Introduction

The task assigned for this coursework is to create a Machine Learning algorithm that correctly determines whether a tweet is real or fake.

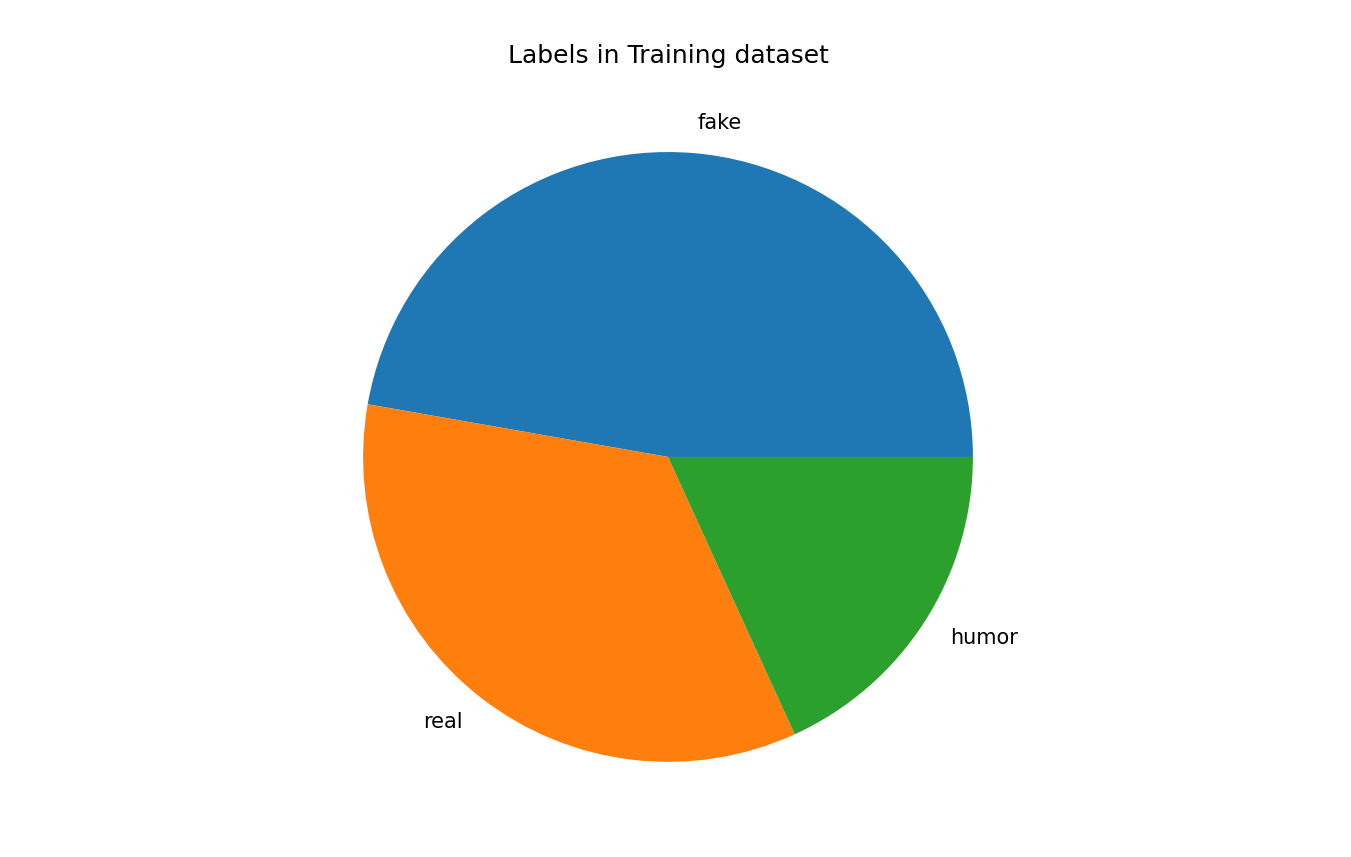
The data provided for this challenge are the tweetId, tweetText, userId, username, timestamp and the label, which gives the classification (true or fake). The number of tweets provided for training the algorithm is 14483. There is a separate set of data intended for testing purposes and evaluating how our algorithm performs.

The main characteristics of the tweets I chose to analyze are:

* Its label
* Language
* Punctuation (exclamation marks, question mark, ellipsis and other)
* Tweet content (emojis, hashtags, location, URL links)
* Word content (verbs, nouns, adjectives)
* Mean word and character counts
* Subjectivity
* Polarity

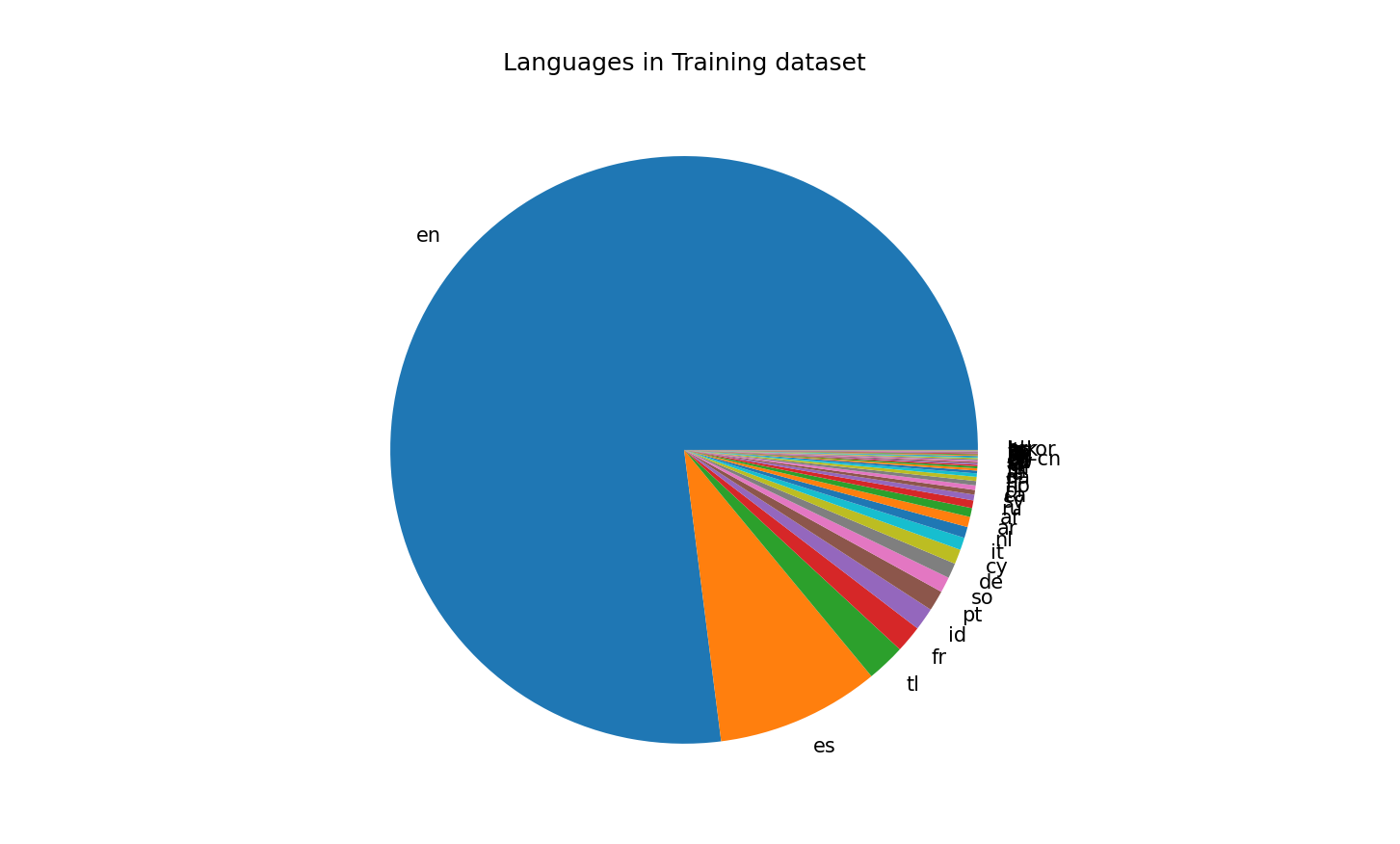
Labels

The label composition for all the tweets in our training dataset are divided to three groups: real, fake and humor. The real tweets are 34.5% from all the tweets in the training dataset, leaving the fake and humor ones approximately 65.5% of the dataset. This is not ideal, as an approach of always guessing a tweet to not be real would make a 65.5.% success rate.

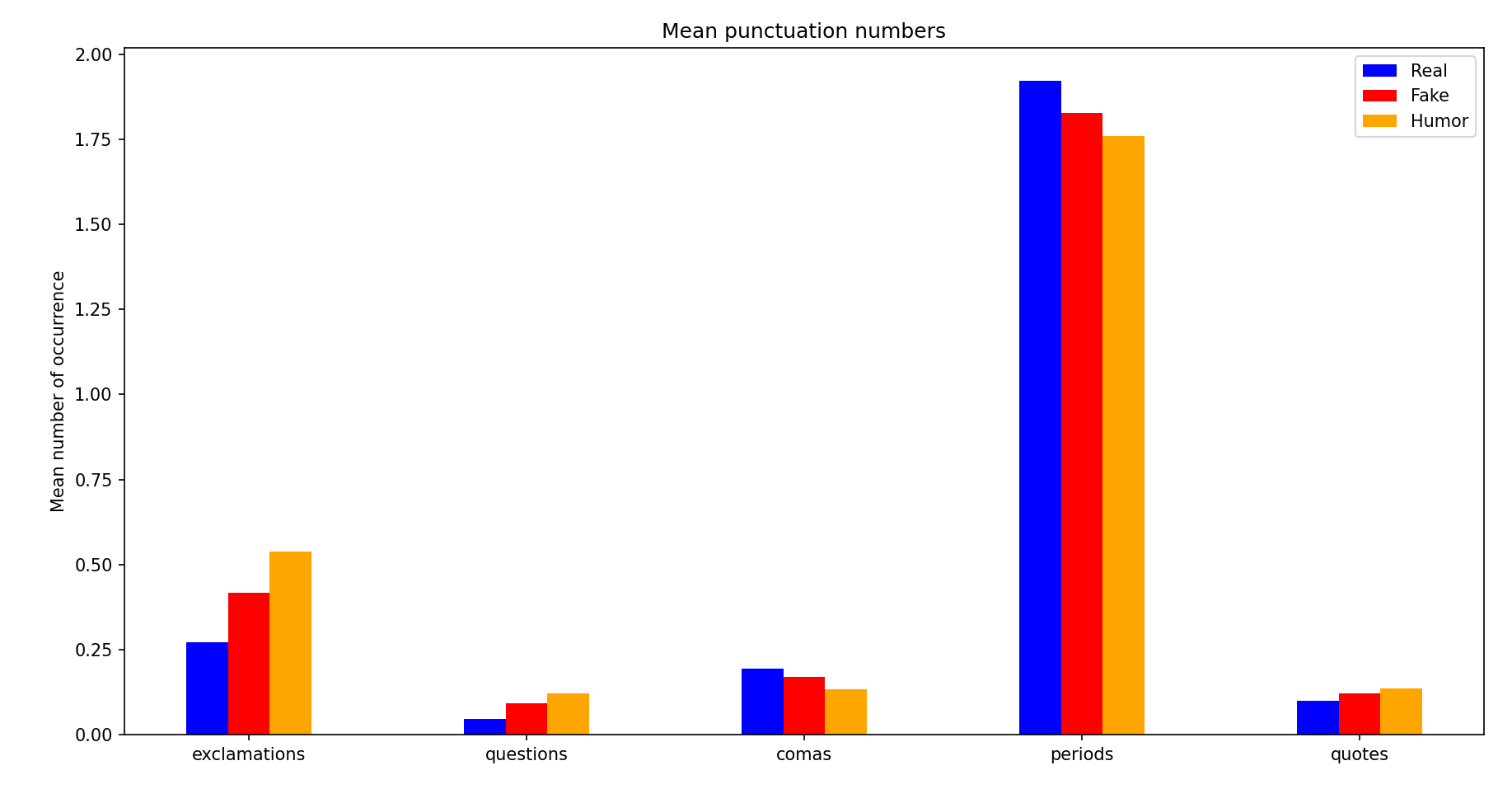


Language

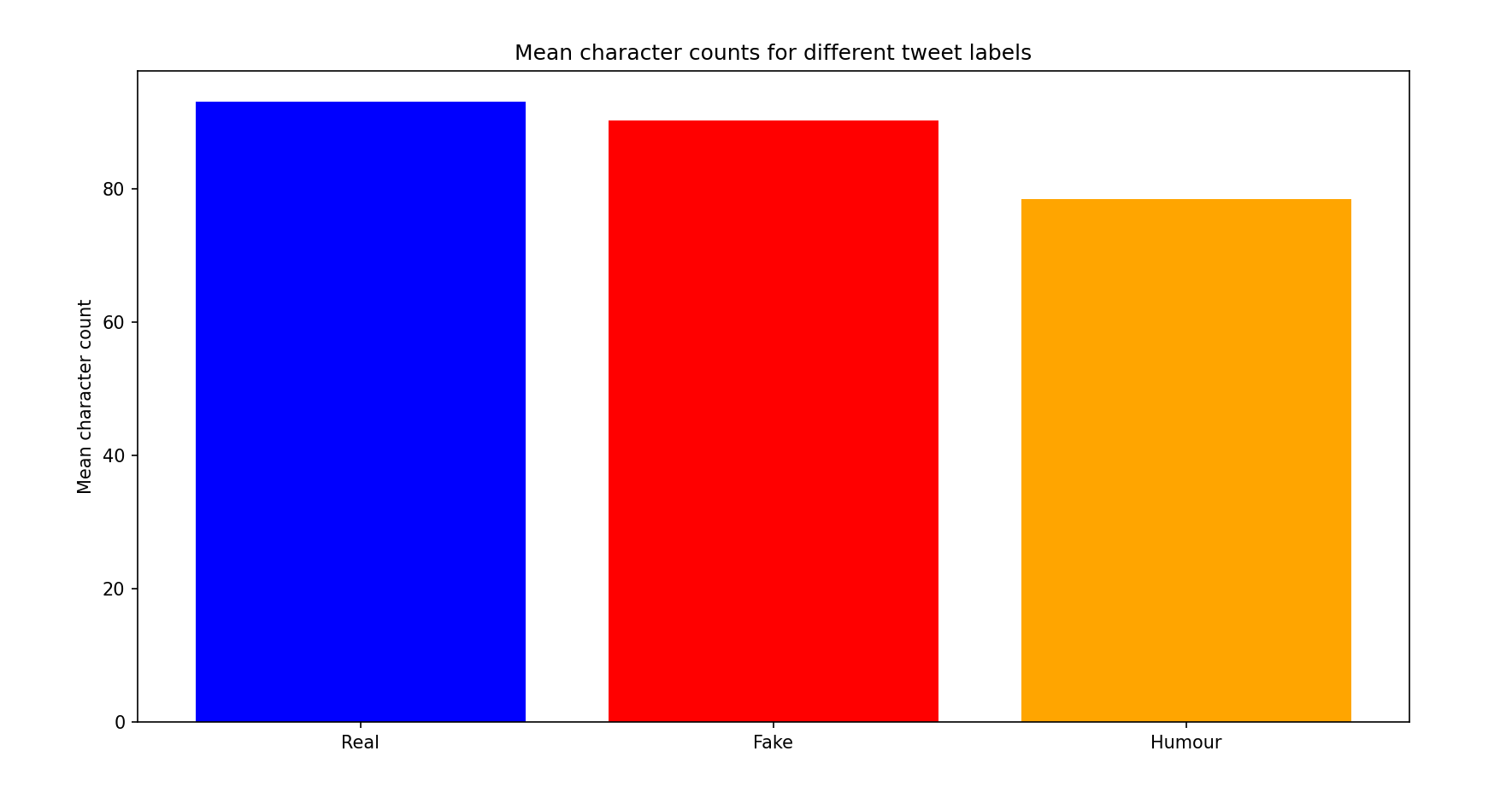
The most used languages in the tweets are English (77%), Spanish (9%), Tagalog (2%), French (1.5%) and other languages, which are no more than 1 % from all the languages. In order to analyze all of the tweets in the most equal way, it was better to translate almost all of them in English, as it is the most commonly used language (77% use).

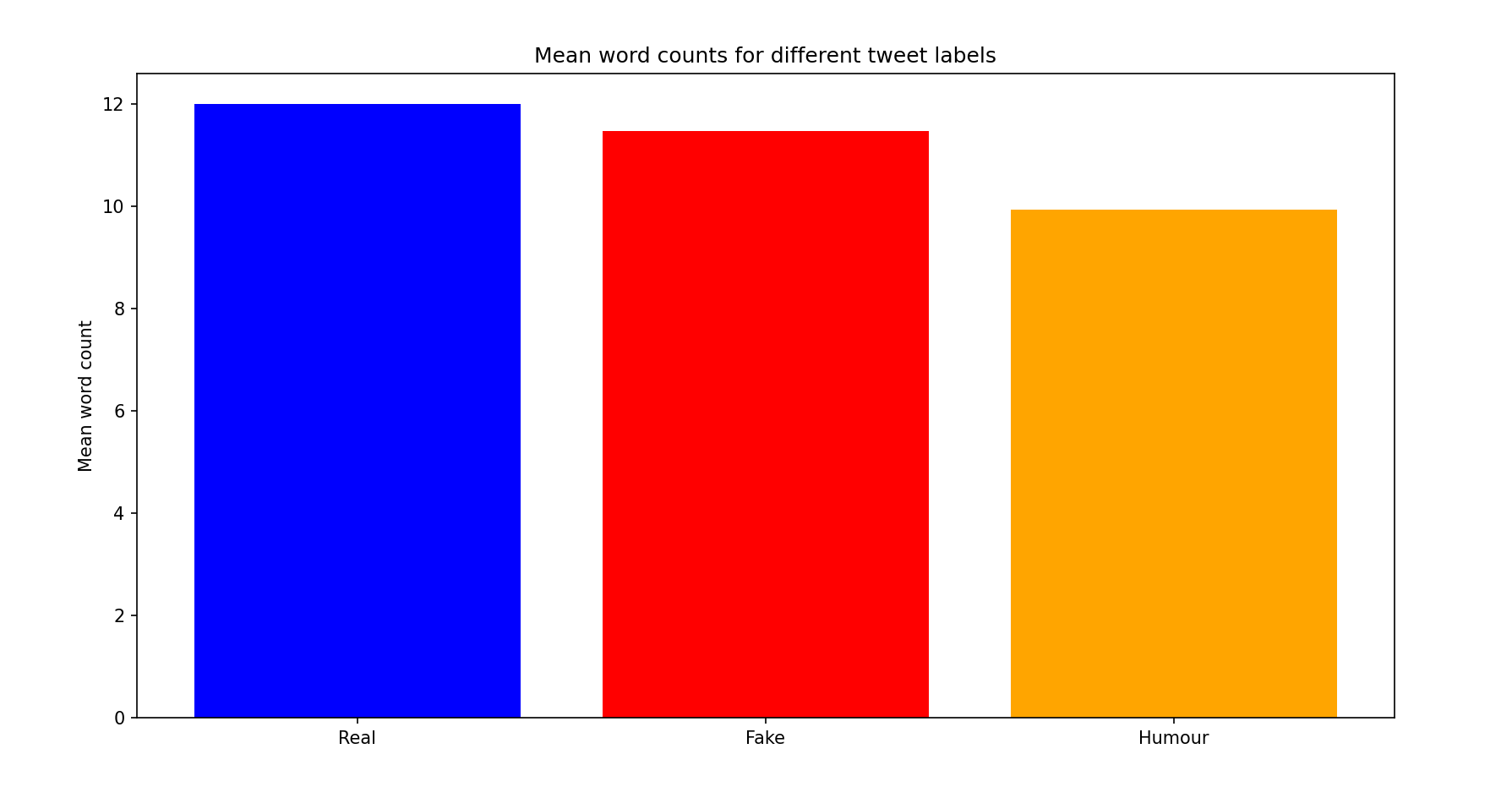


Punctuation



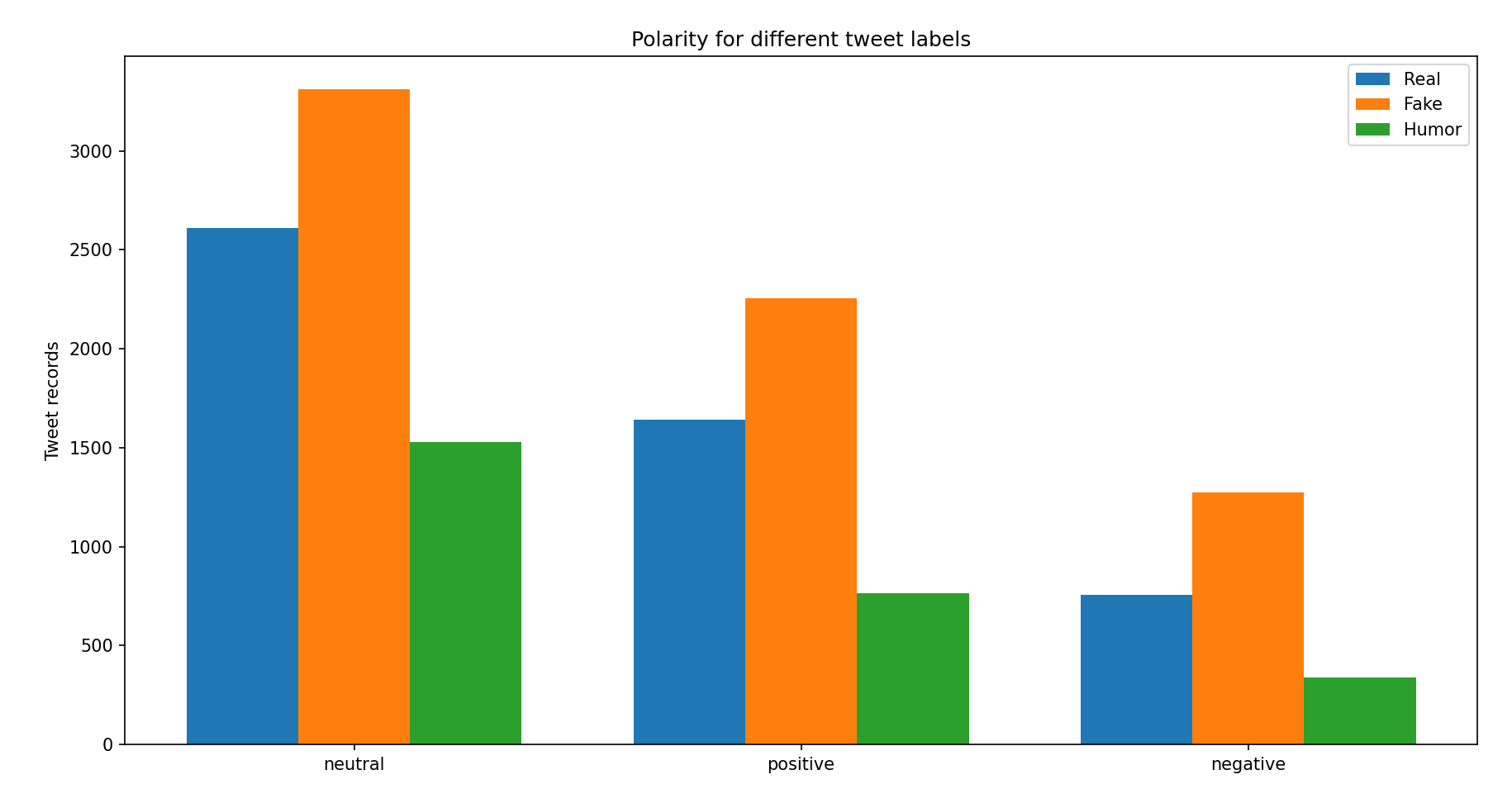
Word and character counts





Polarity

The polarity grade is given based on a numeric result taken from TextBlob, a library for processing text data. It gives an adequate evaluation on the opinion expressed by the text. If the grade is less than 0, we classify it as a negative statement. If equal to 0 – neutral and if higher than 0 – positive. From the data below we can see that the polarity evaluation does not play a great role in determining the label of the tweet, as the ratio is quite similar in all the polarity classifications.



Subjectivity

Like the polarity chart, a score is given to evaluate how subjective or objective a tweet is. This time the score is from 0 to 1, so 0.5 would come as neutral. From the chart below we can conclude that similarly, to the polarity case, there are big differences between the number of tweet records for real and unreal tweets.

