**Munchkin**

Ajitesh Bansal(group leader)

Ishwa Shah

Weizhi Zhang

## Problem and Goal

Let's ask ourselves a simple question, do we trust all the news present on social media? How can we detect fake news from real news? It's a tricky question. So what is fake news? Fake news is a piece of news that is not true and is deliberately designed to mislead people. It is usually spread via social media or other online platforms.

Fake news can be politically driven to give advantages or disadvantages to a political party. Such news items may contain false and exaggerated claims and, because of certain algorithms, trap users in a filter bubble. It is creating different issues, from sarcastic articles to fabricated news and planned government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge complications in our society.

So, our goal is to detect fake news in order to prevent the spreading of misleading news stories that come from non-reputable sources. We are seeking to create a model that can predict whether a given news is fake or real. We can detect fake news using supervised machine learning methods. In the end, we are expecting to differentiate fake news from real ones.

## Formalization

This is a classification problem where we are classifying news as fake or real. We will be using two machine learning algorithms to try to solve the problem and then check which algorithm performed well. The **two** algorithms are as follows:

- ● *Multinomial* **Naive Bayes Classifier:** A Naive Bayes classifier is a probabilistic machine learning model that's used for the classification task. The Multinomial Naive Bayes classifier is suitable for classification with discrete features.
- ● **Passive Aggressive Classifier:** Passive-aggressive algorithms are a family of algorithms for large-scale learning. They are similar to the Perceptron in that they do not require a learning rate. However, contrary to the Perceptron, they include a regularization parameter.
  First we will create random training and testing subsets using the dataset that we will discuss in the next section. Next, we will create a matrix of token count from the text document. Following this, our goal will be to create a linear model that will be used to classify real news from fake news. Finally, we will calculate the model's accuracy by testing the model using the test data.
- ● Splitting a dataset is an important task whereby we can train our model and then test it on real data (in our case) to predict the accuracy of our model for real-world use. In this project, we will use 70% of the data for training and 30% for testing.
- ● Following this will be the feature selection. In this phase, we will look at the following:

    - ○ Stop words such as “and,” “the,” and “him” are assumed to be uninformative in representing the content of a text. Therefore, they can be removed to avoid being interpreted as a signal for prediction. Similar words can be helpful for prediction in some cases, such as classifying writing style or personality.
    - ○ When creating the vocabulary, we will exclude terms with a document frequency that is strictly greater than the given threshold. If the parameter is a float, it represents a percentage of documents; otherwise, it means absolute counts. If the vocabulary is None, this parameter is ignored.
- ● Next, we will initialize and apply the classifiers to the training data and test it on the test data as well.

**Data**

In this project, we’ll get the data from two different sources:
First, we will use the news dataset from Kaggle.
The second dataset we’ll create ourselves using the News API. We will use this API to load some data and then append that data to the other dataset.

**Schedule**

| **Tasks** | **Dates of Completion** |
|---|---|
| 1. Collecting and preprocessing data | Mar 7, 2023 |
| 2. Implementation of Algorithm 1 | Mar 10, 2023 |
| 3. Implementation of Algorithm 2 | April 15, 2023 |
| 4. Evaluating and comparing algorithms | April 20, 2023 |
| 5. Writing report | April 20, 2023 |
| 6. Creating presentation slides | April 30, 2023 |