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

# 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

# Data analysis and visualisation

# 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