Fake News Detection in Social Media

Kelly Stahl

California State University Stanislaus

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 on 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 details how methods of Naive Bayes Classifier can be used to detect fake news. The results show that Bernoulli and Multinomial Naive Bayes methods are not sufficient enough to properly detect fake news on social media, due to the simplicity of the methods and that fake news detection on social media is a very complex problem.

*Keywords:* fake news, false information, deception detection, social media, information manipulation, Network Analysis, Linguistic Cue, misinformation, Data Representation, Bayes Theorem, Naive Bayes Classifier, Bernoulli Naive Bayes, Multinomial Naive Bayes, Deep Syntax, Semantic Analysis, Sentiment Analysis, Fact-checking, Linked Data, Social Network Behavior, Bias

Fake News in Social Media

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 when one fake news site is taken down, another will promptly take its place. In addition, fake news can become indistinguishable 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 contrary, 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 everyday. 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., ended up dramatically fluctuating that day due to one false news report that had been mistaken for authentic news reporting (Rubin, 2017).

 However, the biggest reason why false information is able to thrive continuously is that humans fall victim to Truth-Bias, Naive 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 any other written documents that have the sole purpose to shape 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, “62 percent of U.S. adults get news on social media in 2016, while in 2012, only 49 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. People cannot help but favor what they like to hear and have a predisposition for confirmation bias.

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, Sentiment Analysis, and Naive Bayes Classifier.

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).

In addition, 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).

Finally, the last linguistic approach is Naive Bayes Classifier, which will be discussed in greater detail later in this research.

**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, in order to demonstrate how some of these approaches work, the methods chosen as an example to detect fake news are of the Naive Bayes Classifier**.** The two approaches being explored are Bernoulli Naive Bayes and Multinomial Naive Bayes. First, it will be beneficial to discuss the fundamentals, such as what is Bayes Theorem.

**Bayes Theorem**

Bayes Theorem is used for calculating conditional probability, which is 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 its past knowledge. The formula for calculating the conditional probability is as follows:

P(H|E) = P(E|H) \* (PH) / P(E)

Where P(H) is the probability of the hypothesis H being correct, or also known as the prior probability. P(E) is the likelihood of the evidence being true, excluding the prior probability. P(E|H) is the probability of the evidence, including the prior probability, being true. Finally, P(H|E) is the likelihood of the prior probability, including the evidence, is true (Saxena, 2017). Now that Bayes Theorem has been discussed, let’s move into the explanation of how the Naive Bayes Classifier works.

**Naive Bayes Classifier**

 Furthermore, the Naive Bayes is a type of classifier that incorporates Bayes Theorem. This method 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). For example, consider a dataset with three hundred records and three distinct classes. The three classes are all related to animal types. The three possible animal types, in this example, are cat, owl, or snake. The predictor features set contains four different features that are fly, swim, blue color, or sharp claws. Located below is a frequency table of the data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Fly* | *Swim* | *Sharp Claws* | *Blue Color* | *Animal Type* |
| **0%** (0/100) | **40%** (40/100) | **100%** (100/100) | **0%** (0/100) | Cat |
| **100%** (100/100) | **10%** (10/100) | **100%** (100/100) | **0%** (0/100) | Owl |
| **0%** (0/100) | **100%** (100/100) | **0%**(0/100) | **60%** (60/100) | Snake |

*Figure 1*. Frequency Table of Data. Adapted from *How The Naive Bayes Classifier Works In Machine Learning,*by R. Saxena, 2017,RetrievedOctober 20, 2017, from https://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/

Consider one record, shown below, that has feature values specified but the animal type needs to be predicted. We want to predict the animal type, or class, by using the feature values known.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Fly* | *Swim* | *Sharp Claws* | *Blue Color* | *Animal Type* |
| **False** | **True** | **True** | **False** | ? |

*Figure 2*. Record of Known Data. Adapted from *How The Naive Bayes Classifier Works In Machine Learning,*by R. Saxena, 2017,RetrievedOctober 20, 2017, from https://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/

We need to use the Naive Bayes Approach in order to predict the animal type, which is stated as:

P(H| Multiple Evidences) = P(E1|H)\*P(E2|H)…\*P(En|H)\*P(H)/P(Multiple evidence)

The evidence, in this case, is swim and sharp claws, since these are the values that are known. Now, we must calculate the probabilities for each animal type and see which animal type has the largest probability. The one with the greatest probability will be the predicted animal type for the record (Saxena, 2017).

For Cat, the equation is as follows:

P(Cat| Swim, Sharp Claws) = P(Swim| Cat) \*P(Sharp Claws| Cat)\* P(Cat)/ P(Swim, Sharp Claws)

= .4 **\*** 1\* 0.333/P(Swim, Sharp Claws)

**= 0.1332/P(Swim, Sharp Claws)**

For Owl, the equation is as follows:

P(Owl| Swim, sharp claws) = P(Swim| Owl) \*P(Sharp Claws| Owl)\* P(Owl)/ P(Swim, Sharp Claws)

= .1\*1\*0.333/P(Swim, Sharp Claws)

**=.0333/P(Swim, Sharp Claws)**

For Snake, the equation is as follows:

P(Snake| Swim, Sharp claws) = P(Swim| Snake) \*P(Sharp Claws| Snake)\* P(Snake)/ P(Swim, Sharp Claws)

= 1\*0\* 0.333/ P(Swim, Sharp Claws)

**=0**

Therefore, the value of P(Cat| Swim, Sharp Claws) is the largest of the three classes, hence the Naive Bayes method predicts the record is a cat.

**Types of Naive Bayes Algorithms**

Furthermore, the two types of Naive Bayes Algorithms that will be examined and tested are Bernoulli Naive Bayes and Multinomial Naive Bayes. The Bernoulli Naive Bayes method is “used on data that is distributed according to multivariate Bernoulli distributions” for instance it needs features to “be binary valued” (Saxena, 2017).

In contrast, the Multinomial Naive Bayes method should be used on data that is “multinomially distributed” (Saxena, 2017). A multinomial distribution allows the outcomes of the experiment to be more than two. A good way to think about the two methods is that a multinomial distribution would be able to express the “results of tossing two dice” since each die can have up to six different outcomes. Instead, the binomial distribution would show the “results of a coin toss” since “there are only two possible results” each toss could have, heads or tails (Staff, 2010).

In addition, what makes these methods desirable is that they are relatively fast and highly accessible techniques. Depending on which one a person chooses, they can be used for binary or multiclass classifications, making them an excellent choice for “Text Classification problems” (Saxena, 2017). Also, Bernoulli Naive Bayes and Multinomial Naive Bayes are straightforward algorithms that only really rely on performing many counts. Thus, they can be “easily trained on a small dataset” (Saxena, 2017).

However, the biggest downfall of these methods is that they deem all the features to be separate. Hence, there is no relationship learned among the features (Saxena, 2017).

**Bernoulli and Multinomial Naive Bayes Experiment**

Furthermore, this section will discuss how Bernoulli and Multinomial Naive Bayes algorithms were put to the test and used to detect fake news in Svärd and Rumman’s “Combating Disinformation” experiment. In their study, they used 201 distinct mainly American news articles that were gathered and labeled appropriately. 120 articles were identified as fake news and the other 81 were tagged as real news articles. The fake news articles were retrieved from an “online corpus” that was constructed by “data gathered from around 244 websites” (Svärd & Rumman, 2017). These websites were “tagged by the Google Chrome-extension ‘BS detector’” and the real news articles were retrieved “by hand” (Svärd & Rumman, 2017). These were determined true if the articles did not contain any facts that contradicted another article generated by a dependable source that had “reported on that same news story” (Svärd & Rumman, 2017).

Moreover, a “test sample with known labels is compared to the models prediction” (Svärd & Rumman, 2017). For the first test, the authors used 10 fake and 10 real articles, which were then stored into one array and the matching classifications into another. They kept the order unharmed by having the real news articles be the first 10 and the fake news articles be the last 10 in order to make the coding more simplistic (Svärd & Rumman, 2017). If the prediction of the method was true and the actual value of article was true, then it was labeled as True Positive. In contrast, if the prediction of the method was False and the actual value of article was false, then it was labeled as True Negative. Furthermore, if the method predicted an article to be true but it was actually false then it was considered False Positive. Finally, if the method predicted an article to be false but it was actually true then it was considered False Negative (Svärd & Rumman, 2017). The results of the first test are displayed below along with the evaluation:

*Figure 3*. Smaller Test. Retrieved from *COMBATING DISINFORMATION,*by M. Svärd, & P. Rumman, (2017)

*Figure 4*. Smaller Test Evaluation Scores. Retrieved from *COMBATING DISINFORMATION,*by M. Svärd, & P. Rumman, (2017)

Likewise, a second test was conducted using a larger sample size. The only difference was that instead of 10 fake news articles, there were now 100 fake news articles. The results of the second test are displayed below along with the evaluation:

*****Figure 5*. Bigger Test. Retrieved from *COMBATING DISINFORMATION,*by M. Svärd, & P. Rumman, (2017)

****

*Figure 6*. Bigger Test Evaluation Scores. Retrieved from *COMBATING DISINFORMATION,*by M. Svärd, & P. Rumman, (2017)

**Experiment Conclusion**

In conclusion, the two methods were mediocre at detecting fake news and “fairly accurate within specific areas,” pertaining to the true positive and true negative categories (Svärd & Rumman, 2017). The Bernoulli Naive Bayes method seems to have decent accuracy rates in both tests. However, the Multinomial Naive Bayes method performs much worse in the larger test than in the smaller test. Its accuracy goes from 60% to 44.54%, which is unacceptable for fake news detection.

Overall, the results were not great which is not shocking since these algorithms are so simple. This was a good attempt at detecting fake news but would not be instrumental enough to use alone for fake news detection, not to mention fake news detection in social media.

**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 accurate ways to detect false information in this booming fake news infested domain. Researchers are unsure of which methods or combination of methods that should be used in the field of social media. This paper analyzes how Naive Bayes Classifier methods can be used to detect fake news. However, these methods are not a good consideration for fake news detection in social media and should not be used alone since there are many areas of social media that these approaches do not cover.

It is important that we have some mechanism for detecting fake news or at the very least awareness that everything we read on social media may or may not be true and 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:**

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.

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.

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., 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., 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*.

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.

Staff, I. (2010). Multinomial Distribution. Retrieved October 20, 2017, from https://www.investopedia.com/terms/m/multinomial-distribution.asp

Svärd, M., & Rumman, P. (2017). COMBATING DISINFORMATION: Detecting fake news with linguistic models and classification algorithms (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-209755