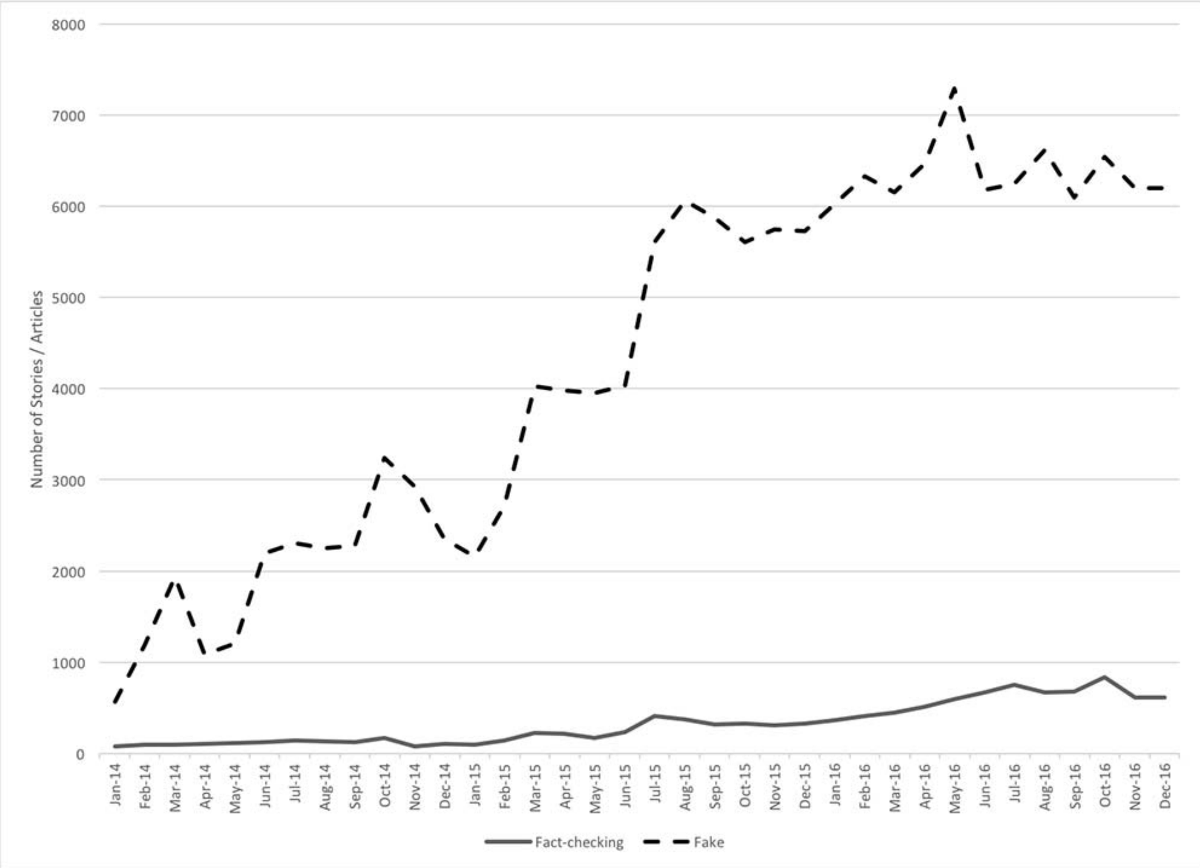
To analyze the impact of the figure of speech, sentiment, and social media on the detection and propagation of fake news over social media.

The spread of dishonest information has grown into a severe problem in the era of social media. Fake news has the potential to influence public opinion, shape election outcomes, and various other things. False narratives have the power to divide people into opposing camps, control the conversation, and undermine faith in mainstream media.



The number of fake articles vs factual articles over time

This study uses Big Data analysis to evaluate the effects of figures of speech, their sentiment, and social media content on the identification and spread of fake news to address this problem. This research intends to inform the creation of tactics that can successfully counter the proliferation of false information on social media platforms by examining the complex interaction between linguistic nuances, sentiment, and the propagation of misinformation. The analysis will include speech analysis, word sentiment analysis, analysis of various tweets, and how fake news is propagated. We'll lay out our initial concerns, pinpoint relevant data sources, explain the methods in data processing, evaluate the suitability of the data, and suggest a course of action for additional investigation and for depicting the result that the project aims for.

The data set being used for this project has been taken from various sources. Together the data set consists of various fake and true articles, various tweets posted on Twitter, and articles that have an influencing figure of speech. The data is provided in various forms such as .csv files, .json files, and tabular entries.

Question 1: How can we effectively detect and deal with the proliferation of fake news on social media platforms?

Ans: The data sources for this analysis would be various articles and how they are classified based on their content. We will also be using the tweets data set for consideration of fake news.

Next for the data cleaning and making it ready to be processed, we will perform various checks like keeping the data in a homogenous form (e.g., Lowercase), removal of unwanted data elements (like stop word), and finding out missing entries in our data and if present then finding a way to correct it.

For the data assessment, we will evaluate the coverage and representativeness of the fake articles' dataset using a variety of machine learning techniques (for instance- Naïve Bayes Classifier, Logistic Regression, Random Forest, etc). Later we will assess the quality of social media data, considering biases, noise, and potential misclassifications. Hence, we will be able to identify fake news and learn what provokes these articles' growth.

Question 2: Do sarcastic or ironic tweets have a higher likelihood of provoking the generation and spread of fake news?

Answer- We will be using the tweets data set completely to classify the figure of speech used in the tweets.

For data pre-processing and cleaning, we can first identify repeating data entries and remove them so that we can make our prediction more accurate. We can even identify and correct spelling errors in tweets as they may hinder our training data.

Based on the content of the tweets, we can identify the figure of speech being used and then classify the tweet. We can assess potential biases or limitations in the sarcasm/irony annotations, which can be an important factor in fake news generation, as the user means something else but says something else, and nowadays a slight push and lead someone to spread a piece of fake news using such figure of speech. We can utilize a range of machine learning methods, like the Naïve Bayes Classifier, Support Vector Machines (SVM), Random Forest, etc.

Backup Questions

Question B1: Can sentiment analysis of words help identify potential indicators of fake news?

Answer- In this analysis, we will make use of both the tweet and the fake news articles datasets. Also, we will be inheriting the results from the previous analysis.

Again, for the data cleaning, we will be using the same techniques used till now to clean and tidy our dataset.

Now we will apply sentiment analysis techniques to determine the sentiment polarity of words/phrases in the datasets. We will also evaluate the coverage and representativeness of sentiment-related data in the datasets. Later we will assess the accuracy and relevance of sentiment analysis along with the tweets containing similar content to determine whether the tweet is related to fake news or not.

Question B2: Can machine learning algorithms effectively differentiate between fake news and genuine news articles before they get published to the public?

Answer: For this analysis, we will be additionally using the true dataset which contains only the genuine articles, along with the fake news dataset.

For the data cleaning, we will be standardizing formats, removing irrelevant information, and performing text normalization in the datasets.

For the model training and evaluation, we will train and evaluate multiple algorithms of machine learning (e.g., SVM, Random Forest, or deep learning models) using the datasets to identify whether a news story is real or fraudulent. Using all the learning done in this task, we will be easily able to distinguish future articles as fake or not.

References:

1- www.kaggle.com. (n.d.). *Fake and real news dataset*. [online] Available at: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

2- Bostonia. (n.d.). *Fake News Influences Real News*. [online] Available at: https://www.bu.edu/bostonia/2017/fake-news-influences-real-news/ [Accessed 12 Jun. 2023].

3- www.kaggle.com. (n.d.). *Tweets with Sarcasm and Irony*. [online] Available at: https://www.kaggle.com/datasets/nikhiljohnk/tweets-with-sarcasm-and-irony [Accessed 12 Jun. 2023].

4- www.kaggle.com. (n.d.). *News Headlines Dataset For Sarcasm Detection*. [online] Available at: https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection.