

FAKE NEWS DETECTION

Submitted by:

CHITRAK Satyarthi

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to my mentor as well as the company for giving me the golden opportunity to do this wonderful project on the topic **FAKE NEWS DETECTION**, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them.

**INTRODUCTION**

* Business Problem Framing

News media has become a channel to pass on the information of what’s happening in the world to the people living. Often people perceive whatever conveyed in the news to be true. There were circumstances where even the news channels acknowledged that their news is not true as they wrote. But some news has a significant impact not only on the people or government but also on the economy. One news can shift the curves up and down depending on the emotions of people and political situation.

It is important to identify the fake news from the real true news. The problem has been taken over and resolved with the help of Natural Language Processing tools which help us identify fake or true news based on historical data. The news is now in safe hands!

* Conceptual Background of the Domain Problem

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

The sensationalism of not-so-accurate eye-catching and intriguing headlines aimed at retaining the attention of audiences to sell information has persisted all throughout the history of all kinds of information broadcast. On social networking websites, the reach and effects of information spread are however significantly amplified and occur at such a fast pace, that distorted, inaccurate, or false information acquires a tremendous potential to cause real impacts, within minutes, for millions of users.

* Review of Literature

As time flows, the amount of data, especially text data increases exponentially. Along with the data, our understanding of AI also increases and the computing power enables us to train very complex and large models faster. Fake news has been gathering a lot of attention worldwide recently. The effects can be political, economic, organizational, or even personal. This paper discusses the approach of natural language processing and machine learning in order to solve this problem. Use of bag-of-words, n-grams, count vectorizer has been made, TF-IDF, and trained the data on five classifiers to investigate which of them works well for this specific dataset of labelled news statements. The precision, recall and f1 scores help us determine which model works best.

* Motivation for the Problem Undertaken

The widespread problem of fake news is very difficult to tackle in today’s digital world where there are thousands of information sharing platforms through which fake news or misinformation may propagate. It has become a greater issue because of the advancements in AI which brings along artificial bots that may be used to create and spread fake news. The situation is dire because many people believe anything they read on the internet and the ones who are amateur or are new to the digital technology may be easily fooled. A similar problem is fraud that may happen due to spam or malicious emails and messages. So, it is compelling enough acknowledge this problem take on this challenge to control the rates of crime, political unrest, grief, and thwart the attempts of spreading fake news. Text, or natural language, is one form which is difficult to process simply because of various linguistic features and styles like sarcasm, metaphors, etc. Moreover, there are thousands of spoken languages and every language has its own grammar, script and syntax. Natural language processing is a branch of artificial intelligence and it encompasses techniques that can utilize text, create models and produce predictions. The aim of this work is to create a system or model that can use the data of past news reports and predict the chances of a news report being fake or not.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Train test is the best way to get the solution of these kind of problems as that is the easiest and the efficient way to solve this problem.

* Data Sources and their formats

<bound method DataFrame.info of Unnamed: 0 id headline \

0 0 9653 Ethics Questions Dogged Agriculture Nominee as...

1 1 10041 U.S. Must Dig Deep to Stop Argentina’s Lionel ...

2 2 19113 Cotton to House: ’Do Not Walk the Plank and Vo...

3 3 6868 Paul LePage, Besieged Maine Governor, Sends Co...

4 4 7596 A Digital 9/11 If Trump Wins

... ... ... ...

20795 20795 5671 NaN

20796 20796 14831 Albert Pike and the European Migrant Crisis

20797 20797 18142 Dakota Access Caught Infiltrating Protests to ...

20798 20798 12139 How to Stretch the Summer Solstice - The New Y...

20799 20799 15660 Emory University to Pay for ’100 Percent’ of U...

written\_by \

0 Eric Lipton and Steve Eder

1 David Waldstein

2 Pam Key

3 Jess Bidgood

4 Finian Cunningham

... ...

20795 NeverSurrender

20796 Rixon Stewart

20797 Eddy Lavine

20798 Alison S. Cohn

20799 Tom Ciccotta

news label

0 WASHINGTON — In Sonny Perdue’s telling, Geo... 0

1 HOUSTON — Venezuela had a plan. It was a ta... 0

2 Sunday on ABC’s “This Week,” while discussing ... 0

3 AUGUSTA, Me. — The beleaguered Republican g... 0

4 Finian Cunningham has written extensively on... 1

... ... ...

20795 No, you'll be a dog licking of the vomit of yo... 1

20796 By Rixon Stewart on November 5, 2016 Rixon Ste... 1

20797 posted by Eddie You know the Dakota Access Pip... 1

20798 It’s officially summer, and the Society Boutiq... 0

20799 Emory University in Atlanta, Georgia, has anno... 0

* Data Preprocessing Done

round((df.isnull().sum().sort\_values(ascending=False) \* 100 /len(df)),2)

* Data Inputs- Logic- Output Relationships

written\_by 9.41

headline 2.68

news 0.19

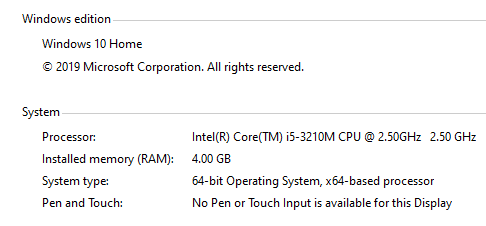
label 0.00

id 0.00

Unnamed: 0 0.00

dtype: float64

* Hardware and Software Requirements and Tools Used



**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

CountVectorizer

MultinomialNB

* Testing of Identified Approaches (Algorithms)

train\_test\_split

classification\_report

* Run and Evaluate selected models

PassiveAggressiveClassifier

**CONCLUSION**

* Key Findings and Conclusions of the Study

precision recall f1-score support

FAKE 0.92 0.89 0.90 2072

REAL 0.87 0.89 0.88 1585

accuracy 0.89 3657

macro avg 0.89 0.89 0.89 3657

weighted avg 0.89 0.89 0.89 3657

* Learning Outcomes of the Study in respect of Data Science

Alpha: 0.0, Score : 0.8851517637407711

Alpha: 0.05, Score : 0.8925348646431501

Alpha: 0.1, Score : 0.8919879682800109

Alpha: 0.15000000000000002, Score : 0.8922614164615805

Alpha: 0.2, Score : 0.8919879682800109

Alpha: 0.25, Score : 0.8922614164615805

Alpha: 0.30000000000000004, Score : 0.8917145200984413

Alpha: 0.35000000000000003, Score : 0.8914410719168717

Alpha: 0.4, Score : 0.8911676237353021

Alpha: 0.45, Score : 0.8925348646431501

Alpha: 0.5, Score : 0.8928083128247197

Alpha: 0.55, Score : 0.8930817610062893

Alpha: 0.6000000000000001, Score : 0.8933552091878589

Alpha: 0.65, Score : 0.8933552091878589

Alpha: 0.7000000000000001, Score : 0.8936286573694285

Alpha: 0.75, Score : 0.8933552091878589

Alpha: 0.8, Score : 0.8930817610062893

Alpha: 0.8500000000000001, Score : 0.8933552091878589

Alpha: 0.9, Score : 0.8936286573694285

Alpha: 0.9500000000000001, Score : 0.8930817610062893

* Limitations of this work and Scope for Future Work

There will always the scope of improvement in any study. As far as the current study is concerned it gives all the insights with the solution of the problem statement given according to my knowledge.