**COMP40020: Human Language Technologies  
Assignment 1**

**Gus Boothman**

**17350796**

**Research Question:** Is New Zealand really that much happier than Ireland? A sentiment and lexical analysis of titles of posts and user comments from 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. This out of the 150 countries. The report is based on surveys of citizens. economic, social, and personal factors are all covered (Helliwell, et al., 2022). This report coupled with the large amounts of young Irish people emigrating to countries like New Zealand provides the rationale for the research question.

Both the subreddits chosen are home to discussion about news, politics, culture, history, and society. Thus, giving a wide variety of opinions from different people in society. The anonymous nature of Reddit lends itself to more open and honest conversation compared to other sites like Twitter. R/ireland current subscriber number is 641,000. While r/newzealand sits at 390,000 subscribers. New Zealand and Ireland are both English speaking countries which makes results easier to compare and contrast as no translation is required.

Sentiment analysis used to categorise the posts by both sub reddits. This is a natural language processing method to determine whether a piece of text is either positive, negative, or neutral (Pand & Lee, 2008). Lexical analysis is used to further break down the text and to see if it is possible to decipher the mood from looking at common words and n-grams.

**Method**

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. NLTK, Pandas, and Matplotlib libraries are used for the sentiment and lexical analysis. Furthermore, code from Martin and Koufos (2020), Omonyi (2022) and Singh (2020) were all used for the sentiment analysis throughout the python notebook.

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 have 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 a sample of “negative” headlines for r/Ireland are show below:

A picture containing text

Description automatically generated

The value counts for each category as raw values and as percentages are then shown.

Text

Description automatically generated

Table

Description automatically generatedStatics on all scores are displayed using pandas describe method:

Chart, histogram

Description automatically generatedThe results are then visualised using the python library matplotlib library. Bar charts are used to visualise the amount of headlines placed in either neutral, negative or positive. Compound scores are then visualised using histograms and box plots:

Text

Description automatically generatedThe headlines are then tokenised to look at the positive and negative headlines in more detail. As you can see a regex method is used to treat ‘New Zealand’ as one token. The stop words are also removed for this stage.

A frequency plot is drawn for the top 20 positive and negative word for both subreddits.

Text

Description automatically generatedThe above steps is then repeated for the top 900 posts and also all the comments in the top 10 posts. As is shown in the below image the code is adapted from lab 3 for extracting comments. The comments are again stored in a set:

**Results**

In the results section, only a sample of the charts and graphs are shown. All results can be seen in the notebook. Furthermore, all results are also taken into account in the discussion section.

**With regard the latest 900 post titles:**

**Chart, bar chart

Description automatically generated**Bar chart:

r/ireland has a mean compound value of -0.006848 with a standard deviation of 0.358226.

r/newzealand has a mean compound value of -0.010231 with a standard deviation of 0.342465.

Histograms:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

**Looking at the top 900 post titles:**

Chart, bar chart

Description automatically generatedBar chart:

r/ireland has a mean compound value of 0.064415 with a standard deviation of 0.370326.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedr/newzealand has a mean compound value of 0.075785 with a standard deviation of 0.364761.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedLooking at frequency plots:

**Looking at all comments from top 10 posts:**

**Discussion**

From looking at the results

Setting the threshold for labelling the comments or headlines as either positive, negative, or neutral was also challenging.

Negativity tends to be more prevalent on social media.

There are important caveats when carry out sentiment analysis and lexical analysis on social media posts. As reddit is largely anonymous, there is no way to looks at proportion of data regarding gender, age, or socio-economic status. It is also not possible to obtain data regarding the connection users have to either subreddit. Information regarding how many users are actually living in the respective countries is unknown. R/ireland is also nearly twice as big as r/ireland in terms of subscriber numbers. Culturally how both nations use language is not controlled. These limitations greatly decrease the generalisability of the research.

Future research could look into other countries on the list. People from Finland who are ranked as number in the world happiness report, social media use could be compared to those lower on the list. Translation techniques would have to be used.

In conclusion, Python provides powerful libraries for natural language processing.

**Total Word Count =**

**Bibliography**