Seasonal Sentiment Analysis on r/vancouver

By: Matthias Zelleke, Erik Kho

Introduction

This final project aimed to analyze the sentiment of Reddit submissions and comments using machine learning (ML) and statistical tests (namely the chi-squared test). We were trying to analyze *whether the season/time of year affects the sentiment of Reddit submissions and comments in the r/vancouver subreddit*.

We classified the text into one of 2 categories, positive or negative sentiment. In our case, sentiment refers to whether the Reddit submissions/comments would be perceived as having a positive or negative tone.

To obtain our current question, we first considered the project idea “What makes a submission good/bad? Perhaps time of day or sentiment or readability score affect how a post is perceived?”. We were intrigued by the concept of sentiment analysis and detecting emotions using ML. Moreover, since we knew how Vancouver’s weather patterns varied throughout the year (sunny in the summer, usually cloudy otherwise) as well as how the outside weather could influence people’s moods, we pondered whether those factors could affect the expression of people’s moods, with regards to their sentiment, through Reddit.

Within the r/vancouver subreddit, we use the chi-squared test and its p-value to determine whether the time of year that a submission is posted affects the likelihood of that submission having a positive or negative sentiment, (and the same with comments). The submissions and comments we are investigating are from 2021 to 2023.

Our hypotheses are:

Null Hypothesis: The season/time of the year does not affect the sentiment of Reddit submissions and comments in the r/vancouver subreddit.
Alternate Hypothesis: The season/time of the year does affect the sentiment of Reddit submissions and comments in the r/vancouver subreddit.

Data Collection and Cleaning

We obtained all of our data from SFU’s Reddit Cluster using a Spark engine within Python. First, we collected submissions/comments from only the r/vancouver subreddit. Secondly, we filtered our data to contain submissions and comments between 2021 and 2023. Moreover, we only sampled 10% of the submissions and comments from this period to control the data’s size.

Thirdly, we removed empty and deleted submissions and comments since they did not give us meaningful information and essentially represented missing values.

Lastly, although we used a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer to convert the text data in the submissions/comments to numerical data for our ML model, it also performed some important data cleaning steps, including: converting all text into lowercase, removing common words like “and”, “the”, and “of”, and tokenizing the input text by word (splitting the text by word before applying the ML model).

Data Analysis

Machine Learning

After collecting and cleaning the required data, we needed to predict the sentiment of our obtained text. For this, we used an ML pipeline consisting of a TF-IDF vectorizer and a support-vector machine (specifically a LinearSVC), both from Python’s scikit-learn library.

The TF-IDF Vectorizer was used to convert string data (reddit submissions & comments) into numerical values that can be inputted into an ML model. This was achieved by calculating each word’s frequency within each submission/comment and the entire collection of submissions/comments from 2021 to 2023, and then performing arithmetic operations to obtain a matrix that represented each word’s contribution to each submission/comment.

We chose to use support-vector machines (SVMs) for a couple reasons. Firstly, SVMs were designed initially for solving binary text classification problems (positive or negative sentiment), making them a good starting point for our analyses. Secondly, by using kernel-penalized methods in the SVM (LinearSVC in scikit-learn), we were able to make predictions that generalized across datasets from different sources (Chidambaram & Srinivasagan, 2018). This was important because our training data came from Twitter, and we made predictions on Reddit.

Statistics

In order to decide whether the season affected the likelihood of a submission/comment having a positive or negative sentiment, we used a chi-squared test. Specifically, we used the chi2_contingency function from the scipy.stats library to do this test and calculate the appropriate counts and p-values. We used the chi-squared test because the season that a submission/comment is in, as well as whether it has a positive or negative sentiment are categorical variables.

When comparing sentiment across seasons, we split up the months of the year in **two main ways**, either into 4 groups each representing one season (e.g. Dec-Feb for winter, Mar-May for spring, etc…), **or** into 2 groups with one representing the “summer” (Jun-Sep) and the other representing “not summer” (Oct-May).

Before applying the chi-squared test, we used the matplotlib library to plot the number of positive and negative comments for every month to see which kind is more frequent.

If p < 0.05 we reject our null hypothesis. With the Reddit data we collected from the cluster, we categorized the submissions to different seasons and filtered them by negative and positive sentiments by setting a value to both sentiments. Then, we count each season's positive and negative sentiments and place them in a contingency table. A chi-squared test can be done on the contingency table. We do a similar test to see whether the summer season affects the proportion of positive and negative comments.

Results:

With the chi-squared test, we received the p-value for 4 season submissions is 0.21 and the p-value for comments is $2.045 * 10^{-5}$. The p-value for the equivalent 2 season analysis on the submissions is 0.69 and the p-value of the comments is 0.01. Here, we assume June-September as summer because at some point in September 2022 was really warm, and in recent years, September has been warmer than previous years. These p-values indicate that we failed to reject our null hypothesis for submissions, since p>0.05 for both 4-season and 2-season based splitting of the months. However, we reject the null hypothesis for comments since p<0.05 (for both splitting methods) and accept the alternate hypothesis.

Data Visualization:

4-season splitting method

Even though we obtained a significant p-value when comparing comments using this method, this significant p-value doesn't specify which seasons had more positive (or negative) sentiment comments than would be expected under the null hypothesis. Thus, to answer such a question, Figure 1 below shows the relationship between the actual and expected counts of positive comments, and Figure 2 below does the same for negative comments.

Figure 1

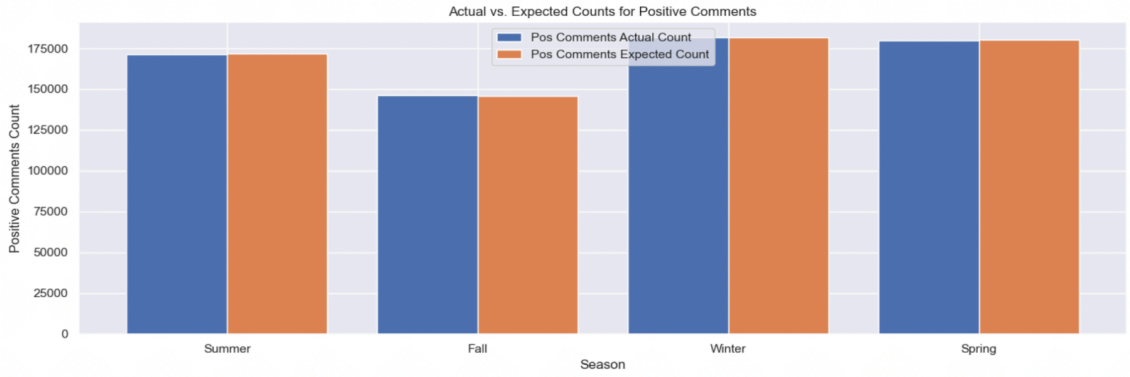
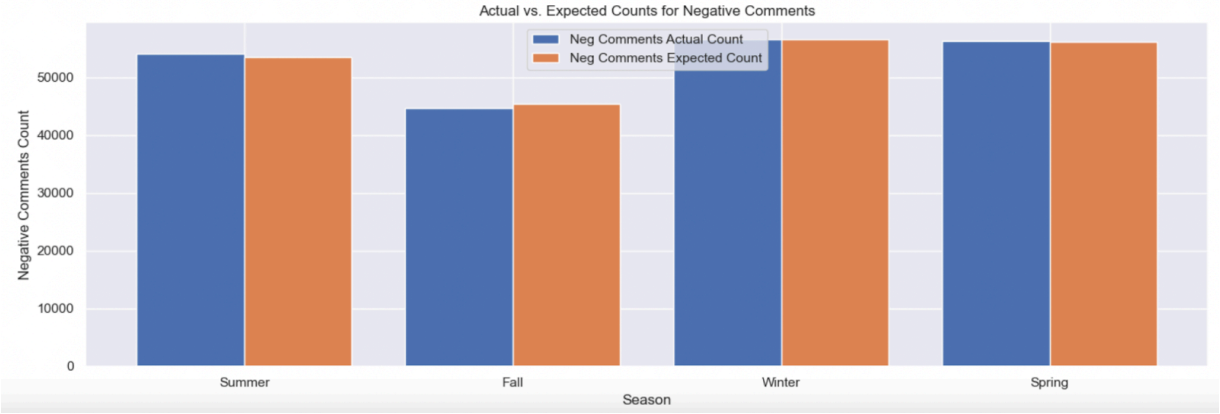


Figure 2



Although the differences are almost indistinguishable (due to the large number of submissions), there are slightly more negative comments in the summer than would be expected under the null hypothesis, and slightly fewer in the fall. There are virtually no differences between the two counts for both the winter and the spring. Those discrepancies in the summer and fall are thus likely the main contributors to the significant p-value obtained when grouping with the 4-season method and examining comments.

2-season splitting method

Since this splitting method grouped 8 months together, the total count of comments was significantly higher than under the 4-season method. Therefore, we will use a table (rather than a plot) to display the differences between actual and expected counts. Table 1 below shows the actual and expected counts of positive (and negative) comments in each season, where the expected counts are what would be obtained under the assumption of the null hypothesis.

Table 1 (Note: Expected counts are rounded to the nearest whole number)

Season	(Actual | Expected) Positive count	(Actual | Expected) Negative count
Summer	220,320 | 220,810	69,342 | 68,852
Non-summer	458,568 | 458,078	142,344 | 142,834

Table 1 shows that there are in fact more negative sentiment comments in the summer than would be expected under the null hypothesis, and that there are more positive sentiment comments during the non-summer months than would be expected.

Limitations

At first, we implemented the Mann-Whitney test to get the p-value. However, we decided to remove it and change it to a chi-squared test because there are only 2 possible values (negative

and positive) and Mann-Whitney tends to sort our data to higher or lower. Moreover, Mann-Whitney is best for a two-tailed test and not good for a categorical like 4 seasons.

A significant limitation was that the data used for training our SVM model (Twitter data) came from a different source than the data we made predictions on (Reddit data). This was because we were unable to find Reddit data that included the r/vancouver subreddit and had sentiment labels. This meant that our model's performance was hindered by factors like how certain words or phrases may be used to convey different meanings between Twitter and Reddit, or how sarcasm and certain types of humor could be expressed differently between the 2 platforms.

If we had more time to complete our project, we would experiment with other stats models like regression. Moreover, we would attempt to implement our code on other social media platforms like Instagram, Twitter, or maybe Quora.

In summary, we conclude that the season does affect the sentiment of reddit comments on r/vancouver, but not reddit submissions.

Project Experience Summaries:

Erik Kho

- Implemented a chi-squared test to validate project hypotheses by calculating p-values.

Matthias Zelleke

- Utilized a support-vector machine model to predict sentiment on reddit data, providing the necessary data to deduce that time of year affects reddit sentiment