Social media analytics

Review - 2

CSE4069

FAKE NEWS ANALYSIS
REVIEW-2

TEAM-21
B. PHANINDRA SAI-19MIA1065
O. NAGA SAI KUMAR-19MIA1071
M. JAY KUMAR PATEL-19MIA1032
T. SIVA NIKHIL-19MIA1086

Introduction

News medium has become a channel to pass on the information of what's happening on 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 have a significant impact not only to the people or government but also 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 the historical data. The news are now in safe hands !

Problem statement

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 user

Objectives

1. Our sole objective is to classify the news from the dataset to fake or true news.
2. Extensive EDA of news
3. Selecting and building a powerful model for classification

About dataset

train.csv: A full training dataset with the following attributes:

- **id**: unique id for a news article
- **title**: the title of a news article
- **author**: author of the news article
- **text**: the text of the article; could be incomplete
- **label**: a label that marks the article as potentially unreliable
  - 1: unreliable
  - 0: reliable

test.csv: A testing training dataset with all the same attributes at train.csv without the label.

submit.csv: A sample submission that you can

Dataset link: https://www.kaggle.com/competitions/fake-news/data

Workflow

1. Introduction
2. Preprocessing and cleaning
3. Story generation and visualization from news
4. Stemming & vectoring
5. Model building: fake news classifier
6. Deep learning LSTM
7. Conclusion

Methodology

Preprocessing and Cleaning

We have to perform certain preprocessing steps before performing EDA and giving the data to the model.Let's begin with creating the output column

Text Processing

This is an important phase for any text analysis application.There will be many unuseful content in the news which can be an obstacle when feeding to a machine learning model.Unless we remove them the machine learning model doesn't work efficiently. Lets go step by step.

Story Generation and Visualization from news

In this section we will complete do exploratory data analysis on news such as ngram analysis and understand which are all the words,context which are most likely found in fake news

Methodology

Stemming & Vectorizing

Stemming the reviews

Stemming is a method of deriving root word from the inflected word. Here we extract the reviews and convert the words in reviews to its root word. for example,

- Going->go
- Finally->fina

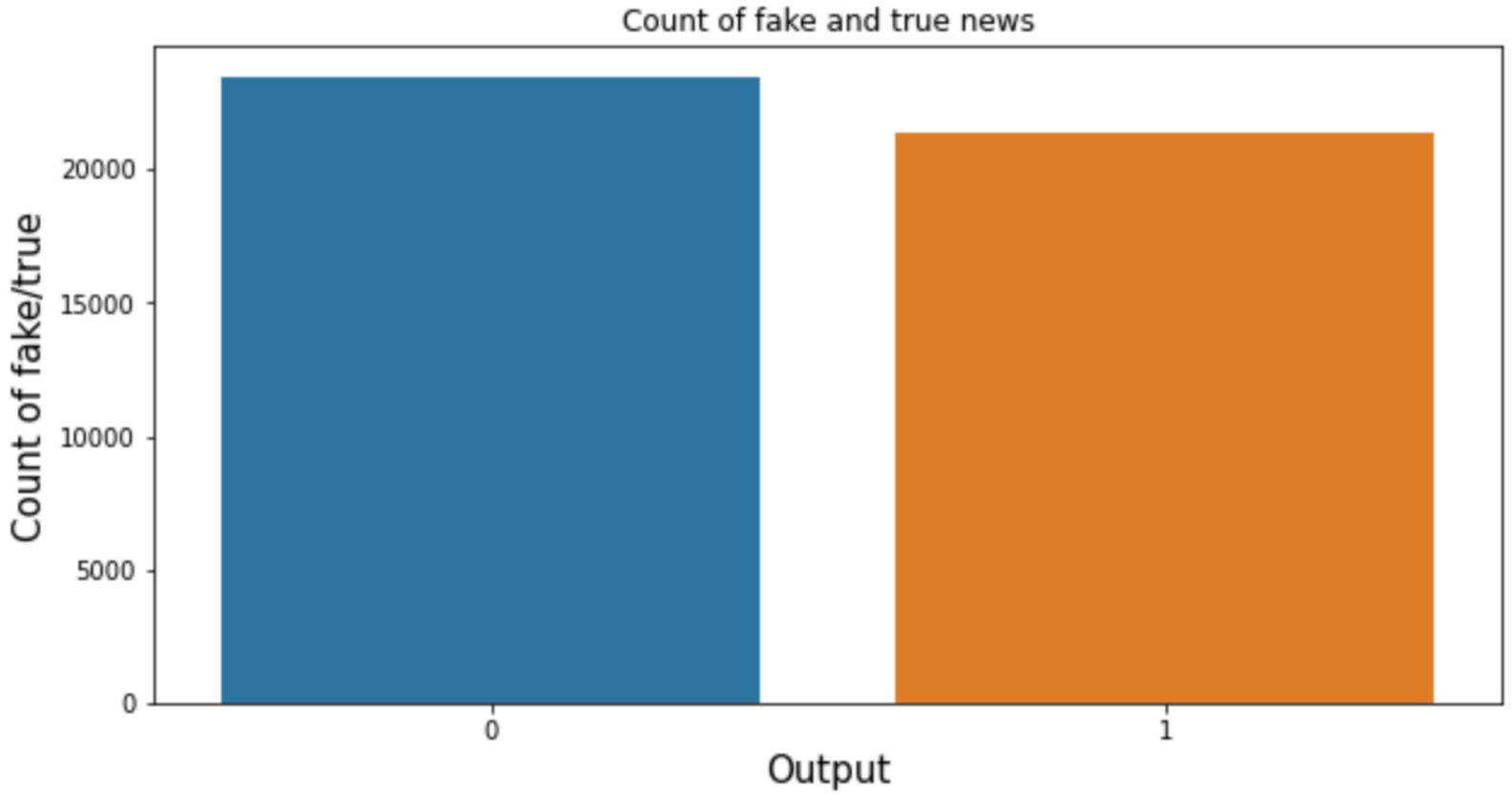
If you notice, the root words doesn't need to carry a semantic meaning. There is another technique knows as Lemmatization where it converts the words into root words which has a semantic meaning. Simce it takes time. I'm using stemming

Deep learning-LSTM

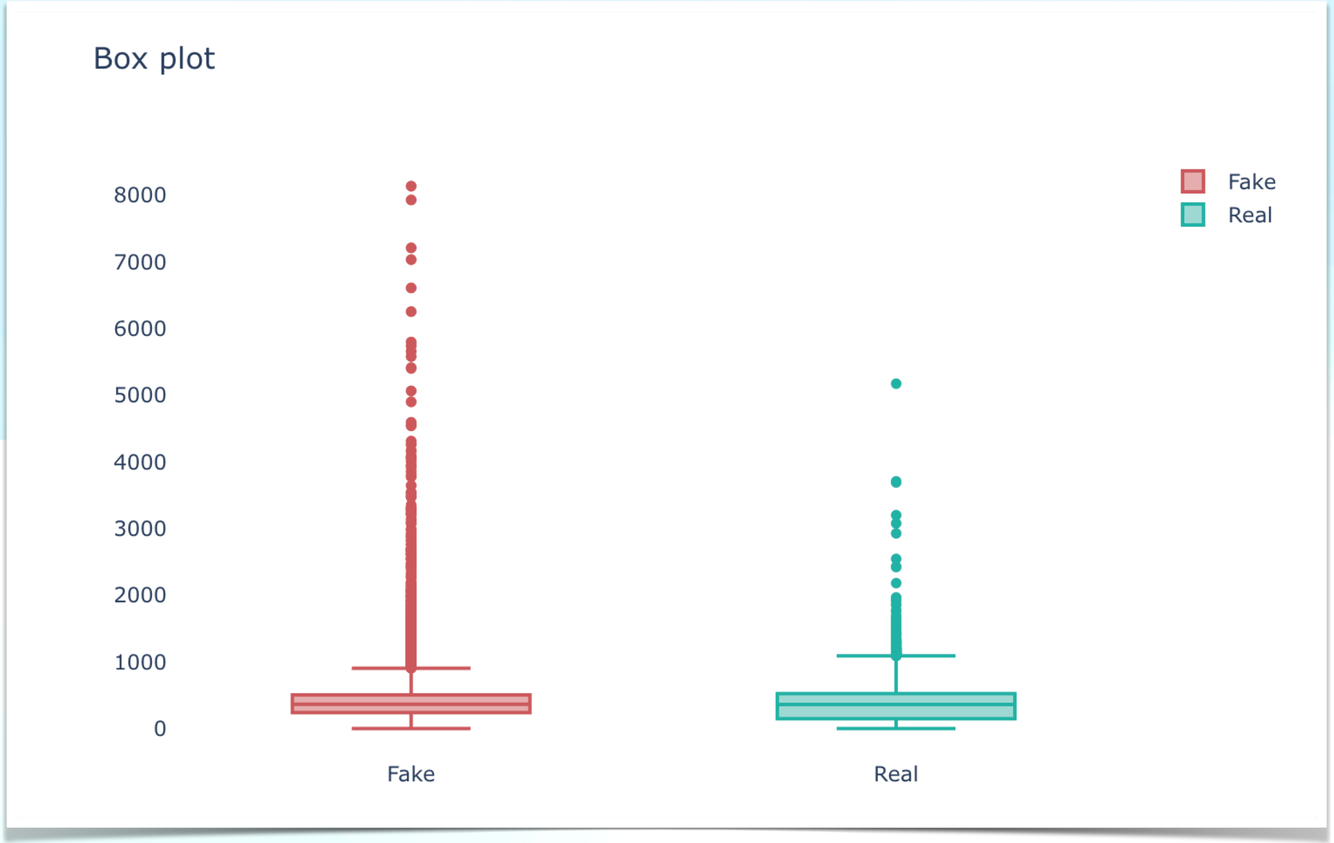
Here in this part we use neural network to predict whether the given news is fake or not.

We aren't gonna use normal neural network like ANN to clasify but LSTM(long short term memory) which helps in containing sequence information.Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. This is a behavior required in complex problem domains like machine translation, speech recognition, and more.

Visualization of data

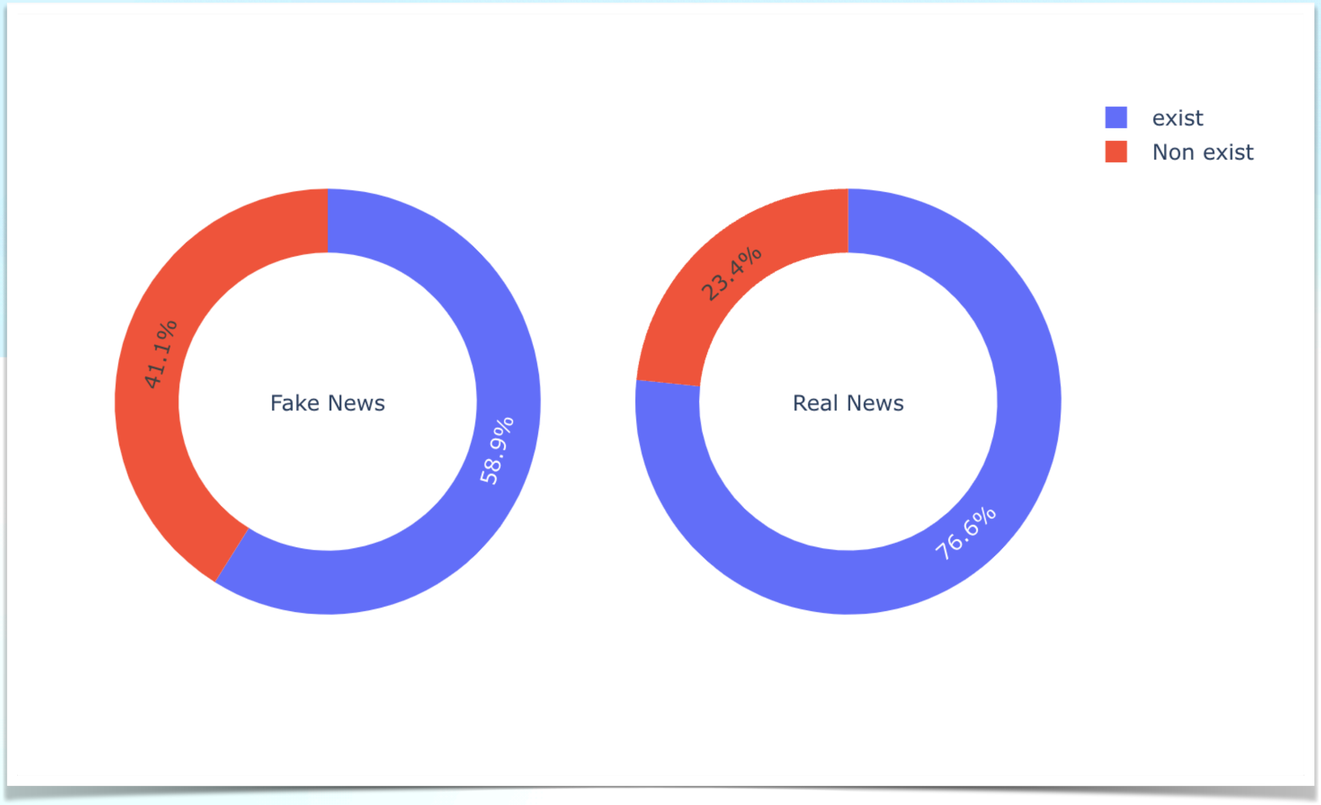


Visualization of data



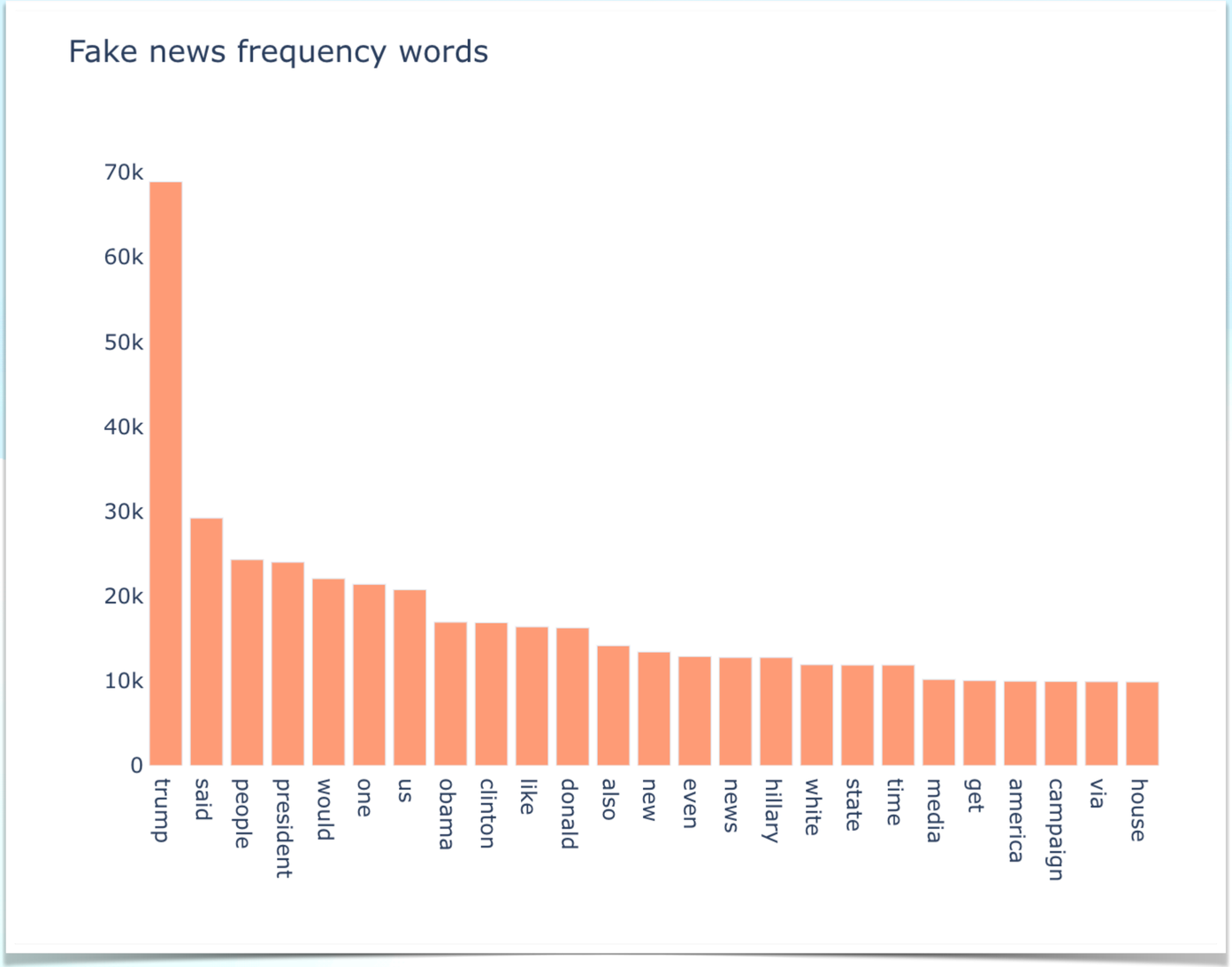
- Fake news in general has a lot more tokens than real ones, which is kind of weird assuming real news has a tendency to bring details about events to inform the reader, however as noted the fact that fakes news is a mix of twitters and news this may justify the fact that they have more words.
- But would that only demonstrate the bias of the dataset? because this can be resolved given that we would only work with fake news that was less than or equal to the news with more tokens, so this problem could be managed hypothetically.
- Perhaps something simple is not being observed, the presence of duplicate news !!! If this is something frequent in the data when the dataset is split, we would have samples both in the training and in the test, so it is possible to present the same samples for the model

Visualization of data

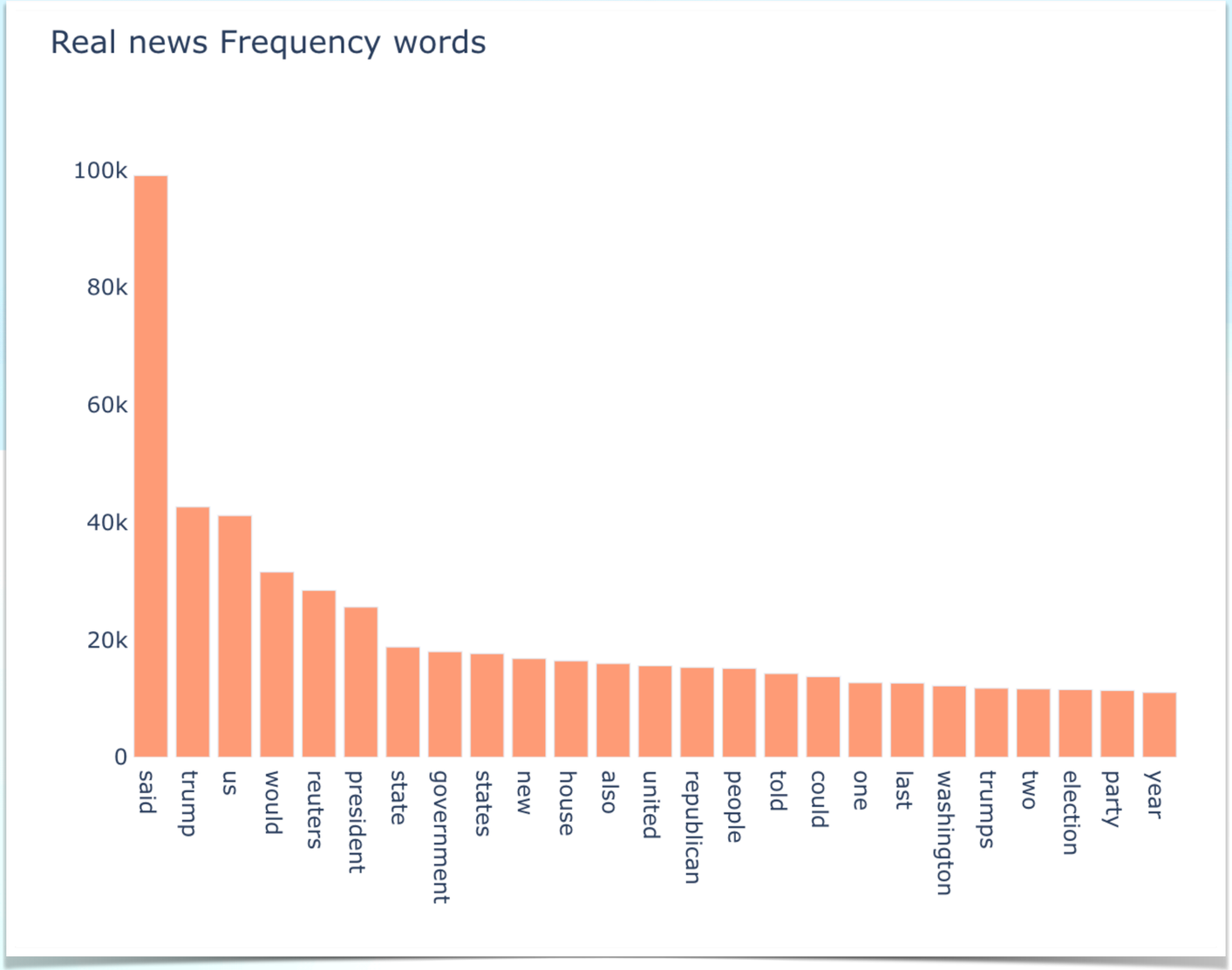


- More than 70% of the words in the news fakes were not found in the dictionary used for verification, it is important to make it clear that it is not a perfect dictionary but that it already brings this section that many words are really misspelled
- but so far everything that has been done has been in the data without any preprocessing, so let's apply a preprocessing that clears some characters and normalizes the text so that we can compare again

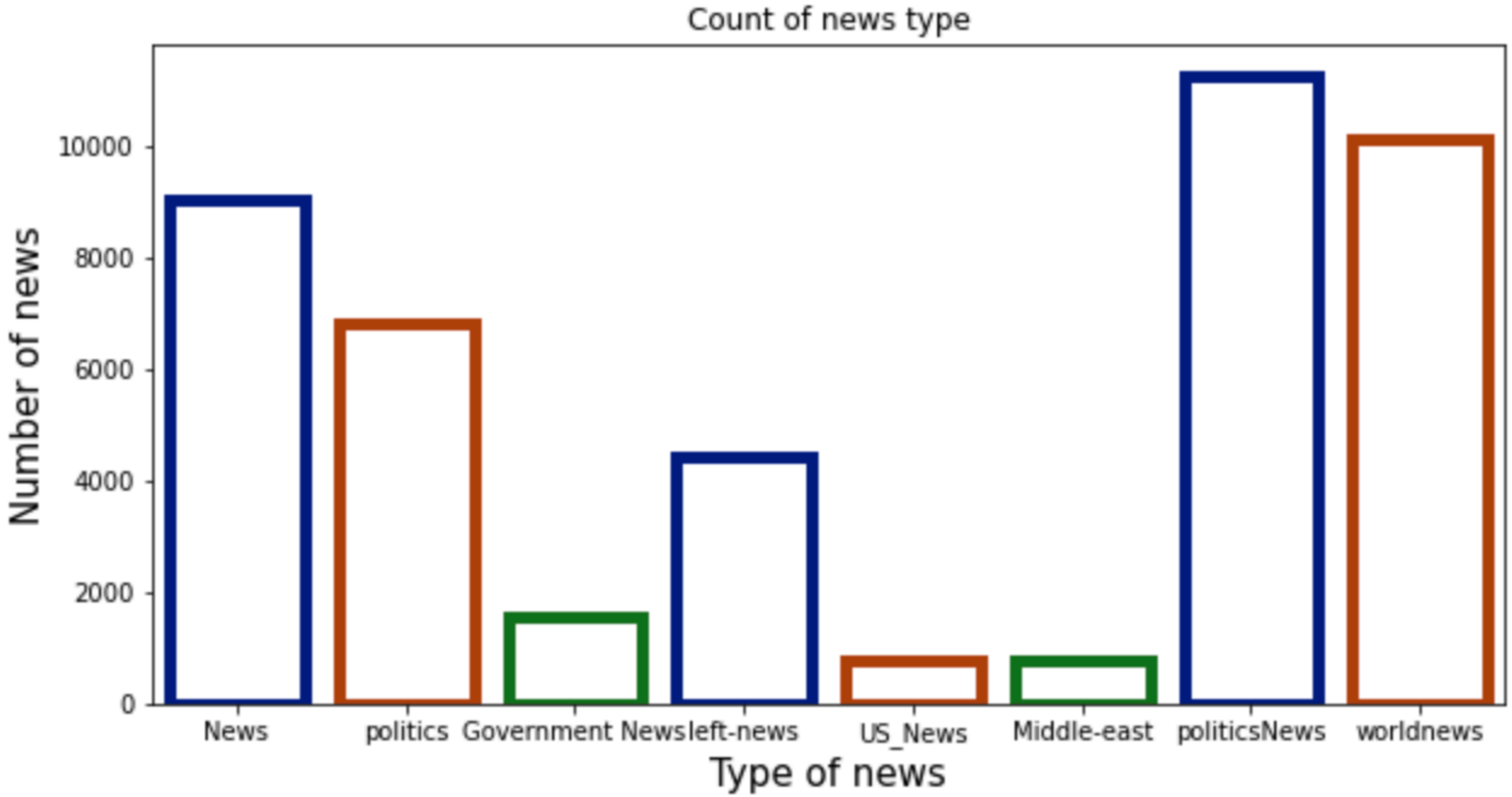
Visualization of data



Visualization of data



Visualization of data



Insights

- Our dataset has more political news than any other news followed by world news
- We have some repeated class names which expresses same meaning such as news,politics,government news etc which is similar to the alternative