Name: sakthivel P

NM ID: au513521106030

Title: fake news detection using NLP

1. Introduction

We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic.

Every news that we consume is not real. If you listen to fake news it means you are collecting the wrong information from the world which can affect society because a person's views or thoughts can change after consuming fake news which the user perceives to be true.

Since all the news we encounter in our day-to-day life is not authentic, how do we categorize if the news is fake or real?

In this article, we will focus on text-based news and try to build a model that will help us to identify if a piece of given news is fake or real.

Before moving to the practical things let's get aware of few terminologies.

2. Terminologies

2.1 Fake News

A sort of sensationalist reporting, counterfeit news embodies bits of information that might be lies and is, for the most part, spread through web-based media and other online media.

This is regularly done to further or force certain kinds of thoughts or for false promotion of products and is frequently accomplished with political plans.

Such news things may contain bogus and additionally misrepresented cases and may wind up being virtualized by calculations, and clients may wind up in a channel bubble.

2.2 Tfidf Vectorizer

TF (Term Frequency): In the document, words are present so many times that is called term frequency. In this section, if you get the largest values it means that word is present so many times with respect to other words. When you get word is parts of speech word that means the document is a very nice match.

IDF (Inverse Document Frequency): in a single document, words are present so many times, but also available so many times in another document also which is not relevant. IDF is a proportion of how critical a term is In the whole corpus.

Collection of word Documents will convert into the matrix which contains TF-IDF features using TfidfVectorizer.

2. Project

To get the accurately classified collection of news as real or fake we have to build a machine learning model.

To deals with the detection of fake or real news, we will develop the project in python with the help of 'sklearn', we will use 'TfidfVectorizer' in our news data which we will gather from online media.

After the first step is done, we will initialize the classifier, transform and fit the model. In the end, we will calculate the performance of the model using the appropriate performance matrix/matrices. Once will calculate the performance matrices we will be able to see how well our model performs.

3. CONCLUSION

The passive-aggressive classifier performed the best here and gave an accuracy of 93.12%.

We can print a confusion matrix to gain insight into the number of false and true negatives and positives

Fake news detection techniques can be divided into those based on style and those based on content, or fact-checking. Too often it is assumed that bad style (bad spelling, bad punctuation, limited vocabulary, using terms of abuse, ungrammaticality, etc.) is a safe indicator of fake news.

More than ever, this is a case where the machine's opinion must be backed up by clear and fully verifiable indications for the basis of its decision, in terms of the facts checked and the authority by which the truth of each fact was determined.

Collecting the data once isn't going to cut it given how quickly information spreads in today's connected world and the number of articles being churned out.

I hope you might find this helpful. You can comment down in the comment sections for any queries.