Issues in Information Systems
Volume 23, Issue 2, pp. 52-61, 2022

DOI: https://doi.org/10.48009/2_iis_2022_105

Analytics and visualization of detecting fake news contents accuracy

Stephanie Molina, *Carson-Newman University, mstephanie159@gmail.com*
Seongyong Hong, *Carson-Newman University, shong@cn.edu*

Abstract

Over the past decade years, the growth of online social media has greatly facilitated the way people communicate with each other. Therefore, the world of the internet has become a very hard place to be able to figure out what information available is factual or not. Recently aggressors of this have been sources publishing articles as real news to sway people to behave a certain way. By training a program to identify what is the difference between a factual piece of writing and the nonfactual, consumers would be able to better navigate the political, spiritual, and physical world around them better informed than they otherwise might have been. In this research, the experiment takes a piece of writing and decodes what words were used and compares its previous learning to find out if the words used lead, it to know that something is nonfactual. It then comes up with a percentage based on that and other smaller details like punctuation to determine the result of the report detailing its factuality percentage based on the machine's findings. The purpose of this research is to help people have a better knowledge of which news sources are the most reliable to be able to make smarter decisions on what to believe in this era of misinformation. It is important that people learn to question and research information for themselves and that is what we hope to promote in this paper.

Keywords: detecting fake news, factuality detection, machine learning

Introduction

Over the past decade years, the growth of online social media has greatly facilitated the way people communicate with each other. Users of online social media share information, connect with other people and stay informed about trending news (Stahl, 2018). The change in the news media is a noticeable one in that it is difficult to gauge which news sources are factual and which are spreading "fake news" just for consumer engagement. This identification of these news outlet's reliability is what is set out to be done in this paper. By training a program using machine learning, it is able to identify the factuality of a given text by combing through for the words and phrases often used (O'Brien, 2018). In doing so, it is possible to graph this numerical representation to compare to the other news websites' articles. This allows consumers of news media to make a better cognitive decision on what to believe when it comes from a particular news source.

The spread of fake news has become much more prevalent today, and it is important to know which news sources can be relied upon for factual information only. By using a python fake news detection program that uses machine learning to inspect an article for the words and phrases used, it is possible to detect how

factual the information presented is (Halgekar and Kulkarni, 2020). This data can then be averaged across several articles for a particular news source website and can then be compared to others to have a better understanding of where media consumers should go if they want "real news". The python program takes reference to many models of word detection such as the one used in the Automatic Detection of Fake News (Perez-Rosas and Kleinburg, et al. 2017). In our research, the news sources obtained are from a variety of local news websites chosen randomly. We will be using Natural Language Processing software to analyze the ten articles chosen (Busioc, et al. 2020). Natural Language Processing, or NLP for short, is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to make programs to analyze and process large amounts of natural language data (Srivastava, 2020).

The rise of social media as an important part of our lives has made it easy for everyone around the world to share and spread any kind of information to the other reaches of the world. This is positive in that it has become easier for citizen journalism if you will, which allows for more raw news to be spread without being sanitized for viewership on the air. Negatively, though, this also allows for unchecked news to be spread as well. False information is fed to the masses without any responsibility for what it could cause. Usually, the reasons for doing this are malicious whether it is to slander someone or a company, or to just to attract people for attention to earn money off the online advertisements. A lot of times they have been used to influence the public's thoughts and opinions on a political candidate for an election or a certain celebrity. Consumers of media have a right to know how factual a news source is, and that news source isn't going to reveal that evidence outright regardless of what people may request. Thus, this is an analysis of a couple of websites that will be shown in a visual way of how factual they really are and the deeper dive into what the found data means.

It is important to people in their daily lives as it is since according to a Pew Research Center article, "Almost three out of four U.S. adults (71%) watch local television news…" This shows just how important it is for everyday people to have access to and understand that the local news channels aren't the only place to get your information, especially in this day and age of the internet. The purpose of this research is to help people have a better knowledge of which news sources are the most reliable to be able to make smarter decisions on what to believe in this era of misinformation. It is important that people learn to question and research information for themselves and that is what we hope to promote in this paper. The interesting part in doing so is finding out how to figure out if something written down is truthful or not. Talking face-to-face in person where you can read body language and hear a person's voice are the usual ways of helping you determine whether or not a person is being truthful, but how do we transfer such ways of deducing into computer language? The key is that we can't, at least not in the same way. To do this in computer talk, we must look at a person's way of writing instead. Everyone has a writing personality that can be seen when reading their works. It's not always easy to see or interpret because words on a paper or screen don't have any visual or auditory emotion to accompany them like you would get in person. This study will be using Python to check the accuracy of 5 local news websites using a handful of random articles recently published to find out the average rate of accuracy each one publishes. And this will allow people to be able to check to find out and see which is the most factual news source at this time based on their most recently published articles from the last few months.

Related Work

In the past years, many researchers recently studied big data analysis as different topics or categories. For example, Reece and Hong (2021) presents big data analytics for smart sports using apache spark. Big data analytics examines large amounts of data to uncover hidden patterns, correlations, and other insights. With today's technology, it's possible to analyze your data and get answers from it almost immediately.

Accordingly, this research focuses on how to detect fake news. Fake news is differentiated by the content that mimics news media in form but not in editorial processes. In a paper called "Automatic Detection of Fake News", they developed computational resources and models to identify fake news and used two datasets they gathered through crowdsourcing and directly from the web (Perez-Rosas, Kleinburg, et al. 2017). They checked the articles for use of words, punctuation, readability, and syntax. They cross examine from many sources. This is similar to the method used in many other research papers listed below in the references. Another example is a paper that also mostly tries to make a model that will detect if a news medium is factual or not based on the words and phrases used in it. They used Natural Language Processing (NLP), which I will define later in this paper if you don't know what it is, classification algorithms in Python to help sort through the articles (Khanam and Alwasel, et al. 2021). Another good research that was researched was this paper that focuses on predicting how factual a source will be. They found that the most trustworthy sites usually have a Wikipedia page and used that to make the finding out if it was reputable easier, but it turns out it is not that useful for political bias that articles may have (Baly and Karadzhov, et al. 2018). In the paper titled "Fake News Detection using Machine Learning Algorithms", the researchers used a similar process listed above to characterize the texts they chose, but with a little difference in that they completely removed and disregarded the punctuation in those texts. It is interesting that they believed that that data was useless for what they were trying to accomplish. (Sharma, et al. 2020). Many other papers such as literature review references show similar workings and ways of processing the entered text data. One researcher checked articles by testing on tasks of increasing difficulty. They differentiate four subject areas to look at which are entertainment, history, and geography. They were able to find that the truth values they got back from their fact checker are positively correlated to the average rating given by human evaluators (Ciampaglia and Shiralkar, et al. 2015). This means that the machine is being about as accurate as any normal human might be, which could be useful. Though, I believe it should try to get different results if they want to see a reason for using this machine instead of just using your own judgment. Something important mentioned in the paper titled "Fake News Detection on Social Media using Geometric Deep Learning" is that a major key challenge faced in detecting factuality using machine learning approaches is that to have the best knowledge on what to look for you need a rich and extensive training dataset to allow it to catch more details (Monti and Frasca, et al. 2019). Another deeper look at Natural Language Processing is explained by a part-of-speech label tagging to sort each word into a category based on part of speech it is (ie: noun, verb, etc.) and are then taken into account similar to the way mine look at the number of certain words used (Collobert and Weston, et al. 2011). A thing I want to focus on for a second regarding the paper by Natali Ruchansky and others is that they accurately wrote out some points that are important. First, they also described how misinformation is a vital part of social media influence (Ruchansky and Seo, et al. 2017). They also say that it is mainly due to the difficulty humans have distinguishing true from false news. They also came into the approach using the ideas of natural language processing and how there are machine learning techniques that rely on specific features to classify certain pieces of text as factual or not. They also tell us that many linguistic characteristics are not fully understood and that the way they are looking at the data is not the only way to lead to fake news detection. In this instance they used three different specific qualities to identify their sources. They used these characteristics, text, response, and source. The text is the words used and analyzing the vocabulary and structure of the articles much like my project sets out to do. The response aspect is gathering data on how people responded or interacted with the text. This is harder to gather because there are many factors that could be used as a response. In the paper they also went over the examples there are such as Facebook likes, YouTube likes, Twitter retweets. etc. The last factor is to integrate, which combines the response and text aspects and then uses the source information from both to classify the article as fake or not. The main reason they are doing it this way is because they don't want to have to rely on social graphs or domain knowledge to be able to take a closer look at the data. They used two real world gathered datasets and they say in their paper that their Capture Score Integrate (CSI) model was able to achieve much higher classification accuracy than other existing models.

Another research I would like to look at is one that goes a bit more into the different ways you can classify data and analyze it. The paper by Ray Oshikawa and two others tells us that one condition for fake news identifiers to have good performance is to have sufficient labeled data (Oshikawa and Qian, et al. 2018). Acquiring this kind of data though is very labor intensive and requires a lot of time. Therefore, they say that many lean towards less supervised methods. Many datasets are also created from already reputable data checking sources such as Polifact and Snopes. These are however usually only used to check the factuality of tweets or rumors in general and not extensively news only.

Methodology

In this research, to calculate the validity of these news sources, we will be taking a random sample of ten of the most recent articles published within February and March 2022 from five local news sources and averaging each's factuality report to visualize into a table and graph. Figure 1 explains how we will do this using the architecture of Natural Language Processing, as I mentioned in the introduction, which is broadly defined as the automatic manipulation of natural language, like speech and text, by software (Dilmegani, 2016).

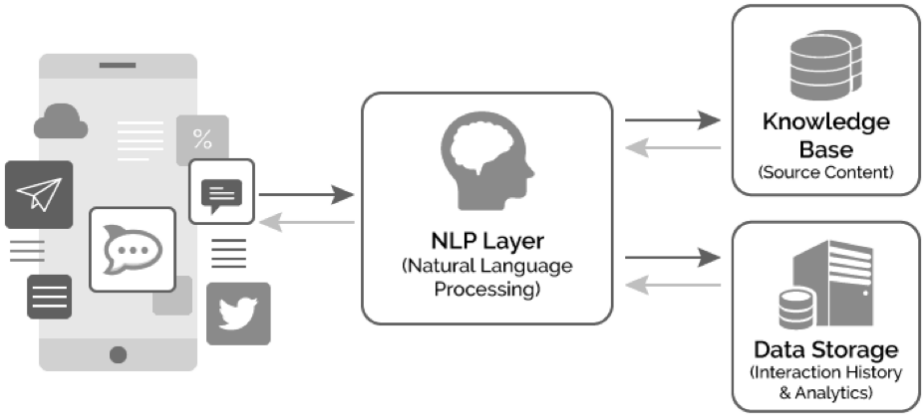


Figure 1: Architecture of NLP Processing

Therefore, to find out the validity of these news source articles we start to look at certain linguistic features such as: Punctuation. The use of punctuation is useful in that it could help to differentiate deceptive news from the actual truthful texts. By looking at punctuation characters such as periods, commas, question marks, exclamation marks, and the use of dashes, we are able to better determine truthfulness using Natural Language Processing. We will also extract features that indicate text understandability. These include content features such as the number of characters, complex words, long words, number of syllables, word types, and number of paragraphs, among others content features. We should also take grammar and such into account when looking at these texts to see if there are errors of any kind present as this would automatically lower the percentage of how factual a given article is. Then we will process and vectorize by using the Python scikit-learn library to perform tokenization and feature extraction of the text data (Pedregosa, et, al. 2011). This data is then encoded as tf-idf values. TF here stands for Term Frequency which is the number of times a word shows up in text.

$$TF(t,d) = \frac{Number\ of\ times\ t\ occurs\ in\ document\ 'd'}{Total\ word\ count\ of\ document\ 'd'}$$

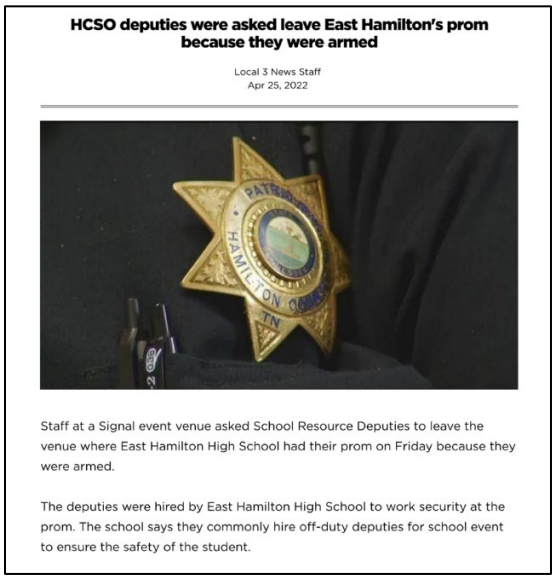
The IDF part stands for Inverse Document Frequency which will refer to words that appear many times in a document but do so in many others as well, so they are more like filler words that are unimportant.

$$IDF(t,d) = \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it})$$

The reason to use this library is because it contains useful tools like Count Vectorizer and Tiff Vectorizer (Hao and Ho, 2019). First, we train the program using pandas to read two separate excel files which contain proven true text articles and made-up fake articles respectively and sklearn to analyze the data (McKinney, 2011).

$$TFIDF(t,d) = TF(t,d) * IDF(t)$$

This is the main way research programs look at the entered text to determine factuality. It is trained this way using the two datasets of the true example data and the false example data. This suggested program is heavily biased towards the articles we used to train it. If there was a certain thing that linked all the truth articles to be similar, then that is what the program will pick up when looking at entered questioned articles. I first start by finding the ten articles for each news source I would like to use and analyze. Figure 2 is an example of an online news article. Then after acquiring all the text from the body of the article, we can move it into an excel file to allow it all to be read by the python library pandas. Figure 3 showing the raw data of excel file, after saving the data, the python program reads in the excel file data and automatically notes the total count for each word read in and can characterize them into being a TF or IDF based on previous learning data analysis using sklearn. After it has finished it can store the result into an array that will print at the end to show us all the report numbers. This number will be the reported factual accuracy of the articles which we can then study. The reported numbers will all be averaged based on the news source and that will provide an average factuality for that specific news source, which will allow us to make a judgment whenever we see articles published by that news source again.



HCSO deputies were asked leave East Hamilton's prom because they were armed

Local 3 News Staff
Apr 25, 2022

Staff at a Signal event venue asked School Resource Deputies to leave the venue where East Hamilton High School had their prom on Friday because they were armed.

The deputies were hired by East Hamilton High School to work security at the prom. The school says they commonly hire off-duty deputies for school event to ensure the safety of the student.

Figure 2: Example of online news article

	B
1	Text
2	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn t d
3	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the assumption, like many of us, that the Christopher Steele-dossier was what prompted the Russia investigation so he s been lashing out at the Department of Jus
4	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered for Homeland Security Secretary in Donald Trump s administration, has an email scandal of his own.In January, there was a brief run-in on a plane between Clark
5	On Christmas day, Donald Trump announced that he would be back to work the following day, but he is golfing for the fourth day in a row. The former reality show star blasted former President Barack Obama for playing golf and now Trump is on track to out
6	Pope Francis used his annual Christmas Day message to rebuke Donald Trump without even mentioning his name. The Pope delivered his message just days after members of the United Nations condemned Trump s move to recognize Jerusalem as the capital of Israel
7	The number of cases of cops brutalizing and killing people of color seems to see no end. Now, we have another case that needs to be shared far and wide. An Alabama woman by the name of Angela Williams shared a graphic photo of her son, lying in a hospital
8	Donald Trump spent a good portion of his day at his golf club, marking the 84th day he s done so since taking the oath of office. It must have been a bad game because just after that, Trump lashed out at FBI Deputy Director Andrew McCabe on Twitter follow
9	In the wake of yet another court decision that derailed Donald Trump s plan to bar Muslims from entering the United States, the New York Times published a report on Saturday morning detailing the president s frustration at not getting his way and how fa
10	Many people have raised the alarm regarding the fact that Donald Trump is dangerously close to becoming an autocrat. The thing is, democracies become autocracies right under the people s noses, because they can often look like democracies in the beginning
11	Just when you might have thought we d get a break from watching people kiss Donald Trump s ass and stroke his ego ad nauseam, a pro-Trump group creates an ad that s nothing but people doing even more of those exact things. America First Policies is set to
12	A centerpiece of Donald Trump s campaign, and now his presidency, has been his white supremacist ways. That is why so many of the public feuds he gets into involve people of color. One of his favorite targets, is, of course, the players in the National Fo
13	Republicans are working overtime trying to sell their scam of a tax bill to the public as something that directly targets middle-class and working-class families with financial relief. Nothing could be further from the truth, and they re getting hammered
14	Republicans have had seven years to come up with a viable replacement for Obamacare but they failed miserably. After taking a victory lap for gifting the wealthy with a tax break on Wednesday, Donald Trump looked at the cameras and said, We have essentia
15	The media has been talking all day about Trump and the Republican Party s scam of a tax bill; as well as the sheer obsequiousness of Trump s cabinet, and then members of Congress, after their tax scam was all but passed. But the media isn t quite saying w
16	Abigail Disney is an heiress with brass ovaries who will profit from the GOP tax scam bill but isn t into f-cking poor people over. Ms. Disney penned an op-ed for USA Today in which she rips the GOP a new one because she has always been cognizant of inco
17	Donald Trump just signed the GOP tax scam into law. Of course, that meant that he invited all of his craven, cruel GOP sycophants down from their perches on Capitol Hill to celebrate in the Rose Garden at the White House. Now, that part is bad enough ce
18	A new animatronic figure in the Hall of Presidents at Walt Disney World was added, where every former leader of the republic is depicted in an audio-animatronics show. The figure which supposedly resembles Jon Voight Donald Trump was added to the collec
19	Trump supporters and the so-called president s favorite network are lashing out at special counsel Robert Mueller and the FBI. The White House is in panic-mode after Mueller obtained tens of thousands of transition team emails as part of the Russian probe
20	Right now, the whole world is looking at the shocking fact that Democrat Doug Jones beat Republican Roy Moore in the special election to replace Attorney General Jeff Sessions in the United States Senate. Of course, Moore s candidacy was rocked by allegat

Figure 3: An excel file of raw data

Experimental Results

This research sets out to show a visualization of the news content's validity therefore, the result will be in a table and accompanying graph. This will allow news media consumers to better understand which news sources can be followed more closely without having to worry about fact-checking the news too much. The results gathered from the process before yields a set of number percentages that we can place in an excel file for calculations to be made as Table 1.

Table 1: Excel data of News Source Articles' Factuality

Article#	News Source A	News Source B	News Source C	News Source D	News Source E
1	50	89	68	48	62
2	48	79	60	66	78
3	39	93	73	49	69
4	25	76	88	54	53
5	40	72	65	62	79
6	59	86	76	70	67
7	62	74	61	53	77
8	43	76	72	42	69
9	48	80	85	55	49
10	24	71	72	49	68

Then we can use this to not only acquire the average truthfulness in Table 2 for each news source we can also visually show it through a line graph as Figure 4 where each news source can be compared to the others. This is a great way for the average person to make informed decisions regarding choosing which news to follow even if they had been a devoted follower of a particular news channel/website.

Table 2: Factuality Averages of 5 News Source Websites based on 10 articles.

News Source	Average Factuality
A	43.8
B	79.6
C	72.0
D	54.8
E	67.1

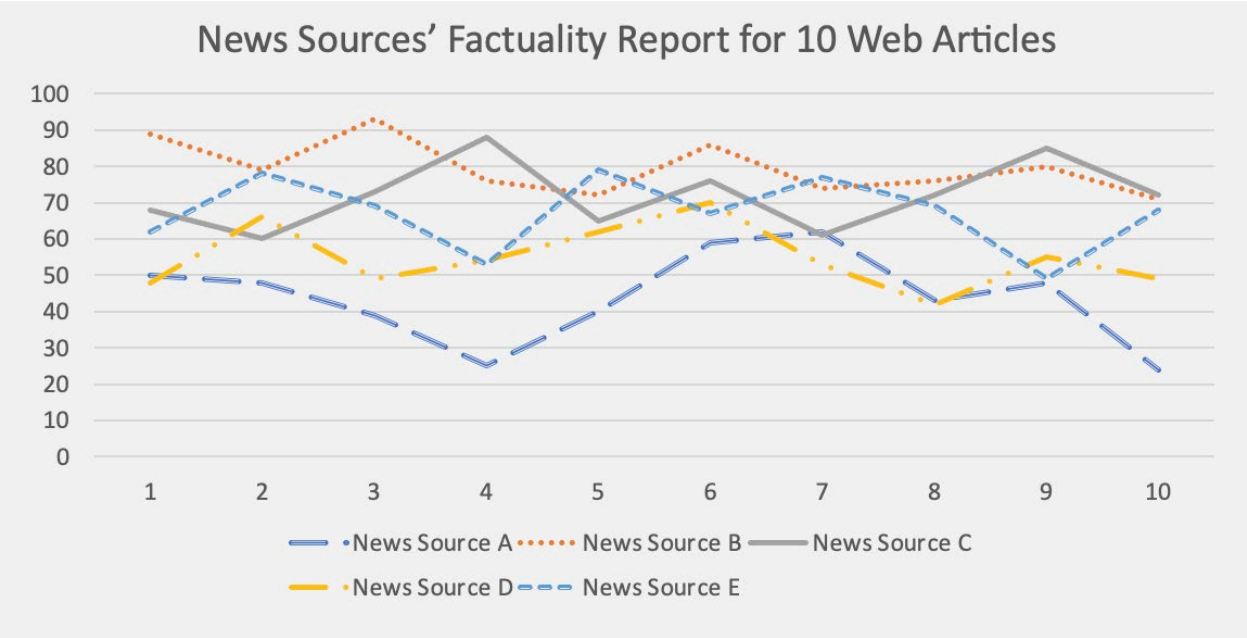


Figure 4: Each News Article Factuality

As illustrated in Figure 4 graph it is interesting to see how in line with each other the news sources all seem to be. There are ones that are worse than others, but it doesn't appear to be much. News source B ended up having the highest average which is visually visible in the graph above. This chart shows us what the averages for each news source ended up being based on the reports of a chosen 10 articles each. It seems like the five local news sources I used as an example are not very reliable in this case. Or perhaps no news outlet is able to be that high in factuality as said before due to human error.

Conclusion and Future Work

Any news source can be checked using the above methods written in this report and we specifically looked at five of my own local news sources for an example to demonstrate how this research can be helpful to media consumers who want news of only facts. I would also bet based on my findings that the probability of any local news source is probably never going to have a score higher than ninety and could perhaps even mean that any news source wouldn't achieve an average factuality score higher than ninety. I would love for someone to use this research or a similar one to show that that is the case. News sources are never going to be one hundred percent correct since we all know that in anything humans do there is human error usually in the form of bias whether intentional or unintentional. This research is by no means perfect, therefore much of any future work could be to improve this. To improve it we would need much more testing and analyzing elements. Not just looking at punctuation and words used but more such as in the papers I have used as reference. It would be helpful to go beyond the article text itself and look into the metadata of the source of the article. Also, checking to see if there are more articles with similar information elsewhere to help prove the facts. I should also mention that if this program was used by the public and became commercialized then perhaps many writers would investigate how to get around being flagged as nonfactual despite not being facts.

References

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting Factuality of Reporting and Bias of News Media Sources. https://arxiv.org/abs/1810.01765

Busioc, C., Ruseti, S., & Dascalu, M. (2020). A Literature Review of NLP Approaches to Fake News Detection and Their Applicability to RomanianLanguage News Analysis. Revista Transilvania, (10).

Chopra, A., Prashar, A., Sain, C. (2013). Natural Language Processing, *INTERNATIONAL JOURNAL OF TECHNOLOGY ENHANCEMENTS AND EMERGING ENGINEERING RESEARCH, 1*(4)*,* 2347-4289.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational Fact Checking from Knowledge Networks. *PLOS ONE 10*(10): e0141938. https://doi.org/10.1371/journal.pone.0141938

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch, *Journal of Machine Learning Research 12*, 2493-2537.

Dilmegani,C. (2016). Top 5 NLP APIs & Comparison in 2022. Retrieved March 2022 from https://research.aimultiple.com/natural-language-platforms/"

Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017). Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1803-1812. https://doi.org/10.1145/3097983.3098131

Halgekar, M. A., & Kulkarni, V. (2020). A Review: Fake News Analysis and Detection.

Hao, J., & Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. Journal of Educational and Behavioral Statistics, 44(3), 348-361.

Khanam, Z., Alwasel, B.N., Sirafi, H., & Rashid, M. (2021). Fake News Detection Using Machine Learning Approaches. *IOP Conf. Ser.: Mater. Sci. Eng., 1099*(012040). https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/pdf

Liddy, E. D. (2001). Natural Language Processing. *In Encyclopedia of Library and Information Science*, 2nd Ed. NY. Marcel Decker, Inc.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake News Detection on Social Media using Geometric Deep Learning. https://arxiv.org/abs/1902.06673

Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A Survey on Natural Language Processing for Fake News Detection. (v1). https://arxiv.org/abs/1811.00770

O'Brien, N. (2018). Machine learning for detection of fake news (Doctoral dissertation, Massachusetts Institute of Technology).

Perez-Rosas, V., Kleinburg, B., Lefevre, A. , & Mihalcea, R. (2017). Automatic Detection of Fake News. https://arxiv.org/abs/1708.07104

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Reece, J., & Hong, S. (2021). Big data analytics for smart sports using apache spark. Issues in Information Systems, 22(3).

Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A Hybrid Deep Model for Fake News Detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797-806. https://doi.org/10.1145/3132847.3132877

Sharma, U., Saran, S., & Patil, S. M. (2020). Fake News Detection using Machine Learning Algorithms. *International Journal of Creative Research Thoughts (IJCRT), 8*(6). https://www.academia.edu/46799010/IJERT_Fake_News_Detection_using_Machine_Learning_Algorithms?from=cover_page

Srivastava, A. (2020). Real time fake news detection using machine learning and NLP. Int. Res. J. Eng. Technol.(IRJET), 7(06).

Shu, K., Wang, S., & Liu, H. (2019). Beyond News Contents: The Role of Social Context for Fake News Detection. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 312-320. https://doi.org/10.1145/3289600.3290994

Stahl, K. (2018). Fake News Detection in Social Media. https://www.csustan.edu/sites/default/files/groups/University%20Honors%20Program/Journals/02_stahl.pdf

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. Unsupervised Fake News Detection on Social Media: A Generative Approach, *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19).*