Developing Machine Learning Model for Predicting Social Media Induced Fake News

David Langley, Caoimhe Reidy, Mark Towey, Manisha Manisha, Denis Dennehy

Developing Machine Learning Model for Predicting Social Media Induced Fake News

David Langley[1], Caoimhe Reidy[1], Mark Towey[1], Manisha[1], Denis Dennehy[1]

[1]National University of Ireland, Galway, Galway, Ireland

Abstract. Fake news has been associated with major global events such as Covid-19 and the political polarisation of the US presidential election in 2016. This paper investigates how fake news has affected society and advance understanding of the nature of its impact in the future of democratic societies. Taken from large datasets consisting of over 23,000 fake news story words and over 21,000 true news story words we use descriptive and predictive analytics, partly analysing more than 350 words during the selected period of October 2016 to April 2017. The findings show that Trump was the most popular word for both true and fake news. In this study, we compare and contrast the words used and the volume of true versus fake news stories related to the election and the inauguration. This study makes an important contribution as it develops a predictive model that highlights the severity of political polarization and its consequences in democratic societies, which inevitably have implications for inclusive societies in the 21st century.

Keywords: Fake News, Social Media, Echo Chambers, Filter Bubbles, Machine Learning, Polarization.

1 Introduction

Fake News can be defined as online publications that are intentionally false in order to mislead readers [1]. Although fake news is not a new a phenomenon, social media has intensified its' severity due to the rate in which news can be spread, regardless of whether it is true or false [1]. The need for fake news detection is greater than ever, as the implication of this false information is becoming increasingly dangerous for democratic societies [2]. Fake news publications through news outlets and social media have had major influence in the outcome of many worldwide events such as the US presidential election in 2016 and more recently the COVID-19 pandemic. The way in which algorithms work on social media platforms, such as Twitter and Facebook, can facilitate the creation of '*echo chambers'* (e.g., situations where individuals "hear their own voice") [17] and *'filter bubbles* (e.g. whereby like-minded individuals are not exposed to contrary perspectives or opinions, which can lead to tunnel vision and enabling confirmation bias [2]. The impact of echo chambers includes excluding alternative perspectives [18] and political chaos in many contexts [8]. Should the public lose trust in media outlets, it is very damaging to society, thus in this study, we build a model to

predict whether a publication is true or false in a bid to restore faith in news sources [3].

The aim of this study *is "to use advanced analytics to identify and predict whether news is fake or true".* This will be achieved through the means of machine learning. Machine learning models will be compared to determine an optimal model.

To achieve this aim, we seek to answer to interrelated questions:

1. What are the most commonly used words in fake news posts?
2. Does a major news event increase or decrease the amount of fake news created?

The paper is structured as follows. First, a review of background literature is presented. Next, the research methodology used to extract and clean data for the purpose of analysis is outlined. Then, discussion of key findings follows. The paper ends with a conclusion.

2 Background Literature

The prevalence of social media (e.g., WeChat, Facebook, LinkedIn, Twitter) has been a catalyst to inducing a polarized society [5]. Fake News is a global issue with its consequences becoming more severe by the day that the World Economic Forum (2018) raises concern that is the greatest threat to society due to the speed at which 'digital wildfires' spread on a global scale. More recently, Tim Cook (CEO, Apple) criticized the facilitating role that technology companies play by prioritizing conspiracy theories and violent incitement because of their high rates of engagement [7].

In response, it has led researchers to explore social media induced polarization from different theoretical lenses to study specific social media platforms and particular attention given to fake news [6]. Researchers' fake news and polarisation in varied contexts, such as politics [8], new framing [9], and modelling the combination of bias and polarisation to examine the impact of misinformation in social media networks [10].

[2] investigates the issue of fake news in the context of social media polarisation. It is apparent from this paper that although there is increased access to information, this does not lead to better informed citizens. It is argued that it leads to increased societal polarisation, this is because of people settling into 'ideological neighbourhoods,' whereby individuals experience opinions and views of like-minded individuals [2]. Xu argues that democracy is in danger. Techniques such as filter bubbles and echo chambers paired with fake news stories may distort the assumption that society is well-informed and may amplify confirmation bias [2]. A diffusion drift model used in this paper shows that increased information access contributes to growing polarisation – assuming the presence of confirmation bias and people looking for more outlying content. Technological solutions are proposed to help this issue. A previous study [4] conducted research to understand whether fake news differs systematically from real news in style and language use. The study reports that fake news articles tend to be shorter, use repetitive and less complex language, less punctuation, and less quotes [4]. Building on this body of knowledge, this study is to build on this work by providing a different

approach, by focusing on the differences in vocabulary used. Manifestations of fake news include misinformation about presidential election campaigns [11], immigration [12], religion [13], and pandemics, specifically Covid-19 [14]. We intend to present an in-depth exploration of the differences between real and fake news, with an overall aim to predict whether a news article is true or false.

3 Research Methodology

Cross Industry Standard Process for Data Mining (CRISP-DM) is an industry standard methodology that prescribes a set of guidelines to guide the efficient extraction of information from data [23]. The CRISP-DM methodology consists of six cyclical steps, namely (i) Business Understanding, (ii) Data Understanding, (iii) Data Preparation, (iv) Modeling, (v) Evaluation, and (vi) Deployment (see Fig. 1 below). It is a comprehensive and well-structured methodology that covers all the aspects of our project effectively. It is iterative where necessary, which is an important feature in an aim to monitor the models and the analysis to keep it up to date.

Fig. 1 CRISP-DM Methodology [21]

The **six stages** of the **CRISP-DM** model in the appropriate context are explained below:

Business Understanding: The first step is to gain an understanding of the topic of the research in a business context. A comprehensive analysis on the issue is crucial to attain this clear understanding and form research objectives. In order to achieve this understanding, the challenges posed by Fake News in the past is investigated, how it has affected society today and realize the prevalence of the issue by the way in which it can impact the future of democratic societies.

Data Understanding: The dataset chosen is analyzed in depth as to what it contains and represents, to best understand how to proceed. This step is vital in establishing the

objectives when combined with the business understanding process [21]. The main focus of this stage are the questions that could arise in the dataset, in an aim to produce intriguing findings and actionable insights in the future. Our data is split between two datasets – one for fake news and one for true news. The dataset was sourced from Kaggle, with verification of the quality of the data carried out.

Data Preparation: the chosen dataset(s) is prepared for analysis through cleaning and formatting techniques. This involves the transformation from the initial dataset to the final dataset. Using Excel and Python, the data will be extracted from the topics to be able to count the number of words per source. The data will be cleaned, to prepare it for analysis. It is then formatted, in an aim to classify these words as being more likely to be used in 'Fake' vs 'True' news. All null and unknown values are removed to reduce uncertainty.

Modelling: Various models are built and analyzed in this step. As there will be multiple iterations of this phase, all models with low levels of error are run to ensure the optimal model is selected and built [21]. The model is then technically assessed. To fulfil our objectives, a predictive model is required – to predict whether an inputted topic or story is likely to be fake or true. We used Support Vector Machines (SVM), which was built using Python. The model is repeatedly run to gain comprehensive findings for examination in the next phase.

Evaluation: This phase focuses on the analysis of the modelling in the context of the business needs and understanding [22]. This is vital, as it is the basis of the actionable decisions, whether the model needs to be iterated or proceed to the next stage. The accuracy and the fit of our model is evaluated – a low level of error is essential in producing conclusive results. The model outputs will be evaluated and compared against each other to enhance data-driven decisions. One optimal model is selected, and the findings are analyzed in the context of our pre-determined objectives. We apply different visualization techniques to illustrate our findings using Tableau and R.

Deployment: This phase is the main test of any study, as it decides whether actionable insights can be drawn from findings and conclusions [23]. A review of the study must be performed for future endeavors. In the Deployment section we will be discussing how the model can be deployed to identify and predict Fake News, to add value to society. Through this section, we will provide insights into how our model can tackle the issue of Fake News and reduce the severity of its consequences.

Table 1 lists the analytical tools used to extract, structure, analyze, and visualize the data.

Table 1 Analytical tools used in the data collection and analysis.

Tools used	Description of tools
Kaggle – Dataset Source	One of the largest dataset free source websites.
Advanced Microsoft Excel - Data Manipulation	A great application for storing data files and then using that current information and changing it do be user friendly, from adding in columns to break down dates to its lower forms to creating columns that further describe the data being discussed.
Python – Data Extraction and Prediction	Python is known to be the most preferred programming language for data manipulation, data analysis and data visualization.

Tableau – Data Visual Tool	One of the best visualization applications for displaying data through means of graphs, charts and more.
R – Statistical Analysis Programming Language	R is a statistical programming language, which we also plan to use due to its high efficiency with regards to its visualization and data wrangling abilities.

3.1 Data Extraction Process

The news headline data is imported into Python for the data extraction process. The data is stripped to facilitate the separation of each headline onto separate rows. Preprocessing is then carried out to clean the data for analysis, which consists of removing punctuation, numbers, symbols, stop words and involved the changing of all words to lowercase, to avoid duplication.

The news headline data is differentiated by the column 'Truth?', where it is characterized by 'FAKE' or 'TRUE' news. Tokenization is performed to count the words used in each news article headline. From this, a dictionary template of all the unique words is created which allowed for outputs; each word column contain 0's if the word is not present in the headline or the count number if the word is present, as shown in Fig. 2.

title	text	subject	date	Truth?	Day	Month	Year	Month & Ye	administr	admits	adviser	america
Republican lawmaker R	WASHING	politicsNews	April 30, 2017	TRUE	30	April	2017	Apr-17	0	0	0	0
Trump says China could	WASHING	politicsNews	April 30, 2017	TRUE	30	April	2017	Apr-17	0	0	0	0
Trump could target 'carı	WASHING	politicsNews	April 30, 2017	TRUE	30	April	2017	Apr-17	0	0	0	0
Trump invites leaders o	WASHING	politicsNews	April 30, 2017	TRUE	30	April	2017	Apr-17	0	0	0	0
Trump celebrates first 1	HARRISBU	politicsNews	April 30, 2017	TRUE	30	April	2017	Apr-17	0	0	0	0
Toned-down White Hou	WASHING	politicsNews	April 30, 2017	TRUE	30	April	2017	Apr-17	0	0	0	0
EPA says website under	WASHING	politicsNews	April 29, 2017	TRUE	29	April	2017	Apr-17	0	0	0	0
Trump to order a study	WASHING	politicsNews	April 29, 2017	TRUE	29	April	2017	Apr-17	0	0	0	0

Fig. 2 Sample dataset

4 Key Findings

This section presents key findings relating to the two aforementioned research questions, namely, 'What are the most commonly used words in fake news articles?' and 'Does a major news event increase or decrease the amount of fake news created?' During our research, we established a further discovery, which consisted of us looking at whether there was a correlation between a headline word count and their article content word count. The findings reveal that there is a correlation.

When thinking of fake news reports, click-bait headlines come to mind, ones whereby all the content is in the headline and near to nothing in the article body. Due to these news articles being fake, there is little to no evidence to support them, so, to get their polarizing message across, the headlines are crowded with content. The aim for publishers of fake news is to spread it far and wide and at speed. With this in mind, a python script was written to count the number of words in the news headlines and the article text. After exporting these to excel, a correlation scatterplot was devised, the

initial assumptions proved to be true. As shown in Fig. 3, there is a clear difference between the true and fake news headlines. The true news headline's word count varies from 4 – 17 words, with their article content varying from 22 – 2000 words. In total contrast, the fake news headline's word count varies from 4 – 42 words, with their content varying from 0 – 4900 words. Fig. 3 demonstrates that a correlation between the headline and their article content is present, and this can be drawn upon when looking to define the typing of a news headline. Upon taking a deeper look into the frequency of the article lengths, a clear observation was extracted (Fig. 4). The distribution for true news is fairly even, with two peaks at 80 and 400 words per article. The most substantial finding in the fake news section is the fact that there are 444 articles without a single word in their article. This is a clear sign that by adapting a machine learning algorithm to take the word length into account, creates a clearer distinction between the types of news headlines.

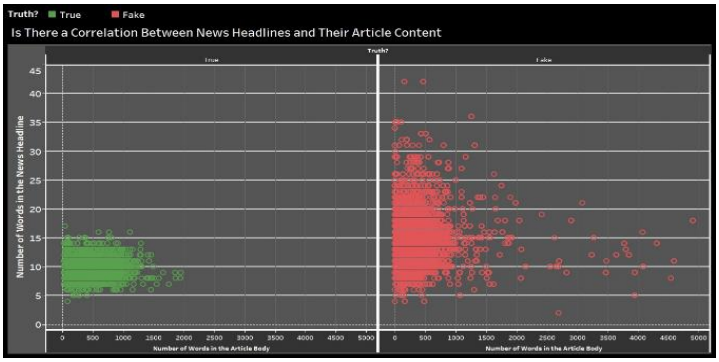

Fig. 3. Correlation between News Headlines their Article Content

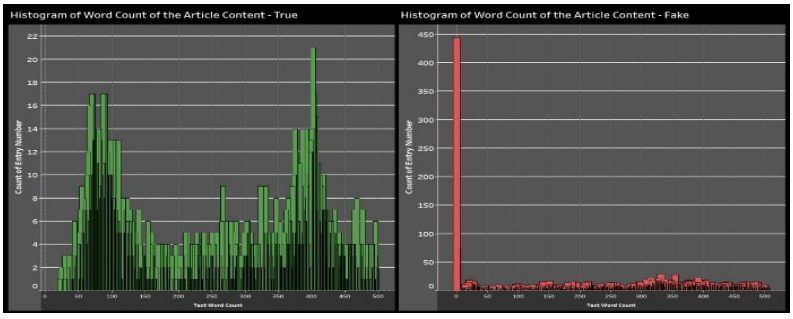

Fig. 4. Frequency Distribution of Article Lengths

The major event that was tracked in our dataset was the 2016 US Presidential Election, between Hillary Clinton and Donald Trump. The voting period for this event took place during November 2016, while the inauguration took place in January 2017. In Fig. 5 [Left], there is a significant increase to the total news headlines during the months of November and January. This increase comes from the rise in true news (Fig. 5 [Right]), as the closer to the event you get the more real news stations and articles are going to be spreading true news.

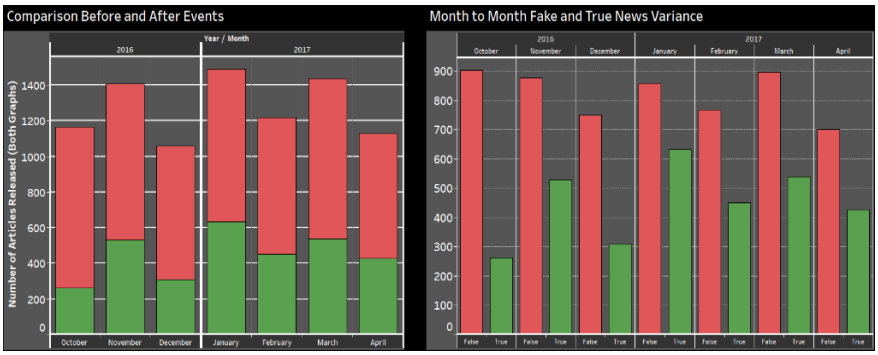

Fig. 5. Comparison of (Left) Total Headlines & (Right) Different Types of News Over Time

Leading up to the event, the percentage of October's Fake vs True News is 77.51% (Fig. 6), this shows depending on what you read, or who you follow, you may become persuaded to change your mind in your decision, thus causing polarization.

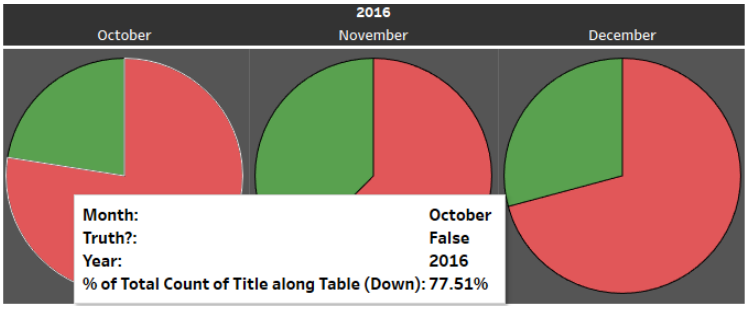

Fig. 6 Percentage of Fake VS True News Before an Event

The results indicate that during the period of October 2016 – April 2017, Trump (also including Trumps) was the most popular word for both true and fake news. The words after this then differed, with 'US', 'says', 'house' and 'White' being in the top 5 for true news headlines (Fig. 7 [Right]). While for fake news headlines the top five consisted of, 'Video', 'Obama', 'Hillary' and 'watch' (Fig. 7 [Left]), this shows that there are clear differences when it comes to which words result in fake or true news and thus will be used in our machine learning model to predict if a news headline is fake or true.

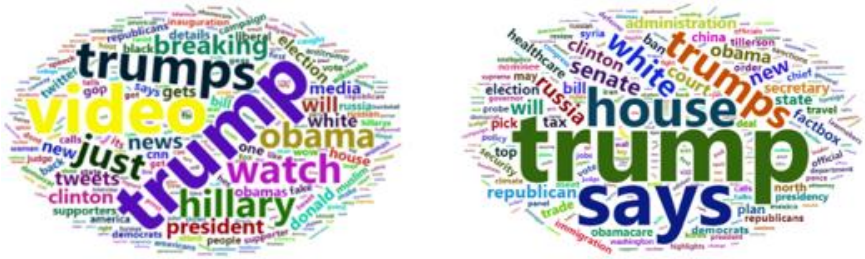

Fig. 7. Word Clouds of the Words used in Fake (Left) and True (Right) News Headlines

5 Model Creation

In our earlier analysis of the most common words in fake/real news, a Bag of Words approach was used to convert text data to numerical data. Bag of Words is the simplest way of converting text to numbers. The main limitation of the Bag of Words model is that it assigns equal value to the words, irrespective of their importance. [15]

A TF-IDF approach was then considered to convert text data to numerical data for our model creation. TF stands for Term Frequency and is the frequency of a word in a document divided by the total number of words in the document. IDF stands for Inverse Document Frequency and is the Log of the total number of documents divided by the number of documents that contain a particular word. The TF and IDF values are multiplied together to give a TF-IDF value for each word. Put simply, this means that "the words that occur less in all the documents and more in individual documents contribute more towards classification" [15]. After optimizing and comparing 8 different Machine Learning Models, the 4 models with the highest accuracy scores were Logistic Regression, Support Vector Machines, Neural Networks and Decision Tree Classification, which are discussed below.

5.1 Logistic Regression

Logistic regression is a linear classifier which is used to predict a binary outcome based on a set of independent variables. [16] For example, the output can be 0/1, True/False, Yes/No, Approved/Declined etc. It is essentially used to predict the probability of a binary event occurring. The binary event in our case, is if a particular news headline is fake news or not. For logistic regression to work, the dependent variable must be dichotomous, i.e., it can only fit into one of two categories. The dependent variable is predicted based on a set of independent variables. Independent variables are variables which may affect the dependent variable. Independent variables can either be continuous data, discrete ordinal data or discrete nominal data. [16]. Logistic regression is easier to train and implement compared to many other machine learning models and it works very well with a linearly separable dataset. Logistic regression is not very accurate with small datasets however and using logistic regression on a small dataset can often result in overfitting which means that the model is too closely fit to the training data and cannot accurately classify the test data [16].

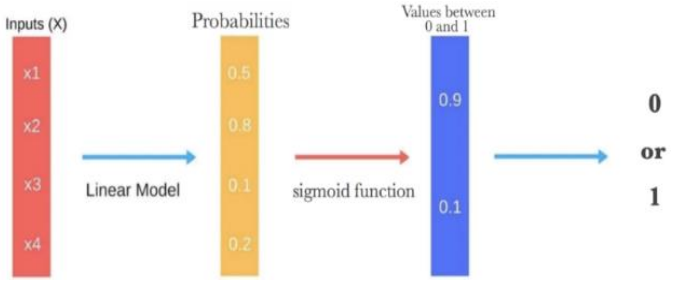

Fig. 8. - Linear Classification Model

5.2 Support Vector Machines (SVM)

SVM is a linear model that solves classification and regression problems as well as linear and non-linear problems [24]. The SVM algorithm creates a line or hyperplane which separates the data into their different classes(characteristics). Creating the hyperplane sounds easy, it could go anywhere in between the grouped variables, but creating the optimal hyperplane can be difficult. For the optimal hyperplane, you need it to be equidistant from the closest support vectors, which are the variables closest to the line originally. The challenge is to maximize the margin, distance from support vectors to the hyperplane. Once the hyperplane has been optimized, then depending on where the variable lies either side of the hyperplane, will determine its class.

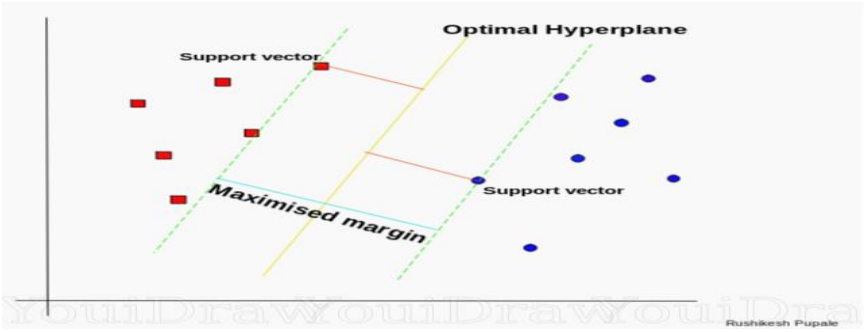

Fig. 9 Optimal Hyperplane using SVM Algorithm.

5.3 Neural Networks

A neural network is a framework that trains to make predictions or generate forecasts by going through the following steps (i) taking the data from the input, (ii) making a logical prediction, (iii) taking the forecast and comparing it to the desired result, and (iv) changing its embedding layer to correctly predict the next time.

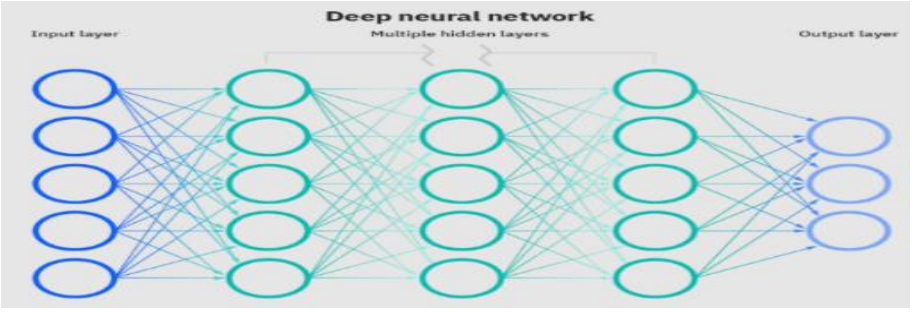

Fig. 10 Neural Network Model

Since neural networks can adapt and change input, they can produce the best possible outcome without requiring the output parameters to be redesigned. Neural networks are

made up of components such as vectors, layers, and linear regression. The data is stored as vectors, which are then stored in arrays in Python. The data from the previous layer is transformed by each layer. Since each layer extracts some representation of the data that came before it, and each layer works as a feature engineering phase.

The neural network uses the activation function. The method of training a neural network is close to that of trial and error.

5.4 Decision Tree Classification

Decision Tree Classification is a supervised machine learning algorithm. It performs a breakdown of the dataset into attributes, whereby each attribute forms a node. The setup of the tree begins at the Root Node with the attribute that provides the most information gain, the attributes with descending importance are positioned along the tree to the final node, the Leaf Node, which contains the result of the tree. When evaluating, each node is addressed as an iterative process, adhering to the decision outcome at each node, leading to a classified dataset [20].

There are two types of decisions trees; there is Categorical Variable and Continuous Variable decision trees. The dataset in question uses a categorical dependent variable, therefore, a categorical variable decision tree is implemented in this model. A key consideration is the ease of overfitting the tree – a limit is placed on the depth of the tree to prevent this; however, this introduces the risk of the results being not absolute [25]. Decision Tree's ease and speed of use outweigh any of its downfalls.

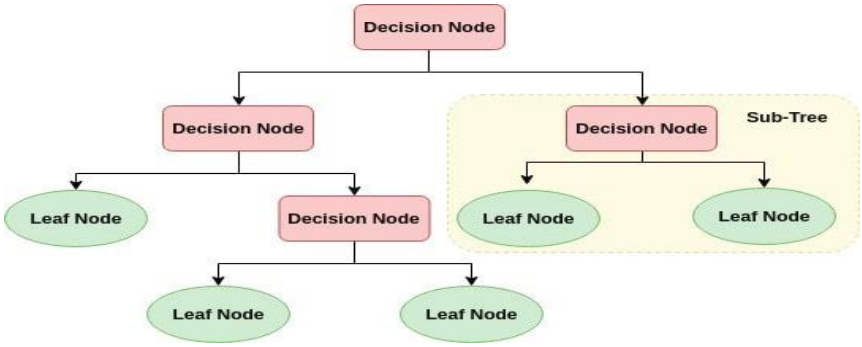

Fig. 11 Visualization of Decision Tree [20]

5.5 Model Choice Conclusion

The decision for which model that would be used was easy. By comparing the accuracy of each model, a clear winner was found, Support Vector Machines (SVM). With an accuracy score of 94.83% compared to the accuracy scores of the other Machine Learning models of, 93.04% for Logistic Regression, 94.03% for Neural Networks, and 89.26% for Decision Tree Classification. The data used was the total News Headlines between October 2016 and April 2017. After considering the possibility of better re-

sults coming from an evenly weighted dataset of truths vs fakes, the number of headlines were equated and the models were run again with only the Logistic Regression model receiving an increased accuracy score of 93.53%, but still not enough to surpass SVM as the winner of the most efficient Machine Learning model for the dataset. To make sure that this gives the greatest results in determining if a headline is a true or fake headline, the confusion matrixes were examined. These backed up the already satisfied result that SVM was the better model, with a larger percentage of correct true headlines for the SVM model vs the Neural Networks model, this doubled down the confirmation of 'best in class', which was the Support Vector Machine model.

6 Discussion, Limitations & Future Research

As our study used the 2016/17 US presidential election datasets to develop our model, Donald Trump was the big hot topic of conversation. His name was number 1 for both the true and fake news, and with how controversial he has been, this was expected [26]. In contrast, the other words in the top 5 differed between the two types, leading to the realization that we could use these words to help predict if an article is fake or true.

The uncovered correlation question when thought about made sense, but it was not until the script was being created and fine-tuned that the answer became clear. With fake news articles they just want to front load all their misinformation into the title because it is catchy and would be read by the public. Whereas with true news articles, they have the content and sources to back up their claims and only need to catch people's attention enough to get them to open the article. They do not need to have 20+ words in their articles to get noticed. There were an alarming 444 articles that had 0 words in their body, all of these being fake news ones. This demonstrates that these articles had no supporting sources and just wanted to impose malice information onto the general public.

Our proposed model could be implemented by organizations across varying industries. Social media platforms like Twitter and Face-book are the most obvious organizations which could find great value in implementing this model. Social media platforms have begun flagging election related posts which may be misleading in the wake of Donald Trump's comments regarding the 2020 U.S election [27] but this model could easily be used to flag potentially fake posts regarding a multitude of varying topics. Links to articles about a wide range of topics appear across social media on a daily basis and fake news articles can easily be flagged as fake whether these articles may be about politics, the COVID-19 pandemic, business or even celebrity gossip. We believe that if this model is implemented correctly, it will protect people from consuming fake news and will prevent people from being exploited by companies and individuals that are sharing fake news. This can benefit businesses, governments and the overall well-being of a society if people are aware of what is well informed, accurate news and what is misleading, fake news.

Our predictive model demonstrates that upon reading fake news, it can have impact the opinion of the reader. The correlation between the headlines and the content of a news article has proved that fake news headlines are likely to have more content than

true news headlines. This can have serious repercussions for society, such as the creation of an echo chamber within an individual's social media newsfeeds. Our analysis shows that we can predict that news is fake or true with the higher accuracy by using the machine learning algorithm. This can add real value to society by revealing the truth about fake news stories, resulting in the possible change of opinion (polarization) in a certain topic. A person's opinion is fundamental to democracy, especially during times of elections or voting, and should that be wrongly influenced, it could cause significant damage to democratic societies.

As with all research, however, we acknowledge this study has three limitations. First, we are novice data scientists who are developing a novel model. Future research could apply to the proposed model to more complex scenarios to assess its utility. For example, apply sentiment analysis to economic news to reveal fake business news leading to instability in the stock market. This is another way of value creation from the paper, as fake news stories can cause serious economic issues to the stakeholders involved. Second, we did not take into consideration to real-world context those who are susceptible to fake news. Future research could focus on contextual factors such as the role of national culture and its influence on polarization [18, 19]. Additionally, future studies could use fake news detection and prediction in the corporate world regarding tarnishing company brand and reputation, and fake news regarding various market impacting mergers and acquisitions that could lead to a global crash in stock markets.

To conclude, by focusing on the differences in vocabulary used, this study advances understanding of filter bubbles and echo chambers, which can contribute to the polarization of societies, rather than inclusive societies.

7 References

1. Allcott, H., Mentzkow, M.: Social Media and Fake News in the 2016 Election, *Journal of Economic Perspectives*, 31(2), 211-236 (2017).
2. Xu, J., Li, T., Abdelzaher, Ji, H., Szymanski, K., Dellaverson, J.: The paradox of information access: on modeling social-media-induced polarization, (2021).
3. Tandoc, E., Z. Wei Lim and R. Ling., R.: Defining "Fake News",*Digital Journalism*, 6(2), 137-153 (2017).
4. Horne, B., Adali., S.: This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News, *Proceedings of the Intternational AAAI Conference on Web and Social Media*, 11(1), (2017).
5. Spohr, D.: Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, *34*(3), 150-160 (2017).
6. Brummette, J., DiStaso, M., Vafeiadis, M., Messner, M.: Read all about it: The politicization of "fake news" on Twitter. *Journalism & Mass Communication Quarterly*, *95*(2), 497-517 (2018)
7. Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, A., Nielsen, R.: Reuters Institute digital news report 2017. Available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf (Retrieved on 22nd Jan 2021).

8. Kim, Y., Kim, Y.: Incivility on Facebook and political polarization: The mediating role of seeking further comments and negative emotion. *Computers in Human Behavior*, *99*, 219-227 (2019).
9. Fisher, C.: What is meant by 'trust'in news media?. In *Trust in media and journalism* (pp. 19-38). Springer VS, Wiesbaden. (2018).
10. Sikder, O., Smith, R. E., Vivo, P., & Livan, G.: A minimalistic model of bias, polarization and misinformation in social networks. *Scientific Reports*, *10*(1), 1-11 (2020).
11. Guess, A., Nyhan, B., & Reifler, J.: Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council*, *9*(3), 4 (2018).
12. Jaramillo-Dent, D., Pérez-Rodríguez, M. A.: #MigrantCaravan: The border wall and the establishment of otherness on Instagram. *New Media & Society*, 23(1), 121-141 (2021).
13. Said, EW.: Covering Islam: How the Media and the Experts Determine How We See the Rest of the World. London: Random House (2018).
14. Laato, S., Islam, A. N., Islam, M. N., & Whelan, E.: What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems*, *29*(3), 288-305 (2021).
15. Usman, M: Python for NLP: Creating TF-IDF Model from Scratch. *StackAbuse.com (2021)*
16. Thankda, A: What is Logistic Regression? A Beginner's Guide. *CareerFoundry.com (2020)*
17. Brugnoli, E., Cinelli, M., Quattrociocchi, W., & Scala, A. (2019). Recursive patterns in online echo chambers. *Scientific Reports*, *9*(1), 1-18.
18. Gillespie, T., Boczkowski, P. J., & Foot, K. A. (Eds.). (2014). *Media technologies: Essays on communication, materiality, and society*. MIT Press.
19. Gupta, M., Esmaeilzadeh, P., Uz, I., & Tennant, V. M. (2019). The effects of national cultural values on individuals' intention to participate in peer-to-peer sharing economy. *Journal of Business Research*, 97, 20-29.
20. A. Navlani, "Datacamp," 28 December 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/decision-tree-classification-python. [Accessed May 2021].
21. R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data," Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining., vol. 1, pp. 29-39, 2000.
22. datascience-pm, "Datascience," 2020. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/. [Accessed 03 2021].
23. sv-europe, "Crisp dm methodology," 2021. [Online]. Available: https://www.sv-europe.com/crisp-dm-methodology/. [Accessed 03 2021].
24. R. Pupale, "Towards Data Science," 2018. [Online]. Available: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989. [Accessed 7 4 2021].
25. A. Chakure, "Medium," 6 July 2019. [Online]. Available: https://medium.com/swlh/decision-tree-classification-de64fc4d5aac. [Accessed May 2021].
26. Francia, P: Free Media and Twitter in the 2016 Presidential Election: The Unconventional Campaign of Donald Trump. Social Science Computer Review. (2017)
27. Fowler, G., Twitter and Facebook warning labels aren't enough to save democracy. The-WashingtonPost.com. (2020)