Unmasking Fake News: An Investigation into the Emotional Underpinnings of Deceptive Information Using Naive Bayes Classification

Anjali Kapoor (akkapoor@princeton.edu)
Department of Computer Science, Princeton University

Abstract

Fake news can have extremely harmful effects on society, such as increasing polarization, instilling distrust in the media and government, inciting violence, and encouraging dangerous health practices (Fleming, n.d.). Psychologists believe that fake news is effective because it incites an emotional response that clouds people's judgement. My hypothesis is that the frequency of emotion-related words may therefore be a way to detect fake news. To test my hypothesis, I trained a Multinomial Naïve Bayes Classifier on a dataset consisting 17,455 fake news articles and 21,192 true news articles converted into a bag-of-words representation, achieving a classification accuracy of 94%. Extracting the top features used for fake versus true classification and applying a sentiment analysis model to produce sentiment scores of the top 100 words associated with each class demonstrated that on average, words associated with fake news expressed 10x more negative sentiment than the words associated with true news. Overall, this model is effective at fake news detection, and the computational results of this study are in line with psychologists' beliefs.

Keywords: Fake News, Emotion, Machine Learning, Naïve Bayes Classification, Sentiment Analysis, Natural Language Processing

Introduction

Fake news is any news that is "inaccurate, biased, misleading, or fabricated" (Grieve & Woodfield, 2023). Fake news has been around since the origination of real news (McIntyre), and can have effects such as increasing polarization, instilling distrust in the media and government, inciting violence, and encouraging dangerous health practices (Fleming, n.d.). Fake news is also extremely pervasive - one study conservatively estimated that the average American encountered between one and three fake news stories during the month before the 2016 election, and another estimated that there are as many as 60 million bots on Facebook (Lazer et al., 2018).

Since fake news is often disseminated by news platforms themselves, it's difficult to trust any source and therefore discern what is true versus fake (Grieve & Woodfield, 2023). The content and style of the articles can be very similar, making it hard to detect fake news; furthermore, fake news is often created intentionally to manipulate people by inciting an emotional response (Horner et al., 2021). Thus, it would be extremely valuable to create a machine learning model to identify fake news, that would not be influenced by manipulative tactics and can recognize patterns that are difficult for a human to detect.

Interestingly, what makes fake news so effective may also be the key to identifying fake news. Fake news is able to evade detection and spread 6x quicker (Vosoughi et al., 2018) as it incites an emotional response that clouds people's judgement. One of the ways it can produce this response is by relying on more negative sentiment and themes of morality (Carrasco-Farré, 2022). While individual words don't necessarily capture the complexity of negative sentiment and themes of morality, there are words associated with emotion such as adjectives, intensifiers, and words related to sex, death, and anxiety (Shariatmadari, 2019). Therefore, I propose using a Multinomial Naïve Bayes Classifier to first, determine whether a model that uses the presence and frequency of words results in accurate predictions, and second, whether the words that it uses to classify fake news are related to emotion.

If the model can classify fake and true news with high accuracy, then it would be very useful for people to determine the authenticity of news, and hopefully reduce the negative impacts of fake news. If the top words, aka the top features used to classify fake news, are related to emotion, this would support the psychologist's theory that fake news manipulates people's emotions more than true news and may further our understanding of the virality of fake news.

Background

I will first dive into cognitive theories on how emotion affects the ability to discern fake news, then explain how a naïve bayes classifier can be applicable to assessing the prevalence of emotional words in fake news, and finally I'll explore machine learning research that has been conducted so far on this topic.

The Effect of Emotion On Detecting Fake News

Psychologists have performed numerous studies that demonstrate that emotion has a large effect on people's judgement, and as a result, inciting an emotional response can hinder people from detecting fake news and encourage people to share fake news. One study, for example, found that participants who felt any emotion (other than anger) from reading a headline were more likely to believe a false headline, than those who felt no emotion (Bago et al., 2022). Similarly, another study found that for nearly every emotion evaluated by the PANAS scale, increased emotionality is associated with increased belief in fake news (Martel et al., 2020), aside from the emotions "interested," "alert," "determined," and "attentive". It's possible that this is because emotions have been found to distract readers from

diagnostic cues such as source credibility (Ecker et al., 2022). Another study took a slightly different angle, studying people's likelihood to share an article rather than focusing on whether they believe an article, and found that negatively biased fake news enhances people's willingness to share the article, while positively biased fake news has no significant effect on virality; furthermore, they found that the potential for virality is mediated by negative emotions, such as anger and fear, but not by positive ones (Corbu et al., 2021). While the findings of these papers all differ slightly in their findings, particularly on the effects of a news article inducing anger, they all have found through studies on human participants that producing some type of emotion increases the likelihood of someone believing and spreading fake news.

Naïve Bayes Classification

A Naïve Bayes Classifier can be utilized to determine whether fake news contains more emotion-related words compared to real news. A Naïve Bayes Classifier is a probabilistic machine learning model based on Bayes' theorem that relies on the assumption that the features are conditionally independent, or in other words the presence of one feature is unrelated to the presence of another feature; in this case, it considers the frequency of each word in an article independently from the frequencies of other words in that article in its classification calculations (Ray, 2017).

To explain more about the math behind the model, Bayes' Theorem calculates the probability of an event based on prior knowledge, using the formula $P(C|X) = \big(P(X|C) * P(C)\big)/P(X)$, where $P(C|X)$ is the probability of class C given observation X, $P(X|C)$ is the probability of observation X given class C, $P(C)$ is the prior probability of class C, and $P(X)$ is the prior probability of observation X. In a Naïve Bayes Classifier, the model assumes that each of the features are conditionally independent, or $P(X|C) = P(x_1|C) * P(x_2|C) * ... * P(x_n|C)$ (Ray, 2017).

To put this formula in context with this situation, to figure out the probability of an article being true or fake, $P(C|X)$, it figures out the probability of each word in the document being associated with the classification of true or false and multiplies these probabilities together to calculate $P(X|C)$. Then it calculates $P(C)$ based on the number of true and fake articles in the training set, and $P(X)$ as $1/|\text{X\_train}|$.

Machine Learning to Identify Fake News

Many researchers have utilized machine learning to identify fake news and analyze emotion in fake news. For example, one team compared the accuracy of three different models at detecting fake news on Twitter - a Naïve Bayes Classifier, a Neural Network, and a Support Vector Machine (SVM) - finding 96.08%, 99%, and 99% accuracies respectively (Aphiwongsophon & Chongstitvatana, 2018). Another study compared the effectiveness of Logistic Regression, SVM, Random Forest (RF), Naïve Bayes, Gradient Boosting, and Passive Aggression, also considering the influence of term frequency-inverse document frequency (TF-IDF) and bag-of-words (BoW), on classifying fake news in Urdu, finding that RF with BoW features performed best with an accuracy of 92% (Rafique et al., 2022). Overall, there is a breadth of research on utilizing machine learning to classify fake news and the efficacy of many types of models have been compared. I chose to use a Naïve Bayes Classifier for my purposes as it has proven to have high classification accuracy for past researchers and would allow me to analyze the sentiment of top features.

In addition to using various machine learning models to detect fake news, researchers have also employed machine learning/AI methods to analyzing the use of emotion in fake news. One study utilized an AI algorithm to compare 150 real and fake news articles and found that fake news titles are substantially more negative than real news titles, and that the text of fake news displays more negative emotions such as disgust and anger and less positive emotion such as joy (Paschen, 2019). Similarly, another study found that positive emotion in a text lowers the likelihood of that text being fake (Nanath et al., 2022). These studies demonstrate that fake news does not just incite emotion among its readers, it also contains more emotion.

Many researchers have also employed machine learning to simultaneously detect emotion and veracity and have found that this combined approach is more accurate at fake news classification than a single task model. One study found that using a multi-task model trained to predict both emotion and legitimacy of a text performs better than a single-task model that only predicts legitimacy in cross-domain settings, thus demonstrating the value of using emotion as a feature in a model (Choudhry et al., 2022). Another study found similar results, a Bidirectional Encoder Representation (BERT) machine learning model performed better at fake news classification with emotional features (Mackey et al., 2021). Though these papers take a different approach than I do, building emotion detection into the model as a feature, rather than creating the model and then evaluating the features it uses, they still come to the same conclusion that there is an imbalance in emotion between real and fake news that can be exploited for fake news detection.

Approach

My approach of using a Naïve Bayes Classifier has already been proven to be highly accurate at fake news classification in numerous past studies (Aphiwongsophon & Chongstitvatana, 2018; Rafique et al., 2022). However, I propose going one step further than these studies, to perform sentiment analysis on the model's top features associated with fake and true classification. Thus, my goal isn't solely to create an accurate model at detecting fake news, but also to determine *why* it is accurate. In other words, I plan to investigate which words the model identifies to be associated with fake versus real news, and whether the words associated with fake news are more related to emotion, and therefore have a more extreme (either positive or negative) sentiment score than the words associated with real news.

Methods

Dataset

I used the ISOT Fake News dataset (*Fake News Detection Datasets*, n.d.), which contains 23,481 fake news articles collected from unreliable websites flagged by PolitiFact and Wikipedia, and 21,417 true news articles from Reuters.com, all collected between 2016 to 2017. I removed all articles that had duplicate text, leaving 17,455 unique fake news articles and 21,192 unique true news articles. While the dataset included the title, text, type, and date of the article, I only used the title and text for this study, combining the two to represent the text of the entire article.

Cleaning the Data

I began by shuffling the rows of the dataset, to ensure that order of the data doesn't introduce any biases. I then cleaned up the text by only allowing words in the English dictionary and removing "(Routers)" tags which were frequently in the true news articles, and "bit.ly" links which were frequently in the fake news articles.

Converting Text to Bag-Of-Words (BoW) Representation

I split up 80% of the data into a training set and 20% of the data into a testing set, using an arbitrary random_state of 42. I then used a CountVectorizer to tokenize the text, build a vocabulary of known words, and count the occurrence of each word in each article. I fit the CountVectorizer to the training data (X_train), transforming the text into a BoW representation. The BoW representation is a sparse matrix where each row corresponds to an article, each column corresponds to a unique word in the vocabulary, and at the intersection is how many times that word appeared in that article. I then transform the test data (X_test) using the vocabulary learned from the training data, ensuring that the same set of features (words) is used for training and testing. The number of vocabulary words in the training data was 24,202.

Utilizing a Naïve Bayes Classifier

I then created and trained a Multinomial Naïve Bayes Classifier using vectorized training data, so that it learns the relationships between the features (word counts) and the labels (classifications) based on the training set. Finally, I use this trained model to make predictions on the vectorized test data.

Extracting Most Important Features

I first extract the feature names from the CountVectorizer. I then get the log probabilities for each feature (word) for each class (Fake versus True). I calculate the feature importance scores by subtracting the log probabilities of the "Fake" class from the log probabilities of the "True" class, the resulting scores representing the contribution of each feature to the likelihood of a news article being classified as "True." Therefore, the most positive features (words) will be most associated with true articles, and the most negative features will be most associated with fake articles.

Evaluating Sentiment Scores of Words

I used the pre-trained sentiment analysis model Valence Aware Dictionary and sEntiment Reasoner (VADER) to analyze the sentiment/emotion associated with each word. Each word is calculated a sentiment score ranging from -1 (most negative sentiment) to 1 (most positive sentiment).

Results

The Multinomial Naïve Bayes Classifier was able to classify fake and news articles in the test set with 94.02% accuracy. Broken down by each class, the model was able to predict fake news with .94 precision, .93 recall, and an f1 score of .93, among 3486 fake articles in the test set. The model was able to predict true news with .94 precision, .95 recall, and an f1 score of .95, among 4244 true articles in the test set. A confusion matrix displays that 3233 fake news articles were correctly classified as fake (true negative), 253 fake news articles were incorrectly classified as true (false positive), 209 true news articles were incorrectly classified as fake (false negative), and 4035 true news articles were correctly classified as true (true positive). Overall, the model performs well at classifying true versus fake news, with high accuracy, recall, and F1-scores for both classes.

After extracting the most important features, aka the features (words) that are most associated with fake versus true classification, it was clear that the top words associated with fake news are less professional and more emotional, and more often adjectives, than the top words associated with true news. Among the top 20 words associated with fake news were "hilarious", "creepy", "idiotic", and "narcissistic" for example. Meanwhile, among the top 20 words associated with fake news, were places like "Barcelona" and "Bali" and more neutral and professional terms like "accession" and "provincial".

Top Features for True Classification	Sentiment Scores	Top Features for Fake Classification	Sentiment Scores
tusk	0.0	bundy	0.0
graft	0.0	hilarious	0.4019
bali	0.0	hilariously	0.0
unionist	0.0	tantrum	-0.4215
shi	0.0	boiler	0.0
dup	0.0	creepy	0.0
barcelona	0.0	gage	0.0
accession	0.0	subscribe	0.0
provincial	0.0	em	0.0
impasse	0.0	idiotic	-0.5574
soe	0.0	uninterruptible	0.0
kang	0.0	lovable	0.6124
extradite	0.0	ca	0.0
ria	0.0	spore	0.0
regulator	0.0	rep	0.0
abbas	0.0	narcissistic	0.0
secessionist	0.0	pundit	0.0
tass	0.0	tapper	0.0
corp	0.0	mooch	-0.4019
parliamentary	0.0	yr	0.0

Figure 1: The top 20 features (words) used to classify fake versus true news, and their corresponding sentiment scores.

After conducting sentiment analysis on the most frequent 20 words linked to true and fake news, a notable pattern emerged. The words associated with true news consistently yielded a sentiment score of 0.0, indicating a neutral

sentiment. In contrast, the words tied to fake news exhibited greater variability in sentiment scores, ranging from negative to positive.

Among the top 100 words associated with fake news, the average sentiment score was approximately -0.071, with a median of 0.0 and a standard deviation of 0.237. In comparison, the average sentiment score for the top 100 words associated with true news was around -0.006, also with a median of 0.0, and a standard deviation of 0.046. This suggests that, on average, words linked to fake news convey sentiment that is ten times more negative than those associated with true news and displays a higher overall variance from neutrality.

I generated a word cloud of the top 100 words associated with fake news, where the size of each word is proportional to the absolute value of its sentiment score. I didn't include the word cloud of the top 100 words associated with true news, as there were only 3 words that had non-zero sentiment scores.

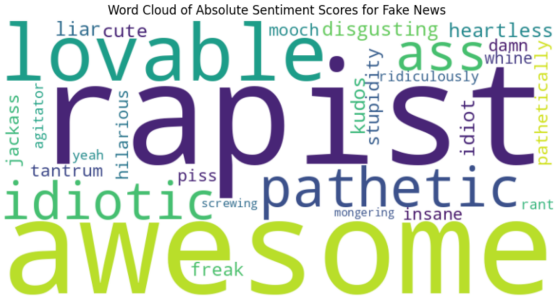

Figure 2: A word cloud of the top 100 words associated with fake news - the size of each word is proportional to the absolute value of its sentiment score.

Discussion

Supporting Hypothesis

My results support my hypothesis, that the frequency of emotion-related words can be used for detecting fake news since a Naïve Bayes Classifier, with BoW representation of news articles as input, was able to classify fake and true news with high accuracy and the top features associated with fake news had higher absolute values of sentiment scores. This means that the model was utilizing the frequency of emotion related words to make its predictions.

Connection to Psychology Studies

As detailed in the Background section, psychologists have found through numerous different studies using human participants that emotions increase the likelihood that people will believe and spread fake news. My findings demonstrate that fake news does in fact disproportionately have words related to emotion than real news, lending evidence as to how fake news can evade detection and gain virality. However, further research needs to be done to understand if simply using emotional words results in an emotional response in readers.

Connection to Other Machine Learning Studies

My findings corroborate many of the findings of other machine learning studies but through a different approach. In line with the studies that found negative sentiment was more prominent than positive sentiment in fake news (Paschen, 2019; Nanath et al., 2022), I found that on average the top 100 words associated with fake news have an average sentiment of -.071 indicating that on average the top 100 words express negative sentiment. My findings are also similar to the studies that found that using emotion as a feature improves a fake news classification model (Choudhry et al., 2022; Mackey et al., 2021), in that it demonstrates that even a model that doesn't explicitly have emotion as a built-in feature ends up detecting a pattern of emotion in fake news and utilizing emotional words to distinguish between fake and real news.

Next Steps

As I mentioned in my introduction, individual words do not capture the complexity of negative sentiment and themes of morality. Sentiment analysis of individual words is not always accurate, as words can have multiple meanings and can be used in figurative senses like metaphors; furthermore, the average sentiment analysis of the words in a text does not necessarily reflect the sentiment of the entire piece. Thus, employing an n-gram Naïve Bayes Classifier that considers sequences of words rather than individual words, may be able to capture the contextual relationships between words and provide more useful sentiment analysis.

Additionally, I could utilize a technique called Term Frequency-Inverted Document Frequency (TF-IDF), to reduce the weights of less significant words or remove stop words all together in my dataset. This didn't appear to be a massive issue, as none of the top 100 words associated with fake or true classification were stop words, however applying TF-IDF or removing stop words may improve the accuracy of the model. Additionally, more data cleaning steps may be needed, despite already having a step of removing words that are not in the English dictionary, as some of the top features that the model used still do not seem to be meaningful English words.

Next, I would like to try using other sentiment analysis and emotional-analysis tools, as VADER is primarily used in the context of social media or informal pieces of text, and therefore may not be the best suited to evaluate sentiment analysis for formal full-length news articles. Additionally, emotional-analysis tools may provide more insight rather than simply positive or negative sentiment of a word.

It would also be valuable to test this exact methodology on other datasets. On the one hand, it's unclear how reliable Polifact.com and Reuter are for fake and real news articles respectively, so this dataset may contain inaccuracies and biases. Regardless, testing this methodology on other datasets

would also confirm that these findings generalize across other datasets.

Conclusion

In conclusion, my study has two key findings. First, it offers a highly accurate machine learning model for classifying fake versus real news. Second, the model's features and weights confirm that emotional words are more prominent in fake news and are key in detecting fake news.

Acknowledgments

Thank you to my professor Tom Griffiths and my preceptor Cody Dong for their input and feedback on my project proposal.

References

Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting Fake News with Machine Learning Method. *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 528–531. https://doi.org/10.1109/ECTICon.2018.8620051

Bago, B., Rosenzweig, L. R., Berinsky, A. J., & Rand, D. G. (2022). Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. *Cognition and Emotion*, *36*(6), 1166–1180. https://doi.org/10.1080/02699931.2022.2090318

Carrasco-Farré, C. (2022). The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, *9*(1), Article 1. https://doi.org/10.1057/s41599-022-01174-9

Choudhry, A., Khatri, I., Jain, M., & Vishwakarma, D. K. (2022). *An Emotion-Aware Multi-Task Approach to Fake News and Rumour Detection using Transfer Learning* (arXiv:2211.12374). arXiv. https://doi.org/10.48550/arXiv.2211.12374

Corbu, N., Bargaoanu, A., Durach, F., & Udrea, G. (2021). *Fake News Going Viral: The Mediating Effect Of Negative Emotions*. *4*, 58–85.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), Article 1. https://doi.org/10.1038/s44159-021-00006-y

Fact check: Why do we believe fake news? – DW – 07/08/2023. (n.d.). Dw.Com. Retrieved December 15, 2023, from https://www.dw.com/en/fact-check-why-do-we-believe-fake-news/a-66102618

Fake News Detection Datasets | ISOT research lab. (2022). Uvic.Ca. https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/

Fleming, J. (n.d.). *LibGuides: Fake News: Consequences of fake news*. Retrieved December 15, 2023, from https://libguides.exeter.ac.uk/fakenews/consequences

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 900–903. https://doi.org/10.1109/UKRCON.2017.8100379

Grieve, J., & Woodfield, H. (2023). The Language of Fake News. *Elements in Forensic Linguistics*. https://doi.org/10.1017/9781009349161

Horner, C. G., Galletta, D., Crawford, J., & Shirsat, A. (2021). Emotions: The Unexplored Fuel of Fake News on Social Media. *Journal of Management Information Systems*, *38*(4), 1039–1066. https://doi.org/10.1080/07421222.2021.1990610

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Linguistic cues could be key to exposing fake news—Department of Literature, Area Studies and European Languages. (n.d.). Retrieved December 15, 2023, from https://www.hf.uio.no/ilos/english/research/news-and-events/news/2022/linguistic-cues-could-be-key.html

Mackey, A. L., Gauch, S., & Labille, K. (2021). *Detecting Fake News Through Emotion Analysis*.

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, *5*(1), 47. https://doi.org/10.1186/s41235-020-00252-3

McIntyre, L. (2021). The Hidden Dangers of Fake News in Post-Truth Politics. *Revue internationale de philosophie*, *297*(3), 113–124. https://doi.org/10.3917/rip.297.0113

Nanath, K., Kaitheri, S., Malik, S., & Mustafa, S. (2022). Examination of fake news from a viral perspective: An interplay of emotions, resonance, and sentiments. *Journal of Systems and Information Technology*, *24*(2), 131–155. https://doi.org/10.1108/JSIT-11-2020-0257

Paschen, J. (2019). Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management*, *29*(2), 223–233. https://doi.org/10.1108/JPBM-12-2018-2179

Rafique, A., Rustam, F., Narra, M., Mehmood, A., Lee, E., & Ashraf, I. (2022). Comparative analysis of machine learning methods to detect fake news in an Urdu language corpus. *PeerJ Computer Science*, *8*, e1004. https://doi.org/10.7717/peerj-cs.1004

Ray, S. (2017, September 11). Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Shariatmadari, D. (2019, September 2). Could language be the key to detecting fake news? *The Guardian*. https://www.theguardian.com/commentisfree/2019/sep/02/language-fake-news-linguistic-research

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559