ML Mini Project 2023:

Fake News Detection

SUBMITTED BY:

HARIKA KONDUR

REHAN REDDY

GOKUL BABU

VINEESH

SRISHANTH

Acknowledgement

We are thankful for the help and cooperation of the college authorities, our ML teacher, Dr. Nagaraj Naik for the successful completion of our investigatory project, ”Fake News Detection“

We are also thankful to the ICAS, Manipal for letting us utilize their resources and lab to complete our project.

We would also like to express our gratitude to our parents and classmates for their well-meant support and for giving us a helping hand whenever required.

Table of Contents

ABSTRACT

Social media consumption has risen drastically over the past decade, as platforms such as Facebook, Twitter, Instagram, and Snapchat have become ever-present in our daily lives. The following are factors that have contributed to the rise of social media consumption:

1. Increase in smartphone usage: The widespread availability of smartphones has made it easier for people to access social media platforms from anywhere and at any time.

2. Proliferation of social media platforms: Over the years, new social media platforms have emerged, offering a wider range of content and features, catering to different audiences and interests.

3. Shift in media consumption habits: Traditional media, such as television and newspapers, are no longer the primary sources of information and entertainment for many people, who now rely on social media for news, entertainment, and social interaction.

4. Personalization and algorithmic curation: Social media platforms use algorithms to personalize users' feeds based on their interests, behavior, and engagement patterns, making the content more relevant and engaging.

5. Influencer culture: The rise of influencer culture has also contributed to the growth of social media consumption, as people follow their favorite influencers and celebrities on social media platforms.

The rise of social media consumption has had a profound effect on the way we communicate, consume information, and interact with others. While it has its benefits, it also raises concerns about privacy, misinformation, as well as on the state of our mental health and well-being. People are increasingly using the internet to stay informed, sharing millions of posts, articles, and videos across platforms such as Instagram, Twitter, and YouTube. The rapid adoption of social media has led to a rise in user information-sharing, with fake news becoming a part of our digital daily routines.

Misinformation is spread in part because social media platforms fail to confirm the reliability of news. This makes it simple to share images and videos which appear to be real but have been strategically manipulated. Misinformation has long been known to have a significant impact on public opinion and discourse.

It can take various forms, such as misleading headlines, manipulated images or videos, and false or exaggerated claims.

It has the ability to be used for propaganda and manipulation. Individuals, groups, or organizations may create and spread fake news to influence public opinion or achieve political goals. Phishing attempts also use simulated information and take advantage of online users' trust. Contact forms that appear to be authentic are used to collect personal data for the purpose of identity theft. Email hoaxes in the form of chain mail, which threaten recipients if they do not share an email, are another frequent occurrence.

Another reason is that fake news can be profitable. Clickbait headlines and sensationalized stories can attract a large number of clicks and views, generating revenue through advertising.

Because of their widespread use and the ease with which information can be spread, social media platforms have become a popular medium for the spread of fake news. Some of the ways that social networking sites contribute to the propagation of misinformation are as follows:

-Lack of fact-checking mechanisms: Social media platforms do not always have procedures in place to verify the legitimacy of news before it is shared, meaning that false stories can spread easily without being challenged.

-Algorithmic bias: Social media algorithms are designed to keep users engaged and often prioritize content that generates clicks and likes, regardless of its authenticity, in turn causing stories to be promoted more widely than accurate news.

-Echo chambers: Social media platforms allow users to self-select the information they see, creating echo chambers where users only see information that confirms their existing beliefs, leading to users only being exposed to a limited set of viewpoints.

Social media platforms have a responsibility to ensure that accurate information is displayed on users’ feeds by implementing fact-checking mechanisms and promoting a diverse array of viewpoints.

Finally, the circulation of fake news can also be attributed to a lack of media literacy and critical thinking skills. With the abundance of information available online, it can be difficult for people to distinguish between real and fake news, particularly when they are not trained to evaluate sources and evidence. One instance of this are the Bitcoin scams, As cryptocurrencies have risen in popularity in recent years, so has the number of scams. To increase public trust, alleged bitcoin traders publicized their services using fake reviews by famous people. Potential investors were deceived and were given a promise of high returns based on the reviews.

Additionally, social media algorithms can contribute to the spread of fake news. These algorithms are designed to show users content that aligns with their interests and beliefs, creating a feedback loop that reinforces certain narratives and ideas, even if they are not based on factual information.

With the exponential rise of online news websites and social media sites in the past decade, there has been a lot of false reporting and consequent sharing of these false articles on social media websites. This has led to multiple arguments, threats, and the rise of conspiracy theories due to the subsequent fallout over social media websites like Twitter and Instagram between users. These theories and arguments, when reported by the traditional media, draw the attention of a huge chunk of the population, thus bringing it into the mainstream media. As of 2023, almost anyone can create a news website that shows up on the top results section of web searches. This, combined with reporting from the traditional media, attracts a lot of people to these websites when people search it online, where they can be fed with arguments in favor of news reports that are essentially false. These people who get convinced talk about the same article with multiple people from their social group, and this sparks a huge chain reaction that becomes difficult to break, as seen during the Covid-19 pandemic.

During the entire duration of the pandemic, there were multiple conspiracy theories stemming from fear, confusion, and uncertainty, and the news from the mainstream media was lost in the overload of information being posted online. This created a perfect storm where the fake news cycle could not be broken, even resulting in mass violence against people of Asian descent. Social media websites like Instagram and Twitter created fact-check badges based on keywords in posts and images but it was too late as the damage had already been done, the badges also falsely labeled information and could be tricked

easily and were popping up on everyone's feed constantly leading to users ignoring the labels most of the time.

This creates the need for a comprehensive fact-checking algorithm that "reads between the line" and understands paraphrased human conversations. This becomes possible through machine learning.

Online disinformation and fake news circulation can have detrimental effects on individuals, society, and global affairs.

Some of the potential consequences of fake news include:

Misinformation: The spread of fake news can lead to the circulation of false information, which can mislead people and cause them to make decisions based on inaccurate information.

Erosion of Trust: The widespread dissemination of fake news can erode people's trust in the media and other sources of information. Damage to Reputation: Fake news can damage the reputation of individuals, organizations, and even entire countries.

Increased Tension and Conflict: Fake news can create tension and conflict between different groups of people, and can even cause political instability, particularly if it is used to spread hateful or divisive messages and sway public opinion.

In some cases, this mistrust results in incivility, protest over imaginary events or violence. In one infamous case in 2016, a fake news story (and the comments people attached to it) moved one man to shoot up a pizzeria that was linked by false claims to human trafficking and a presidential candidate. In this incident known as “Pizzagate," a man with a semi-automatic rifle walked into a Washington, DC pizza joint and fired shots. Alt-right communities first created this piece of fiction, and fake news websites promoted the lie by referencing specific locations. It was then circulated on Twitter by people in the Czech Republic, Cyprus, and Vietnam, as well as many bots, getting the story much additional attention. Fake news, political in nature, influenced the man to fire shots inside this restaurant, nearly killing innocent people. The spread of information that was knowingly false had potentially deadly consequences. In short, fake news can have far-reaching and serious consequences, making it important to be vigilant and mindful about the information we consume and share.

K-NN is a supervised classifier and one of the most fundamental ML algorithms based on supervised learning. It is used to solve classification and regression predictive problems. K-NN considers the similarity of a new case or data to existing cases and places it in the most similar category. The K-NN algorithm not only stores all of the data, but also classifies new data points based on the similarity of the cases or data. This means that if new data is obtained, it can be classified using K-NN.

The K-nearest neighbour algorithm essentially shows that for a given value of K, it will find the K nearest neighbour of an unseen data point and then assign the class to the unseen data point by having the class with the most data points out of all classes of K neighbours.

PROBLEM STATEMENT

Our goal is to develop a reliable model that classifies a given news article as real or fake when it is trained with a certain dataset.

DATASET DESCRIPTION

The dataset has been collected from BuzzFeed News organization that was used in training and testing of the model. This dataset gathered information about the Facebook post and each of which represents a news article from the three main political news pages Politico, ABC News, and CNN.

Employees of the Buzz Feed regularly access and examine the reality of each post and split each post into 4 labels

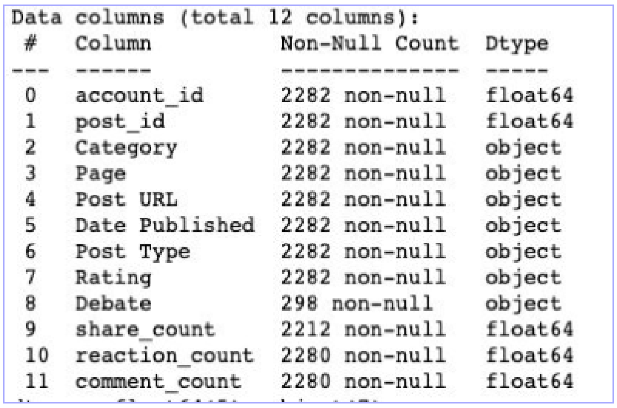
-mostly true
-mostly false
-mixture of true and false -no factual content

Dataset Features:

account_id
post_id
Category
Page
Post URL
Date
Published
Post
Type
Rating
Debate
share_count
reaction_count
comment_count

DATA PREPROCESSING

1. Remove the columns with mostly null values i.e. Debate column, as it only has 298 non-null columns



```
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   account_id      2282 non-null   float64
 1   post_id         2282 non-null   float64
 2   Category        2282 non-null   object
 3   Page            2282 non-null   object
 4   Post URL        2282 non-null   object
 5   Date Published  2282 non-null   object
 6   Post Type       2282 non-null   object
 7   Rating          2282 non-null   object
 8   Debate          298 non-null    object
 9   share_count     2212 non-null   float64
 10  reaction_count  2280 non-null   float64
 11  comment_count   2280 non-null   float64
```

2. Convert string values into numerical values

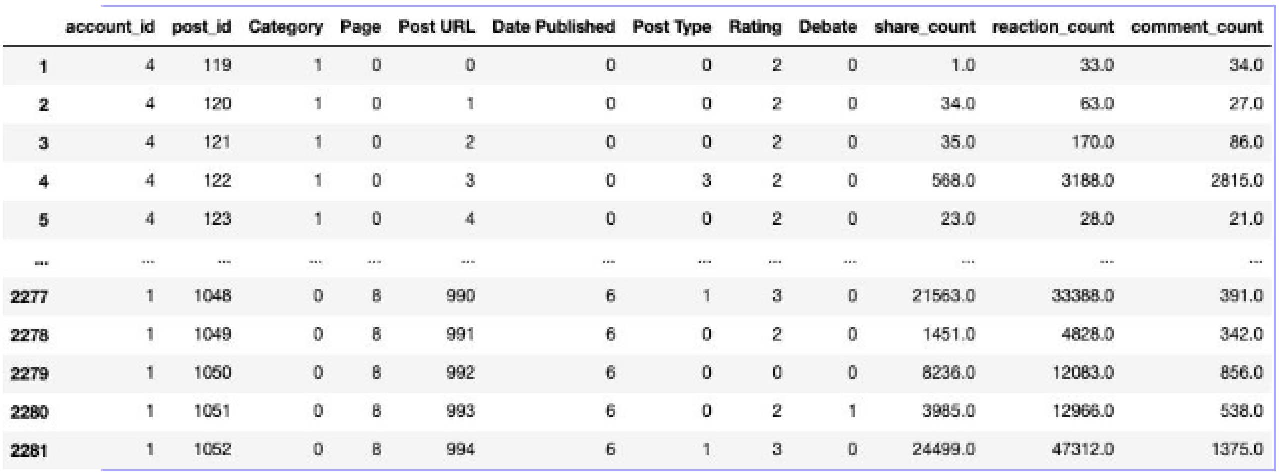
   Eg. Rating

   0:mixture of true and false 1:mostly false

   2:mostly true

   3:no factual content

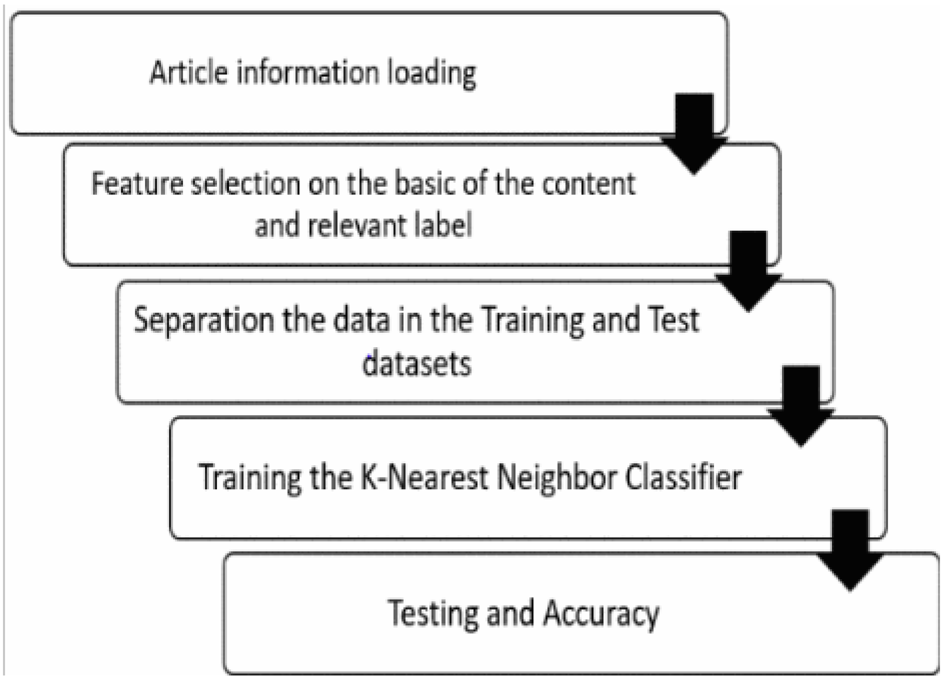
After Data Preprocessing:



	account_id	post_id	Category	Page	Post URL	Date Published	Post Type	Rating	Debate	share_count	reaction_count	comment_count
1	4	119	1	0	0	0	0	2	0	1.0	33.0	34.0
2	4	120	1	0	1	0	0	2	0	34.0	63.0	27.0
3	4	121	1	0	2	0	0	2	0	35.0	170.0	86.0
4	4	122	1	0	3	0	3	2	0	568.0	3188.0	2815.0
5	4	123	1	0	4	0	0	2	0	23.0	28.0	21.0
...	...	...	...	...	...	...	...	...	...	...	...	...
2277	1	1048	0	8	990	6	1	3	0	21563.0	33388.0	391.0
2278	1	1049	0	8	991	6	0	2	0	1451.0	4828.0	342.0
2279	1	1050	0	8	992	6	0	0	0	8236.0	12083.0	856.0
2280	1	1051	0	8	993	6	0	2	1	3985.0	12966.0	538.0
2281	1	1052	0	8	994	6	1	3	0	24499.0	47312.0	1375.0

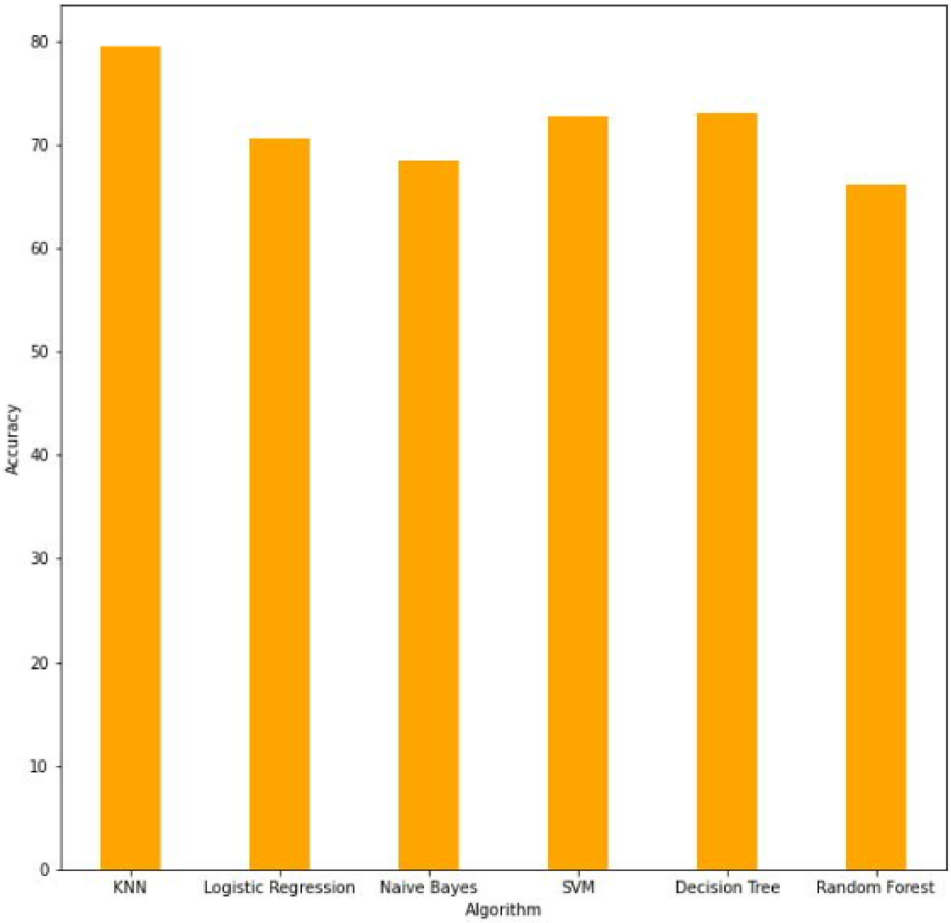
Design Methodology and Approach

The approach and design methodology for a machine learning project aimed at predicting fake news involves several steps. First, the article information is loaded into the system, and a comprehensive dataset of both genuine and fake news articles is gathered. The next step is feature selection, which involves identifying the features that could help distinguish between genuine and fake news articles. After feature selection, the dataset is separated into a training set and a testing set to evaluate the performance of the model. A KNN classifier is then trained on the training set until a suitable n value is found. This n value represents the number of nearest neighbors that should be considered when predicting the label of a new data point. In addition to the KNN classifier, other machine learning algorithms are applied to the dataset, such as random forest, decision tree, naive bayes, and logistic regression. The accuracy of each algorithm is then compared to identify the best algorithm for the task of predicting fake news accurately. Continuous monitoring and refinement of the model would also be necessary to ensure its accuracy and reliability. The dataset may also need to be updated regularly to account for new fake news articles and the changing nature of the news landscape.


Article information loading
Feature selection on the basic of the content and relevant label
Separation the data in the Training and Test datasets
Training the K-Nearest Neighbor Classifier
Testing and Accuracy

Result

The results of the machine learning project aimed at predicting fake news indicate that the KNN algorithm achieved the highest accuracy of 79.5%. Following closely was the SVM algorithm, which achieved an accuracy of 72.7%. The decision tree algorithm achieved an accuracy of 73.0%, while logistic regression had an accuracy of 70.6%. The naive Bayes algorithm achieved an accuracy of 68.5%, and the random forest algorithm had an accuracy of 66.2%. Overall, the KNN algorithm performed the best.



There could be several reasons why the KNN algorithm achieved the highest accuracy for this model. One possible explanation is that KNN is a simple but effective algorithm that can be applied to a wide range of classification problems. It works by finding the k-nearest neighbors to a data point and using their labels to classify the point.

In the case of predicting fake news, the KNN algorithm may be effective because it can capture the underlying patterns and features that distinguish between genuine and fake news articles. Additionally, the choice of K (the number of neighbors to consider) can significantly impact the accuracy of the KNN algorithm. Thus, tuning K until a suitable value is found could contribute to the high accuracy of the KNN model in this case.

Another possible explanation for the high accuracy of the KNN model is that the dataset used for training and testing the algorithm was well-suited to this particular algorithm. It is worth noting that the accuracy of a machine learning model is highly dependent on the quality and relevance of the dataset used for training and testing. Therefore, careful consideration and selection of the dataset used could have contributed to the high accuracy of the KNN algorithm in this project.

CONCLUSION

In today's fast-paced world, where information is easily accessible, the spread of fake news has become a significant problem. Detecting fake news is critical in maintaining reliable and trustworthy information sources. In conclusion, the machine learning project aimed at predicting fake news was successful in developing an accurate model for detecting fake news articles. The approach involved loading article information, selecting relevant features, separating the dataset into testing and training sets, training the KNN algorithm, and comparing its accuracy with other algorithms such as logistic regression, naive Bayes, SVM, decision tree, and random forest. The results of the project revealed that the KNN algorithm achieved the highest accuracy of 79.5%, while other algorithms achieved varying levels of accuracy. The success of the KNN algorithm in this project could be attributed to its simplicity, effectiveness, and the relevance of the dataset used. Overall, the results of this project are promising and demonstrate the potential of machine learning algorithms in detecting fake news articles accurately.

Working on this project pushed us to think beyond our limits and was an enriching experience. Completing this project filled us with immense satisfaction and thus, we would like to thank our professors for giving us the opportunity to work on this project.