Fake news detection in social media

Kelly Stahl *

B.S. Candidate, Department of Mathematics and Department of Computer Sciences, California State University Stanislaus, 1 University Circle, Turlock, CA 95382

Received 20 April, 2018; accepted 15 May 2018

Abstract

Due to the exponential growth of information online, it is becoming impossible to decipher the true from the false. Thus, this leads to the problem of fake news. This research considers previous and current methods for fake news detection in textual formats while detailing how and why fake news exists in the first place. This paper includes a discussion on Linguistic Cue and Network Analysis approaches, and proposes a three-part method using Naïve Bayes Classifier, Support Vector Machines, and Semantic Analysis as an accurate way to detect fake news on social media.

Keywords: fake news, false information, deception detection, social media, information manipulation, Network Analysis, Linguistic Cue, Fact-checking, Naïve Bayes Classifier, SVM, Semantic Analysis

Introduction

How much of what we read on social media and on supposedly "credible" news sites is trustworthy? It is extremely easy for anyone to post what they desire and although that can be acceptable, there is the notion of taking it a step too far, such as posting false information online in order to cause a panic, using lies to manipulate another person's decision, or essentially anything else that can have lasting repercussions. There is so much information online that it is becoming impossible to decipher the true from the false. Thus, this leads to the problem of fake news.

Literature review

What is fake news? Fake news is the deliberate spread of misinformation via traditional news media or via social media. False information spreads extraordinarily fast. This is demonstrated by the fact that, when one fake news site is taken down, another will promptly take its place. In addition, fake news can become indistinguishable from accurate reporting since it spreads so fast. People can download articles from sites, share the information, re-share from others and by the end of the day the false information has gone so far from its original site that it becomes indistinguishable from real news (Rubin, Chen, & Conroy, 2016).

Significance

Using social media as a medium for news updates is a double-edged sword. On one hand, social media provides for easy access, little to no cost, and the spread of information at an impressive rate (Shu, Sliva, Wang, Tang, & Liu, 2017). However, on the other hand, social media provides the ideal place for the creation and spread of fake news. Fake news can become extremely influential and has the ability to spread exceedingly fast. With the increase of people using social media, they are being exposed to new information and stories every day. Misinformation can be difficult to correct and may have lasting implications. For example, people can base their reasoning on what they are exposed to either intentionally or subconsciously, and if the information they are viewing is not accurate, then they are establishing their logic on lies. In addition, since false information is able to spread so fast, not only does it have the ability to harm people, but it can also be detrimental to huge corporations and even the stock market. For instance, in October of 2008, a journalist posted a false report that Steve Jobs had a heart attack. This report was posted through CNN's iReport.com, which is an unedited and unfiltered site, and immediately people retweeted the fake news report. There was much confusion and uncertainty because of how widespread it became in such a short amount of time. The stock of Job's company, Apple Inc., fluctuated dramatically that day due to one false news report that had been mistaken for authentic news reporting (Rubin, 2017).

* Corresponding author. Email: kstahl@csustan.edu

However, the biggest reason why false information is able to thrive continuously is that humans fall victim to Truth-Bias, Naïve Realism, and Confirmation Bias. When referring to people being naturally "truth-biased" this means that they have "the presumption of truth" in social interactions, and "the tendency to judge an interpersonal message as truthful, and this assumption is possibly revised only if something in the situation evokes suspicion" (Rubin, 2017). Basically humans are very poor lie detectors and lack the realization that there is the possibility they are being potentially lied to. Users of social media tend to be unaware that there are posts, tweets, articles or other written documents that have the sole purpose of shaping the beliefs of others in order to influence their decisions. Information manipulation is not a well-understood topic and generally not on anyone's mind, especially when fake news is being shared by a friend. Users tend to let their guard down on social media and potentially absorb all the false information as if it were the truth. This is also even more detrimental considering how young users tend to rely on social media to inform them of politics, important events, and breaking news (Rubin, 2017). For instance, "Sixty-two percent of U.S. adults get news on social media in 2016, while in 2012, only fort-nine percent reported seeing news on social media," which demonstrates how more and more people are becoming tech savvy and relying on social media to keep them updated (Shu et al., 2017). In addition, people tend to believe that their own views on life are the only ones that are correct and if others disagree then those people are labeled as "uniformed, irrational, or biased," otherwise known as Naïve Realism (Shu et al., 2017).

This leads to the problem of Confirmation Bias, which is the notion that people favor receiving information that only verifies their own current views. Consumers only want to hear what they believe and do not want to find any evidence against their views. For instance, someone could be a big believer of unrestricted gun control and may desire to use any information they come across in order to support and justify their beliefs further. Whether that is using random articles from uncredible sites, posts from friends, re-shared tweets, or anything online that does agrees with their principles. Consumers do not wish to find anything that contradicts what they believe because it is simply not how humans function. People cannot help but favor what they like to hear and have a predisposition for confirmation bias. It is only those who strive for certain academic standards that may be able to avoid or limit any biasness, but the average person who is unaware of false information to begin with will not be able to fight these unintentional urges.

In addition, not only does fake news negatively affect individuals, but it is also harmful to society in the long run. With all this false information floating around, fake news is capable of ruining the "balance of the news ecosystem" (Shu et al., 2017). For instance, in the 2016 Presidential Election, the "most popular fake news was even more widely spread on Facebook" instead of the "most popular authentic mainstream news" (Shu et al., 2017). This demonstrates how users may pay more attention to manipulated information than authentic facts. This is a problem not only because fake news "persuades consumers to accept biased or false beliefs" in order to communicate a manipulator's agenda and gain influence, but also fake news changes how consumers react to real news (Shu et al., 2017). People who engage in information manipulation desire to cause confusion so that a person's ability to decipher the true from the false is further impeded. This, along with influence, political agendas, and manipulation, is one of the many motives why fake news is generated.

Contributors of fake news

While many social media users are very much real, those who are malicious and out to spread lies may or may not be real people. There are three main types of fake news contributors: social bots, trolls, and cyborg users (Shu et al., 2017). Since the cost to create social media accounts is very low, the creation of malicious accounts is not discouraged. If a social media account is being controlled by a computer algorithm, then it is referred to as a social bot. A social bot can automatically generate content and even interact with social media users. Social bots may or may not always be harmful but it entirely depends on how they are programmed. If a social bot is designed with the sole purpose of causing harm, such as spreading fake news in social media, then they can be very malicious entities and contribute greatly to the creation of fake news. For example, "studies shows that social bots distorted the 2016 US presidential election discussions on a large scale, and around 19 million bot accounts tweeted in support of either Trump or Clinton in the week leading up to the election day," which demonstrates how influential social bots can be on social media (Shu et al., 2017).

However, fake humans are not the only contributors to the dissemination of false information; real humans are very much active in the domain of fake news. As implied, trolls are real humans who "aim to disrupt online communities" in hopes of provoking social media users into an emotional response (Shu et al., 2017). For instance, there has been evidence that claims "1,000 Russian trolls were paid to spread fake news on Hilary Clinton," which reveals how actual people are performing information manipulation in order to change the views of others (Shu et al., 2017). The main goal of trolling is to resurface any negative feelings harvested in social media users, such as fear and even anger, so that users will develop strong emotions of doubt and

distrust (Shu et al., 2017). When a user has doubt and distrust in their mind, they won't know what to believe and may start doubting the truth and believing the lies instead.

While contributors of fake news can be either real or fake, what happens when it's a blend of both? Cyborg users are a combination of "automated activities with human input" (Shu et al., 2017). The accounts are typically registered by real humans as a cover, but use programs to perform activities in social media. What makes cyborg users even more powerful is that they are able to switch the "functionalities between human and bot," which gives them a great opportunity to spread false information (Shu et al., 2017).

Now that we know some of the reasons why and how fake news progresses, it would be beneficial to discuss the methods of detecting online deception in word-based format, such as e-mails. The two main categories for detecting false information are the Linguistic Cue and Network Analysis approaches.

Linguistic cue methods

In Linguistic Cue approaches, researchers detect deception through the study of different communicative behaviors. Researchers believe that liars and truth-tellers have different ways of speaking. In text-based communication, deceivers tend to have a total word count greater than that of a truth-teller. Also, liars tend to use fewer self-oriented pronouns than other-oriented pronouns, along with using more sensory-based words. Hence, these properties found in the content of a message can serve as linguistic cues that can detect deception (Rubin, 2017). Essentially, Linguistic Cue approaches detect fake news by catching the information manipulators in the writing style of the news content. The main methods that have been implemented under the Linguistic Cue approaches are Data Representation, Deep Syntax, Semantic Analysis, and Sentiment Analysis.

When dealing with the Data Representation approach, each word is a single significant unit and the individual words are analyzed to reveal linguistic cues of deception, such as parts of speech or location-based words (Conroy, Rubin, & Chen, 2015).

The Deep Syntax method is implemented through Probability Context Free Grammars (PCFG). Basically, the sentences are being transformed to a set of rewritten rules in order to describe the syntax structure (Conroy, Rubin, & Chen, 2015).

Another approach, Semantic Analysis, determines the truthfulness of authors by characterizing the degree of compatibility of a personal experience. The assumption is that since the deceptive writer has no previous experience with the particular event or object, then they may end up including contradictions or maybe even leave out important facts that were existent in profiles on related topics (Conroy, Rubin, & Chen, 2015).

Finally, the last linguistic approach, Sentiment Analysis, focuses on opinion mining, which involves scrutinizing written texts for people's attitudes, sentiments, and evaluations with analytical techniques. However, this approach still is not perfect considering that the issues of credibility and verification are addressed with less priority (Rubin, 2017).

Network analysis methods

In contrast, Network Analysis approaches are content-based approaches that rely on deceptive language cues to predict deception. What makes this category different from the Linguistic approach is that the Network Analysis approach needs "an existing body of collective human knowledge to assess the truth of new statements" (Conroy, Rubin, & Chen, 2015). This is the most straightforward way of false information detection by checking the "truthfulness of major claims in a news articles" in order to determine "the news veracity" (Shu et al., 2017). This approach is fundamental for further progress and development of fact-checking methods. The underlying goal is using outside sources in order to fact-check any projected statements in news content by assigning a "truth value to a claim in a particular context" (Shu et al., 2017).

Moreover, the three existing fact-checking methods are expert-oriented, crowdsourcing-oriented, and computational-oriented. Expert-oriented fact checking is intellectually demanding and even time consuming since it is heavily based on human experts to analyze "relevant data and documents" which will lead to them composing their "verdicts of claim veracity" (Shu et al., 2017). A great example of expert-oriented fact checking is PolitiFact. Essentially PolitiFact requires their researchers to spend time analyzing certain claims by seeking out any credible information. When enough evidence has been gathered, a truth-value that ranges from True, Mostly True, Half True, Mostly False, False, and Pants on Fire is assigned to the original claim.

In addition, crowdsourcing-oriented fact checking uses the "wisdom of the crowd" concept which allows normal people, instead of only experts, to discuss and analyze the news content by using annotations which are then used to create an "overall assessment of the news veracity" (Shu et al., 2017). An example of this in action is Fiskkit, which is an online commenting website that aims to improve the dialogue of online articles by allowing its users to identify inaccurate facts or any negative behavior. This enables users to discuss and comment on the truthfulness of certain parts and sections of a news article (Shu et al., 2017).

Finally, the last type of fact-checking is Computational-oriented, which provides "an automatic scalable system to classify true and false claims" and tries to solve the two biggest problems: i). Identifying any "claims that are check-worthy" and ii). Determining the validity of these fact claims (Shu et al., 2017). Any statements in the content that reveal core statements and viewpoints are removed. These are identified as factual claims that need to be verified, hence enables the fact-checking process. Fact checking for specific claims requires external resources such as open web and knowledge graphs. Open web sources are used as "references that can be compared with given claims in terms of both consistency and frequency" (Shu et al., 2017). Knowledge graphs instead are "integrated from the linked open data as a structural network topology" which aspire to find out if the statements in the news content can be deduced from "existing facts in the knowledge graph" (Shu et al., 2017).

Moreover, the two main methods that are being used under the Network Analysis approach are Linked Data and Social Network behavior. In the Linked data approach, the false statements being analyzed can be extracted and examined alongside accurate statements known to the world (Conroy, Rubin, & Chen, 2015). When referring to accurate statements "known to the world" this relates to facts proven to be true and or statements that are widely accepted, such as "Earth is the name of the planet we live in."

Relating to the Social Network Behavior approach, this uses centering resonance analysis, which can be abbreviated as CRA, in order to represent "the content of large sets of text by identifying the most important words that link other words in the network" (Conroy, Rubin, & Chen, 2015). All the previous approaches discussed are the main methods of how researchers have been detecting fake news, however these practices have primarily been used for the textual formats, such as e-mails or conference call records (Rubin, 2017). The real question is how do predicative cues of deception in micro-blogs, such as Twitter and Facebook, differ from those of textual formats?

Therefore, concerning the area of false information in social media, fake news in the field of social media is relatively new. There have only been a handful of research studies completed in this domain, which requires more research to be conducted. In order to address this area, researchers are currently working on creating software that has the ability to detect deception. Deception detection software generally implements the different types of Linguistic cue approaches. However, when dealing with false information detection on social media, the problem is much more complex, using one method is no longer enough. Since linguistic cues are only one part of the problem, there are other aspects that essentially need to be incorporated such as positioning of the message sources in the network, reputation of cites, trustworthiness, credibility, expertise, and the tendency of spreading rumors should all be considered (Rubin, 2017).

Selected methods explored further

Furthermore, the methods to be further explored in relation to fake news detection in social media are Naïve Bayes classifier, SVM, and semantic analysis.

Naïve Bayes Classifier

Naïve Bayes is derived from Bayes Theorem, which is used for calculating conditional probability, the "probability that something will happen, given that something else has already occurred" (Saxena, 2017). Thus we are able to compute the likelihood of a certain outcome by using past knowledge of it.

Furthermore, Naïve Bayes is a type of classifier considered to be a supervised learning algorithm, which belongs to the Machine Language class and works by predicting "membership probabilities" for each individual class, for instance, the likelihood that the given evidence, or record, belongs to a certain class (Saxena, 2017). The class with the greatest, or highest probability, shall be determined the "most likely class," which is also known as Maximum A Posteriori (MAP) (Saxena, 2017).

Another way of thinking about Naïve Bayes classifier is that this method uses the "naïve" notion that all features are unrelated. In most cases, this assumption of independence is outrageously false. Suppose Naïve Bayes classifier is scanning an article and comes across "Barack," in many cases the same article will also have "Obama" contained in it. Even though these two features are clearly dependent, the method will still calculate the probabilities "as if they were independent," which does end up overestimating "the probability that an article belongs to a certain class" (Fan, 2017). Since Naïve Bayes classifier overestimates the probabilities of dependencies, it gives the impression that it would not work well for text classification. On the contrary, Naïve Bayes classifier still has a high performance rate even with "strong feature dependencies," since the dependencies will actually end up cancelling out each other for the most part (Fan, 2017).

In addition, what makes Naïve Bayes classifier desirable is that it's relatively fast and a highly accessible technique. It can be used for binary or multiclass classifications, making it an excellent choice for "Text Classification problems" as mentioned earlier (Saxena, 2017). Also, Naïve Bayes classifier is a straightforward algorithm that only really relies on performing many counts. Thus, it can be "easily trained on a small dataset" (Saxena, 2017).

However, the biggest downfall of this method is that it deems all the features to be separate, which may not always be the case. Hence, there is no relationship learned among the features (Saxena, 2017).

SVM

A support vector machine (SVM), which can be used interchangeably with a support vector network (SVN), is also considered to be a supervised learning algorithm. SVMs work by being trained with specific data already organized into two different categories. Hence, the model is constructed after it has already been trained.

Furthermore, the goal of the SVM method is to distinguish which category any new data falls under, in addition, it must also maximize the margin between the two classes (Brambrick). The optimal goal is that the SVM will find a hyperplane that divides the dataset into two groups.

To elaborate further, the support vectors are "the data points nearest to the hyperplane" and if removed would modify the location of the dividing hyperplane (Brambrick). Thus, support vectors are crucial elements of a data set. In addition, the hyperplane can be thought of as "a line that linearly separates and classifies a set of data" and "the further from the hyperplane our data points lie," the higher the chance that the data points have been accurately classified (Brambrick).

Moreover, the advantages of using the SVM method are that it tends to be very accurate and performs extremely well on datasets that are smaller and more concise. In addition, this technique is very flexible since it can be used to classify or even determine numbers. Also, support vector machines have the capability to handle high dimensional spaces and tend to be memory efficient (Ray, Srivastava, Dar, & Shaikh, 2017).

On the contrary, the disadvantages of using the SVM approach are that it has difficulty with large datasets since "the training time with SVMs can be high" and it is "less effective on noisier [meaningless] datasets with overlapping classes" (Brambrick). In addition, the SVM method will not "directly provide probability estimates" (Ray et al., 2017).

Semantic Analysis

Semantic analysis is derived from the natural language processing (NLP) branch in computer science. As discussed earlier, the method of semantic analysis examines indicators of truthfulness by defining the "degree of compatibility between a personal experience," as equated to a "content 'profile' derived from a collection analogous data" (Conroy, Rubin, & Chen, 2015). The idea is that the fake news author is not familiar with the specific event or object. For example, they have never even visited the location in question, thus they may neglect facts that have been present in "profiles on similar topics" or potentially include ambiguities that semantic analysis can detect (Conroy, Rubin, & Chen, 2015).

Furthermore, a huge reason for using semantic analysis is that this method is able to precisely classify a document through the use of association and collocation (Unknown, 2013). This is especially useful for languages that have words with multiple meanings and close synonyms, such as the English language. Suppose if one decided to use a simple algorithm that is unable to make the distinction among different word meanings, then the result may be ambiguous and inaccurate. Thus, by considering rules and relations when searching through texts, semantic analysis operates similarly to how the human brain functions (Unknown, 2013).

However, in light of the situation of comparing profiles and the "description of the writer's personal experience" discussed above, there are potentially two limitations with the semantic analysis method (Conroy, Rubin, & Chen, 2015). In order to even "determine alignment between attributes and descriptors," there needs to be a great amount of excavated content for profiles in the first place (Conroy, Rubin, & Chen, 2015). In addition, there also exits the challenge of being able to accurately associate "descriptors with extracted attributes" (Conroy, Rubin, & Chen, 2015).

Proposed method

Due to the complexity of fake news detection in social media, it is evident that a feasible method must contain several aspects to accurately tackle the issue. This is why the proposed method is a combination of Naïve Bayes classifier, Support Vector Machines, and semantic analysis. The proposed method is entirely composed of Artificial Intelligence approaches, which is critical to accurately classify between the real and the fake, instead of using algorithms that are unable to mimic cognitive functions. The three-part method is a combination between Machine Learning algorithms that subdivide into supervised learning techniques, and natural language processing methods. Although each of these approaches can be solely used to classify and detect fake news, in order to increase the accuracy and be applicable to the social media domain, they have been combined into an integrated algorithm as a method for fake news detection.

Furthermore, SVM and Naïve Bayes classifier tend to "rival" each other due to the fact they are both supervised learning algorithms that are efficient at classifying data. Both techniques are moderately accurate at categorizing fake news in experiments, which is why this proposed method focuses on combining SVM and Naïve Bayes classifier to get even more accurate results. In "Combining Naive Bayesian and Support Vector Machine for Intrusion Detection

System," the authors integrate both methods of SVM and Naïve Bayes classifier in order to create a more precise method that classifies better than each method individually. They found that their "hybrid algorithm" effectively minimized "false positives as well as maximize balance detection rates," and performed slightly better than SVM and Naïve Bayes classifier did individually (Sagale, & Kale, 2014). Even though this experiment was applied to Intrusion Detection Systems (IDS), it clearly demonstrates that merging the two methods would be relevant to fake news detection.

Moreover, introducing semantic analysis to SVM and Naïve Bayes classifier can improve the algorithm even more. The biggest drawback of Naïve Bayes classifier is that it deems all features of a document, or whichever textual format being used, to be independent even though most of the time that is not the situation. This is a problem due to lowered accuracy and the fact that relationships are not being learned if everything is assumed to be unrelated. As we mentioned earlier, one of the biggest advantages of semantic analysis is that this method is able to find relationships among words. Thus, adding semantic analysis helps fix one of the biggest weaknesses of Naïve Bayes classifier.

In addition, adding semantic analysis to SVM can improve the performance of the classifier. In "Support Vector Machines for Text Categorization Based on

Latent Semantic Indexing," the author shows that combining the two methods improves the efficiency due to "focusing attention of Support Vector Machines onto informative subspaces of the feature spaces," (Huang, 2001). In the experiment, semantic analysis was able to capture the "underlying content of document in semantic sense," (Huang, 2001). This improved the efficiency of SVM since the method would waste less of its time classifying meaningless data and spend more time organizing relevant data with the help of semantic analysis. As outlined earlier, a huge benefit of semantic analysis is its ability to extract important data through relationships between words; hence, semantic analysis is able to use its fundamental benefit to further improve SVM.

Conclusion

As mentioned earlier, the concept of deception detection in social media is particularly new and there is ongoing research in hopes that scholars can find more accurate ways to detect false information in this booming, fake-news-infested domain. For this reason, this research may be used to help other researchers discover which combination of methods should be used in order to accurately detect fake news in social media. The proposed method described in this paper is an idea for a more accurate fake news detection algorithm. In the future, I wish to test out the proposed method of Naïve Bayes classifier, SVM, and semantic analysis, but, due to limited knowledge and time, this will be a project for the future.

It is important that we have some mechanism for detecting fake news, or at the very least, an awareness that not everything we read on social media may be true, so we always need to be thinking critically. This way we can help people make more informed decisions and they will not be fooled into thinking what others want to manipulate them into believing.

References

Brambrick, Aylien, N. (n.d.). KDnuggets. Retrieved February 20, 2018, from https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

Chen, Y., Conroy, N., & Rubin, V. (2015). News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4.

Conroy, N., Rubin, V., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4.

Fan, C. (2017). Classifying fake news. Retrieved February 18, 2018, from http://www.conniefan.com/2017/03/classifying-fake-news

Huang, Y. (2001). Support Vector Machines for Text Categorization Based on Latent Semantic Indexing.

Ray, S., Srivastava, T., Dar, P., & Shaikh, F. (2017). Understanding Support Vector Machine algorithm from examples (along with code). Retrieved March 2, 2018, from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Rubin, V., Chen, Y., & Conroy, N. (2015). Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4.

Rubin, V., Chen, Y., & Conroy, N. J. (2016). Education and Automation: Tools for navigating a sea of fake news. *UNDARK*.

Rubin, V., Conroy, N. J., & Chen, Y. (2015, January). *Research Gate*. Retrieved April 11, 2017, from https://www.researchgate.net/publication/270571080_Towards_News_Verification_Deception_Detection_Methods_for_News_Discourse doi:10.13140/2.1.4822.8166

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. doi:10.18653/v1/w16-0802

Rubin, V. (2017). Deception detection and rumor debunking for social media. *Handbook of Social Media Research Methods*.

Sagale, A. D., & Kale, S. G. (2014). Combining Naive Bayesian and Support Vector Machine for Intrusion Detection System. *International Journal of Computing and Technology*, *1*(3). Retrieved April 1, 2018.

Saxena, R. (2017). How the Naive Bayes Classifier works in Machine Learning. Retrieved October 20, 2017, from https://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22-36.

Unknown. (2013). Why Semantics is Important for Classification. Retrieved March 19, 2018, from http://www.skilja.de/2013/why-semantics-is-important-for-classification/