science Quarterly* (1997) 35-67.

Xu, J., and H. Chen. CrimeNet explorer: a framework for criminal network knowledge discovery. *ACM Transactions on Information Systems*, 23(2), pp.201-226 (2005)

Zimdars, M. False, misleading, clickbait-y, and satirical “news” sources (2016), (available at https://docs.google.com/document/u/1/d/10eA5-mCZLSS4MQY5QGb5ewCbvb3VAL6pLkT53V_81ZyitM/mobilebasic).

Zuber, F. Spread of Unethical Behavior in Organizations: A Dynamic Social Network Perspective. *Journal of Business Ethics*, 131(1), pp. 151-172 (2015)

Appendix 1: Results for the Remaining Existing Methods

We provide the results for all existing methods in Table A1.

Method		True positive rate	True negative rate	Accuracy	F1 score	Matthews correlation coefficient	Youden's J-statistic	Prediction rate
Existing	OpenMind (OM)	63.5%	100.0%	94.3%	77.7%	77.1%	63.5%	99.8%
	ZenMate (ZM)	88.4%	95.6%	94.5%	83.5%	80.4%	84.0%	100.0%
	Fakenewsai (FA)	93.6%	73.4%	76.6%	55.7%	50.5%	67.0%	91.3%
	BS Detector	46.4%	99.1%	90.9%	61.4%	61.0%	45.5%	99.4%
	FNDF	11.6%	100.0%	86.0%	20.8%	31.6%	11.6%	100.0%
	FNDH	30.2%	93.9%	83.8%	37.1%	29.4%	24.1%	100.0%
	FN Blocker	12.8%	100.0%	86.2%	22.7%	33.1%	12.8%	100.0%
	FN Alert	29.4%	100.0%	88.9%	45.5%	51.0%	29.4%	99.8%

Table A1. Existing Methods results