**Detecting Fake News with Natural Language Processing**

**Submitted by :**

**B.NAVEEN KUMAR REDDY**

**au723921104006**

**reddybnaveenkumarreddy@gmail.com**

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.

Do you trust all the news you consume from online media?

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.



**Ready to start your data science journey?**

**Master 23+ tools & learn with guided projects with the Certified AI & ML BlackBelt Plus Program**

[Enroll Now](https://blackbelt.analyticsvidhya.com/plus?utm_source=blog_india&utm_medium=desktop_banner_between_articles&utm_campaign=01-Oct-2023%7C%7C&utm_content=tools)

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.

3. 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.

The practical implementation of these tools is very simple and will be explained step by step in this article.

Let’s start.

**3.1 Data Analysis**

Here I will explain the dataset.

In this python project, we have used the CSV dataset. The dataset contains 7796 rows and 4 columns.

This dataset has four columns,

1. **title**: this represents the title of the news.
2. **author**: this represents the name of the author who has written the news.
3. **text**: this column has the news itself.
4. **label**: this is a binary column representing if the news is fake (1) or real (0).

The dataset is open-sourced and can be found [here](https://www.kaggle.com/c/fake-news/data?select=train.csv).

**Data Preprocessing**

In data processing, we will focus on the text column on this data which actually contains the news part. We will modify this text column to extract more information to make the model more predictable. To extract information from the text column, we will use a library, which we know by the name of ‘**nltk’**.

Here we will use functionalities ofthe **‘nltk**‘ library named Removing Stopwords, Tokenization, and Lemmatization. So we will see these functionalities one by one with these three examples. Hope you will have a better understanding of extracting information from the text column after this.

**Removing Stopwords:-**

These are the words that are used in any language used to connect words or used to declare the tense of sentences. This means that if we use these words in any sentence they do not add much meaning to the context of the sentence so even after removing the stopwords we can understand the context.

**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