Fake News Detection

Introduction:

The main idea is detecting the fake news.

The objective is model can classify the news whether it’s fake or not.

MethodologyDiagram

Description automatically generated

1. Tokenization splitting on space.
2. Text normalization using Case folding.
3. Text cleansing: -
4. Fix Contraction (I’ll -> I will).
5. Remove Stop Words.
6. normalize hyphenated words.
7. Remove suffixes (nlp.com), punctuation, emojis, URLs, HTML tags.
8. Replace user handles with (\_USER\_).
9. Replace currency symbols with ($ => dollar, etc..).

4. Text Lemmatization: Specify POS of each word to get the exact Lemma.

5. Feature Extraction: Using TF-IDF vectorizer.

6. Model: Feed the model with the two columns (title and text) as X.

Data Set Summary:

The dataset: News.csv which is mostly about the USA politics.

The dataset consists of three columns: -

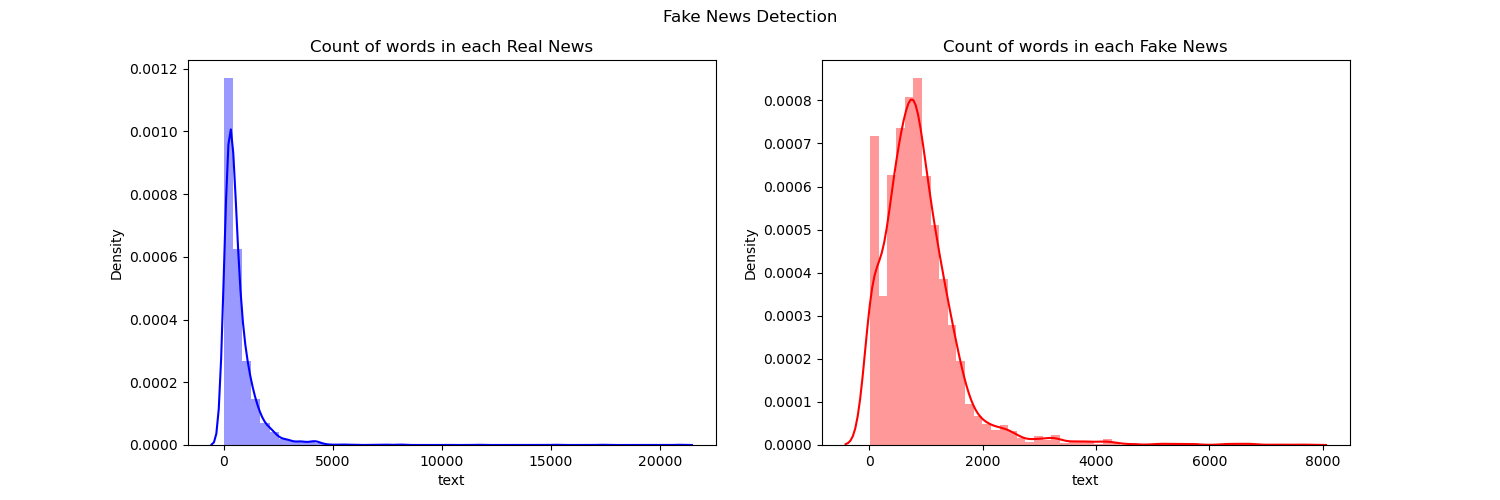
1. Title column: contains text data it defines the title of news.
2. Text column: contains text data it defines the content of news.
3. Label column: it defines the label of news whether it is Fake or Real.

This graph shows the number of fake news compared to the number of real news and we can observe clearly that data is almost balanced:

Chart, bar chart

Description automatically generated

This graph shows the count of words in each fake news compared to the count of words in each real news:



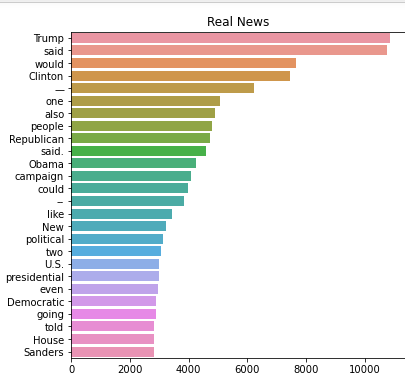
Most common words used in the dataset: -

A screenshot of a computer

Description automatically generated with medium confidence

The common words per class: -

Chart

Description automatically generated 

Results:

1. Passive aggressive classifier: -

Graphical user interface, application

Description automatically generated

1. Naïve bayes: -

Graphical user interface, application

Description automatically generated

3- RBF SVM: -

Graphical user interface, application

Description automatically generated

1. Logistic regression classifier: -

Graphical user interface

Description automatically generated with medium confidence