COMP40020: Human Language Technologies  
Assignment 1

* Investigating language styles from data on Reddit
* Build a corpus from one or more subreddits
* Pre-processing steps that should be used.
* Can you infer the subject of a subreddit from its most common words
* Common n grams capture popular jokes
* Text lexically diverse
* Around 2000 words

**Research Question:**

Is New Zealand really that much happier than Ireland? A sentiment and lexical analysis of titles of posts r/newzealand and r/ireland

According to the annual World Happiness Report for 2022 New Zealand comes in at position 10 while Ireland is placed 3 places lower in position 13 out of the 150 countries. This report coupled with the large amounts of young Irish people emigrating to countries like New Zealand and Australia (position 12) provides the rationale for the research question.

Both these subreddits are home to discussion about news, politics, culture, history and society.

Sentiment analysis is a natural language processing method to determine whether a piece of text is either positive, negative, or neutral. Lexical analysis is used to further break down the text and assess the mood.

New Zealand and Ireland are both English speaking countries which makes results easier to compare and contrast.

**Methods**

The PRAW python library is used to make a connection with the chosen subreddits. The code is similar to what is provided in the lab 3 notebook with some minor modifications.

Text

Description automatically generatedChart

Description automatically generated with medium confidenceFirst, analysis was completed on the headlines. The latest 900 titles were taken from both sub reddits. Each headline was appended to a python set to ensure no duplicates were recorded. Stop words were not removed as they can potentially sentiment bearing meanings. Both sets were then passed through a sentiment intensity analysis function. This takes in the headlines set and loops over each headline. For each headline a polarity score is calculated. This is a method that is part of the sentiment intensity analyser model from the NLTK library. For each headline scores of strength is attributed: positive, negative, and neutral. A compound score is also given which is the sum of the scores of each word in the headline. This measures the intensity of the sentiment (Keita, 2022). These values are stored in a dictionary. This is then updated to contain the original headline. This dictionary is then appended to an array. The array is then transformed into a pandas data frame. Using the function below both data frames are given a label initially set to 0. Then this is modified based on the value of the compound score. If greater than 0.2 it is labelled as 1 (positive) if less than -0.2 it labelled as -1 (negative) with the everything else staying as 0.

It is now possible to view those headlines that are classed as positive, negative, or neutral. For example the “negative” headlines for r/Ireland.

A picture containing text

Description automatically generated

Text

Description automatically generatedThe value counts for each category as raw values and as percentages are then shown.

The results are then visualised with a bar chart using the python library matplotlib library.

The headlines are then tokenised to look at the positive and negative headlines in more detail. A frequency plot is drawn for positive and negative word for both subreddits.

The above steps is then repeated for the top 900 posts. As is shown in the below picture below the method is changed to “top”.

This method is then repeated for all the comments in the top 10 posts for each sub reddit.

**Results**

The results are graphs and plots as follows.

**Discussion**

There are important caveats when carry out sentiment analysis on social media posts.

**Bibliography**