For this project I chose to analyze news datasets to identify true versus fake information, or as it is sometimes described on social media, "Real" news vs "Fake" news. People spend most of their time on the internet so we are more likely to get our news from online articles instead of television. Information is spread quickly and easily through social media but how can we tell if the information we are reading is accurate? Is there a way to flag an article as misinformation? What are the consequences of an article being misrepresented as true? For this paper I will use the term Fake News in reference to articles that are or are suspected to be misinformation and Real News in reference to articles with factual information.

Newspaper articles hold a high standard for the accuracy of information that is published. Authors are required to find reliable sources, backup claims with quotes from individuals or find documentation that can be used as proof that was they write is correct. The author must research their topic using books, online articles, and interviewing individuals. [[1](https://www.thoughtco.com/how-to-write-a-news-article-1857250)] The author must include their sources so the article can be checked for accuracy by anyone who reads it. When we read articles from major newspaper companies, there is a trust, an expectation that we can believe what we are reading without doing the research ourselves.

Information online is abundant through personal and professional blogs, local and global news websites, and free video services like YouTube. According to Siteefy.com, there are 197, 046, 670 active websites as of 9/18/2022 and "175 new websites created every minute". [[2](https://siteefy.com/how-many-websites-are-there/)] If the information we are reading is presented as factual, how do we know that the source is reliable? Can we read something and trust that the author did their research prior to writing the article? Are all articles published on the web held the same standard as a company such as the New York Times?

When thinking about misinformation, how can we determine the intention of the author? Was it the writer's intention to be deceptive and publish information? Did they intend on writing an accurate, factual article or blog but not do their due diligence when researching the topic and can we clearly know their intention without actually speaking to the person?

Facebook has been accused of creating an algorithm that prioritizes negative posts to a user's feed since people are more likely to interact with content that sparks a strong emotional reaction. [1] Using the pandemic as an example, this was such a scary time for all and the spread of misinformation about a new virus was dangerous and potentially deadly. Facebook updated their system to compare information against a fact-checker and flag posts as false. [2] Twitter has also attempted to stop the spread of misinformation by asking users to flag posts that "seem misleading". [3] The existence of fake news is not new and is also not unique to the pandemic information that has been shared on social media in the past two years. What makes this so important today is just how easily information is shared to a large group of people. A system is needed to accurately identify misinformation as quickly as this information is spread and is needed across the web, not just on social media platforms.

I am glad to see these social media companies attempting to identify and stop or slow the spread of misinformation by using fact-checkers and flagging by the community. I would like to learn how the fact-checker and other methods of identifying misinformation work. What is the common thread between Fake News articles and how accurate is the algorithm used to catch the misinformation? I understand that the most accurate way to determine if an article is fake is to run it through a fact checking system or to have a professional editor check the author's sources for accuracy. Most individuals, including myself, do not have access to a fact checking system and are not professional editors who would check the sources of an article we are reading. So what can we do instead?

I plan to build a model that can categorize articles as Real News or Fake News. The purpose of the model is not to check an article for factual accuracy but instead flag an article as possible misinformation or Fake News. This flag can help the reader to make an informed decision with what they are reading. This model will be used with public article datasets found on Kaggle that are assumed to be "real news" to determine accuracy.