This is a rough outline for report – 5 - 10 page (10 max), bag of words =/ n-grams.

<https://simplemaps.com/data/world-cities> -> city names

# Introduction

Our task is to train a Machine Learning (ML) algorithm to correctly classify which tweets are fake and which tweets are real using the tweet’s content as a basis for classification. This task essentially is a binary classification problem, the two classes being fake (humour label is classified as fake) and real. Additionally, I will not be using any data relating to the author of the tweet, like how many followers or their biography description. I am only going to be using the following information to train the machine learning algorithm: tweetId, tweetText, userId, username, timestamp and the label (ground truth). I do have image id’s, but I will not be using them. From that, I will generate more features to feed into the ML algorithm.

# Literature review

There have been multiple approaches before attempting to classify tweets into fake tweets and real tweets. A lot of these approaches use additional data such as how long the author of the tweet has been with twitter <cite>, how many friends the author has <cite>, or how many tweets has been posted by the author <cite>.

Fortunately, these methodologies included what features were extracted and used from the tweet’s content; of which we have access to. These include: length of tweet<cite>, does tweet have URL? <cite>, does tweet have ! or ? <cite>, does tweet have a geographical location?<cite>, does tweet have #words?<cite>, etc . This is key as it shows what features provide useful descriptive power to be able to classify fake and real tweets since these same features were also used in classification.

Previous attempts have used a wide range of ML algorithms and it seems these ML algorithms have been used a lot: support vector machines<cite>, logistic regression<cite>, decision trees<cite>, random forest<cite>, naïve bayes<cite>, k-nearest neighbours<cite>, neural networks <cite>, etc. This is vital information since it shows the effectiveness of each technique and thus must be a reason to use them in this classification task.

The methodologies also mentioned important methods in feature extraction on the tweet’s content via Parts of Speech tagging (POS)<cite>, n-grams <cite> and also Bag of Words approach<cite>.

Talk about:

* Features selected / used
* Feature extraction
* ML algorithm of choice

# Pre-processing

# Data visualisation

Firstly, I had a quick look at the training set.txt file and saw that there were 14483 records and that not all the tweets were English. I used the langdetect library <cite> to determine what language the tweet was written in to find out.

Chart, pie chart

Description automatically generated

As you can see, English, Spanish, Tagalog, French and Indonesian are the most common languages in the dataset in that order. English tweets make the majority of the training dataset (76.93% or 11142 tweets) which I deemed to be large enough for a training dataset. I believed that translating each tweet into English and performing feature extraction on (probably) broken English would yield very little, especially when all the other languages combined only make up 23.07% (or 3341 tweets). Interestingly, langdetect library had trouble identifying tweet id “262974742716370944” due to poor spelling and very little tweet content to analyse. This tweet was counted as error as manually checking what language a tweet is in is far too exhaustive, probably is not the case in the real world.

Chart, pie chart

Description automatically generatedOn this English training dataset, I looked at the ground truth labels to check for any dataset bias.

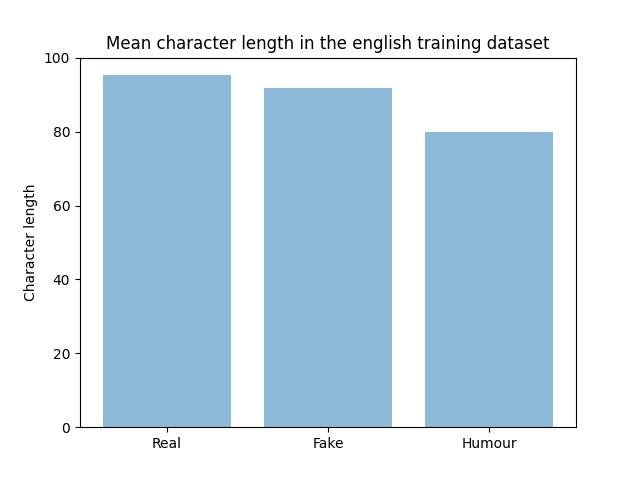
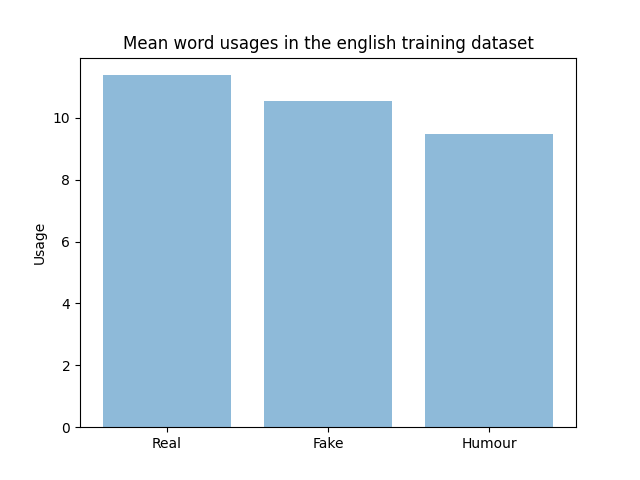
There are 4151 real labels (37.26%), 5275 fake labels and 1715 humour labels. In total, there are 6990 fake labels (62.74%) which is includes the Humour label. This is very important because now we have a large dataset bias i.e. a dumb algorithm that would guess fake for every post would be 62.74% accurate, for this dataset. This means for our training cycles; we would need to even out the dataset labels to avoid this bias.

I then analysed the tweet contents for punctuation, emojis, URLs, hashtags, mentions for each label in the English training dataset.

Chart, bar chart

Description automatically generated

It seems that real tweets use less exclamation marks and question marks however uses more hashtags and mentions when compared to fake and humorous tweets. Emoji , URL and ellipsis usage don’t seem to have much descriptive power.

Further analysis on the length of tweets.

It seems to be that real tweets tend to be longer in both word and character count in comparison to fake and humorous tweets. These features seem to have adequate descriptive power in classifying real from fake tweets.

# Data Analysis

# Algorithm design

Probably going to be using simple and traditional ML Algorithms

SVM

Decision trees -> random forest is probs the best

Lmao try linear or logistic for jokes

# Evaluation

# Conclusion

# References

Zotero gottem baby