**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. The report is based on surveys of citizens from 150 countries. 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.

The data used is the subreddits of the two countries. These subreddits chosen as their purpose is to facilitate discussion about news, politics, culture, history, and society. This gives a wide range of opinions from different people in society about various subjects. A benefit of using data from Reddit is that its anonymous nature lends itself to more open and honest conversation compared to other sites like Twitter. Both subreddits are active, r/ireland current subscriber number is 641,000. While r/newzealand sits at 390,000 subscribers. Furthermore 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 titles of posts and the comments 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. Code from Martin and Koufos (2020), Omonyi (2022) and Singh (2020) were all used and modified for the sentiment analysis throughout the python notebook. NLTK, Pandas, and Matplotlib libraries are used for the sentiment and lexical analysis.

Text

Description automatically generatedFirst, analysis was completed on the headlines of posts. This is the text users title their posts. 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 (Vallantin, 2019). 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.

Chart

Description automatically generated with medium confidenceUsing the code 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 everything else staying as 0 (neutral).

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 shown 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 generatedStatistics 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 labelled as 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.

Graphical user interface, text, application

Description automatically generatedA frequency plot is drawn for the top 20 positive and negative word for both subreddits *(see Results section*). Then the top 10 bigrams and trigrams are shown using these two functions:

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 python notebook. Furthermore, all results are 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.

Values counts: r/ireland = 54.5% neutral. 22.9% positive, 22.5% negative

r/newzealand = 56.6% neutral. 21.1% positive, 22.2% negative

Histograms:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Examples of bigrams and trigrams:

**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.

Values counts: r/ireland = 53.7% neutral. 29.5% positive, 16.7% negative.

r/newzealand = 51.2% neutral. 31.3% positive, 17.4% negative

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedLooking at frequency plots:

Examples of bigrams and trigrams:

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

Chart, bar chart

Description automatically generatedBar chart:

r/ireland has a mean compound value of 0.036248 with a standard deviation of 0.503418.

r/newzealand has a mean compound value of 0.069173 with a standard deviation of 0.529840.

Values counts: r/ireland = 31.3% neutral. 37.0% positive, 31.6% negative.

r/newzealand = 30.1% neutral. 40.3% positive, 29.5% negative

Histograms:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedFrequency plots:

Examples of bigrams and trigrams:

**Discussion**

From looking at the results, the mean compound scores for r/ireland are all lower than r/newzealand except when looking at the most recent 900 post titles. This is clear when viewing the histograms. From looking at the value counts, the most recent titles for r/ireland has a slightly higher proportion of negative titles but also slightly higher proportion of positive titles. Similar results are seen when looking at the top 900 titles, with r/newzealand have more negative titles by less than one percent. Interestingly, for both subreddits over half of the top posts have neutral scores with a greater likelihood of being positive rather than negative. With regards to the comments, r/ireland has slightly more negative comments by one percent.

Looking at the frequency of common words provided some further insight. For example, you can see the prevalence of covid discourse in r/newzealand which was labelled as negative. While in r/ireland words like “rent”, “deposits”, and “house” were prevalent among the negative comments. This analysis quickly how news events and societal topics are viewed and the rate at which they are discussed.

There are important caveats when carry out sentiment analysis and lexical analysis on social media posts with this method. As reddit is largely anonymous, there is no way to looks at proportion of data regarding gender, age, or socio-economic status. These are often important figures when trying to generalise results to a wider population. It is also not possible to obtain data regarding the connection users have to either subreddit. For example, 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 taken into account. For example, sentiment analysis can struggle to detect irony and sarcasm. While negativity levels are relatively similar, there is potential for a negativity bias to be at play here. It has been shown that people are more likely to interact with negative news studies (Kätsyri, et al., 2016). These limitations greatly decrease the generalisability of the results. Furthermore, analysing the top posts gives a better overview of the outlook of the sub reddit compared to looking new posts at an inconsistent time point. This makes the method less replicable.

Setting the threshold for labelling the comments or headlines as either positive, negative, or neutral was also challenging. The threshold was set to greater than 0.2 was considered positive and less than -0.2 was considered negative, with everything else considered neutral. This was set by looking at samples of those labelled positive and negative and adjusting the threshold based on observations. A method using a predefined validation set would have potentially yielded more accurate results.

Future research should involve using more data. For example, it was computationally intensive to extract the comments for an entire post. Thus, comments were only scraped from the top 10 posts. Assessing more comments would yield more generalisable results. Furthermore, other countries on the world happiness report rankings could be explored. For example, people from Finland who are ranked as number one in the world happiness report, social media use could be compared to those lower on the list to see how the report is reflecting social media use. Translation techniques would have to be used in this case. Fine tuning the sentiment analysis function is also an area of importance.

In conclusion, Python provides powerful libraries for natural language processing. It is impressive how with few lines of code a huge amount of text can be processed. I understand why organisations like political parties and advertisers are using such techniques and technologies (Kaczmirek, et al., 2013). Finally, while there is only a slight difference, it appears that Irish reddit users are more likely to post negative views compared to New Zealand redditors. This corroborates with the world happiness report.

**Total Word Count =**

**Bibliography**