Sentiment Analysis of Twitter Data Using a Latent Dirichlet Allocation Topic Model

1 Abstract

This project is the continuation and an additional part of the previous project (project4). In this part of the project, we will do sentiment analysis using the LDA topic model. LDA was proved that is an efficient manner to find emotions in each tweet and summarize the results in a manner that is clearly understood.

2 Introduction

Sentiment analysis is broken into five steps. The basic workflow chart of this experiment is shown in Figure 2.1. The ***syuzhet*** package is used for sentiment analysis. This package extracts sentiment and sentiment-derived plot arcs from the text. There are a variety of dictionaries that exist for evaluating the opinion or emotion in text. The package incorporates three optional sentiment lexicons. The lexicons are: *nrc, afinn, bing*. The *NRC* lexicon involves a list of English words associated with eight emotions – anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. In addition, there are two sentiments: negative and positive. The lexicon includes 13,889 words in total. The *BING* lexicon uses a binary categorization model that sorts words into positive and negative categories. The lexicon includes 6,789 words in total. The *AFINN* lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. The lexicon includes 2,477 words in total. In this paper was used all three lexicons.

Figure 2.1: The basic workflow of the experiments.

3 Data presentation

As previously reported by project4, tweets were used as a data source and collected from time December 30, 2016 to Feb 04 2017, in JSON format. From JSON format the tweets were converted into CSV format, namely, a format appropriate for analysis. The subject of interest of the project was five different music genres: pop, rock, rap, classical, and folk. The dataset was used contains only the unique text tweets, that is, 10583 tweets. The frequency of the received data was almost daily. Specifically, for half an hour a day, the received tweets for the pop, rock and rap music genres approached approximately the two hundred tweets, while for the classical and folk music genres the received tweets were about sixty tweets.

4 Sentiment Analysis

In this chapter, the results of the lexicons will be presented.

4a. NRC lexicon

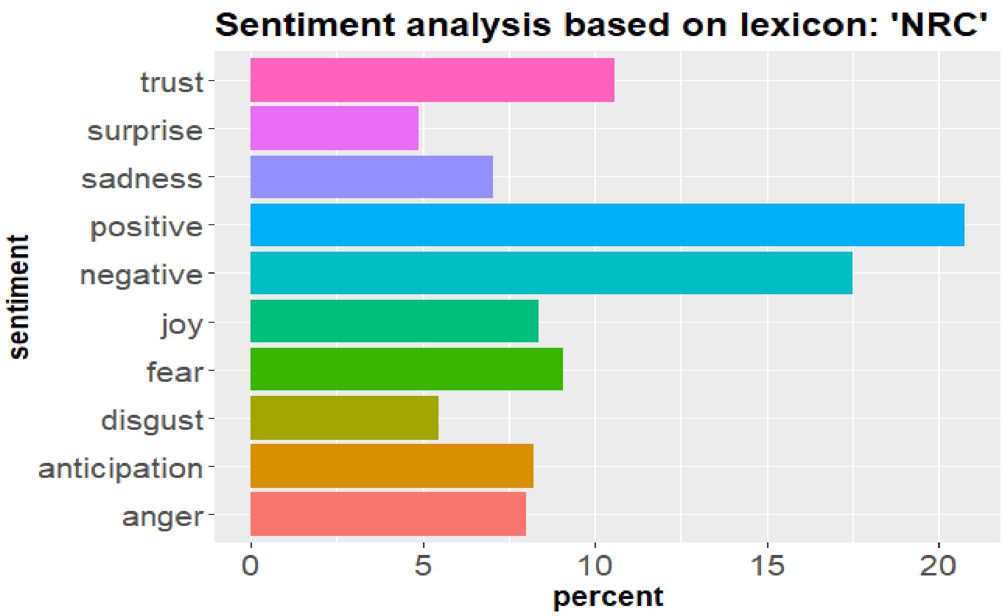

Figure 4.1: Distribution of emotion terms using NRC lexicon.

Table 4.1 indicates that the percentage of the total number of the positive terms in the collected data was 20.8%, while 17.5% corresponds to the number of negative terms. Figure 4.1 shows that both the positive and negative sentiments are high.

Table 4.1: Music Genres Data Sentiment Analysis Results-Two Sentiments.

Sentiment	Count
positive	4030
negative	3400

The emotion terms seem to be mixed. Table 4.2 indicates that emotion terms correspond to 51%. The emotion term trust accounts for 10.6% while the other emotions follow with relatively similar percentages, namely, there is a balance among them with the weakest term "surprise" account for 4.9%.

Table 4.2: Music Genres Data Sentiment Analysis Results-Eight Emotions.

Emotion	Count
trust	2055
fear	1760
joy	1620
anticipation	1600
anger	1555
sadness	1370
disgust	1060
surprise	950

Figure 4.2 indicates ten terms of each emotion category of the lexicon. It can be seen that trust words are more prevalent than emotional categories such as surprise. Additionally, it appears that there are common terms among the categories, such as the term “good” that appears in anticipation, trust and joy categories. Furthermore, it is obvious that there are misattributions of the terms in the categories, such as the term lovely that rated as sadness or the term boy rated as disgust or the term trump, which referred on Donald Trump, and rated as surprise. Thus, I consider that the NRC

lexicon may incorrectly classify some terms and identify a word as an emotional word while it should not.

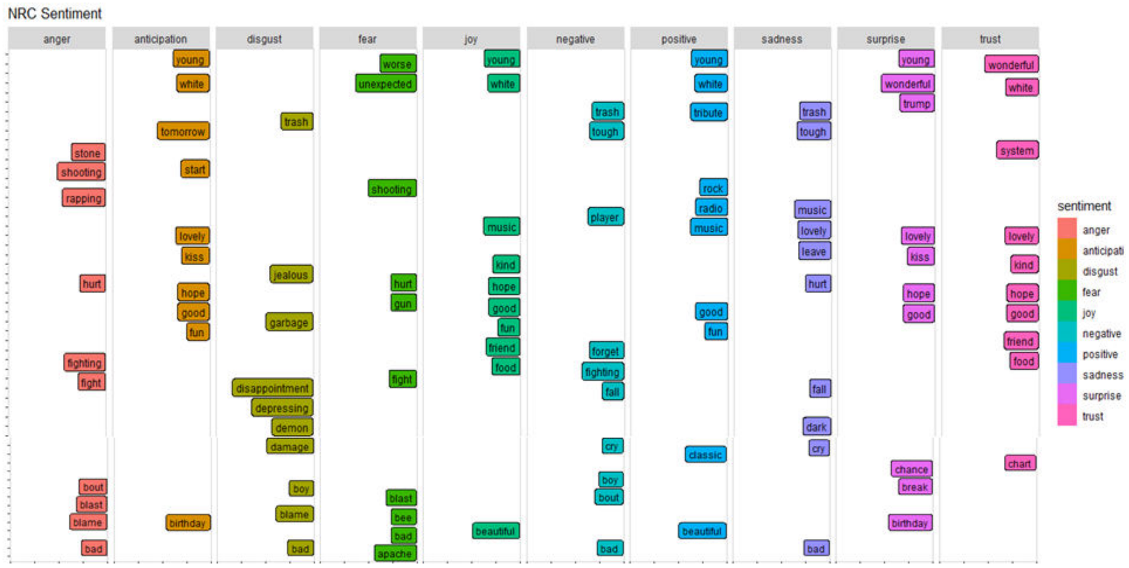

Figure 4.2: Terms for each emotion category.

4b. Bing lexicon

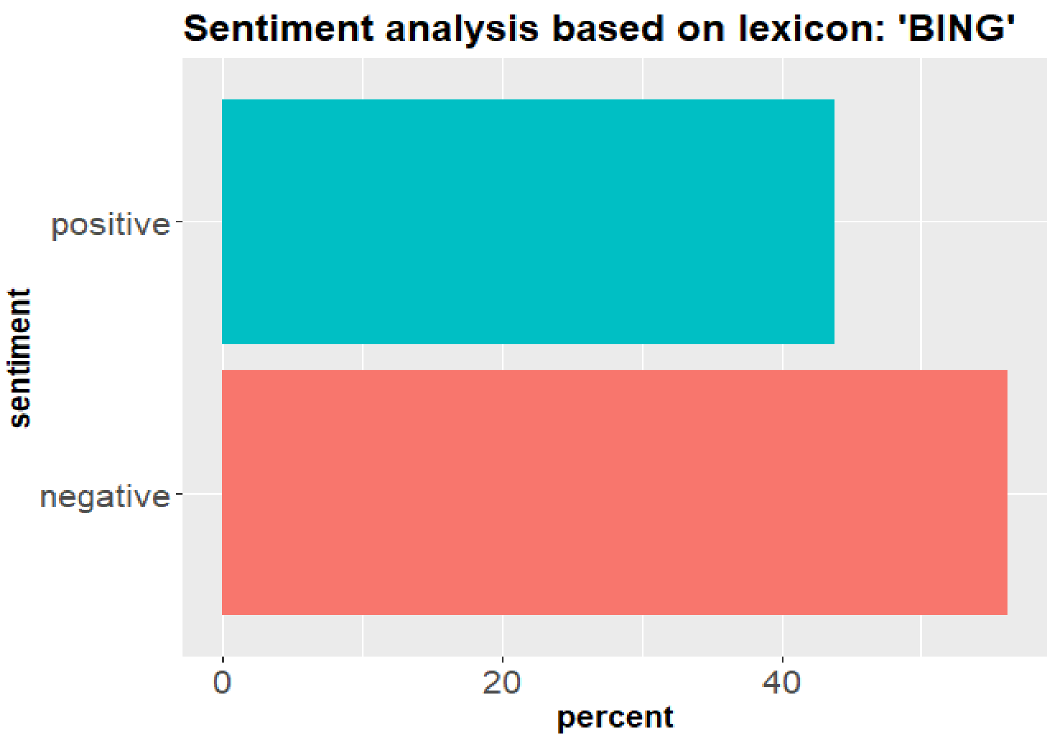

Figure 4.3: Distribution of sentiments terms using BING lexicon.

Table 4.3 indicates that the percentage of the total number of the negative terms was 56.1%, while 43.9% corresponds to the number of positive terms.

Table 4.3: Data Sentiment Analysis Results using BING lexicon.

Sentiment	Count
negative	3295
positive	2565

Figure 4.4 indicates ten words for each sentiment. In both sentiment categories, I observed that the terms were scored more accurately. In addition, in contrast to the NRC lexicon, the terms of the categories are more emotional, are more to the point. There are only three common terms among the two lexicons.

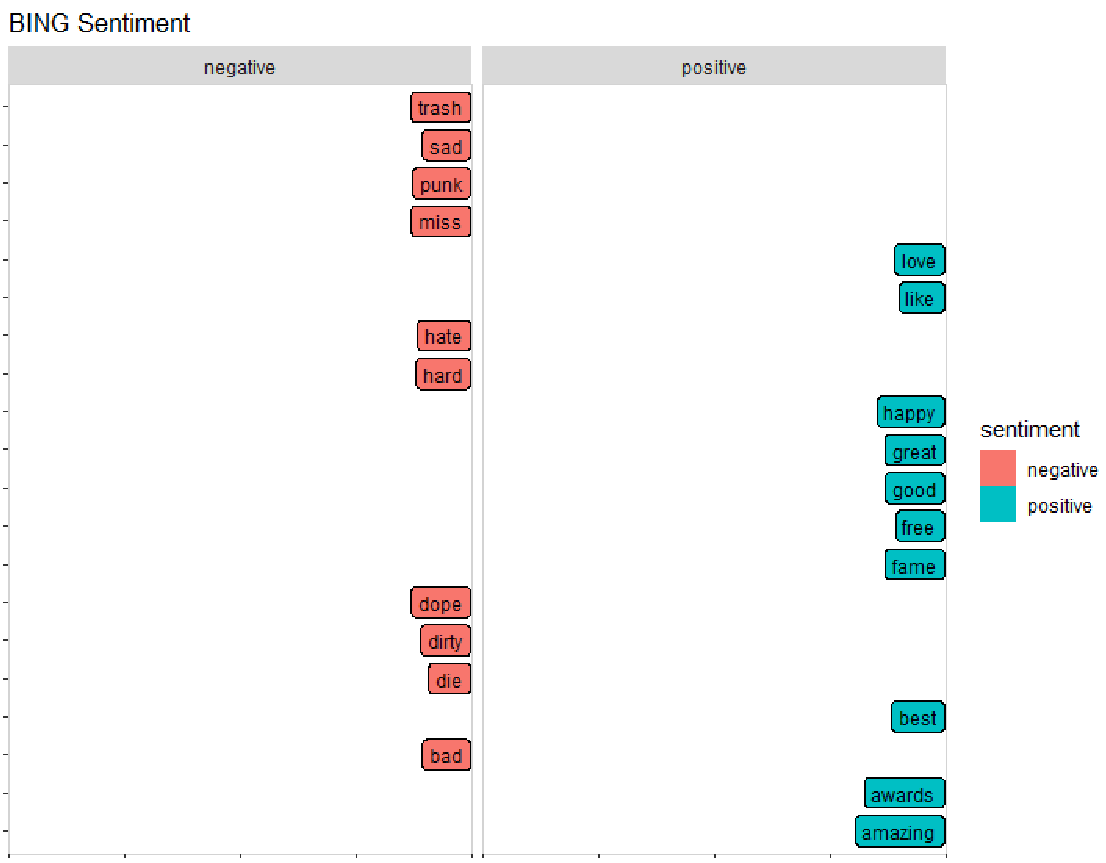

Figure 4.4: Ten words of each sentiment using BING lexicon.

4c. Afinn lexicon

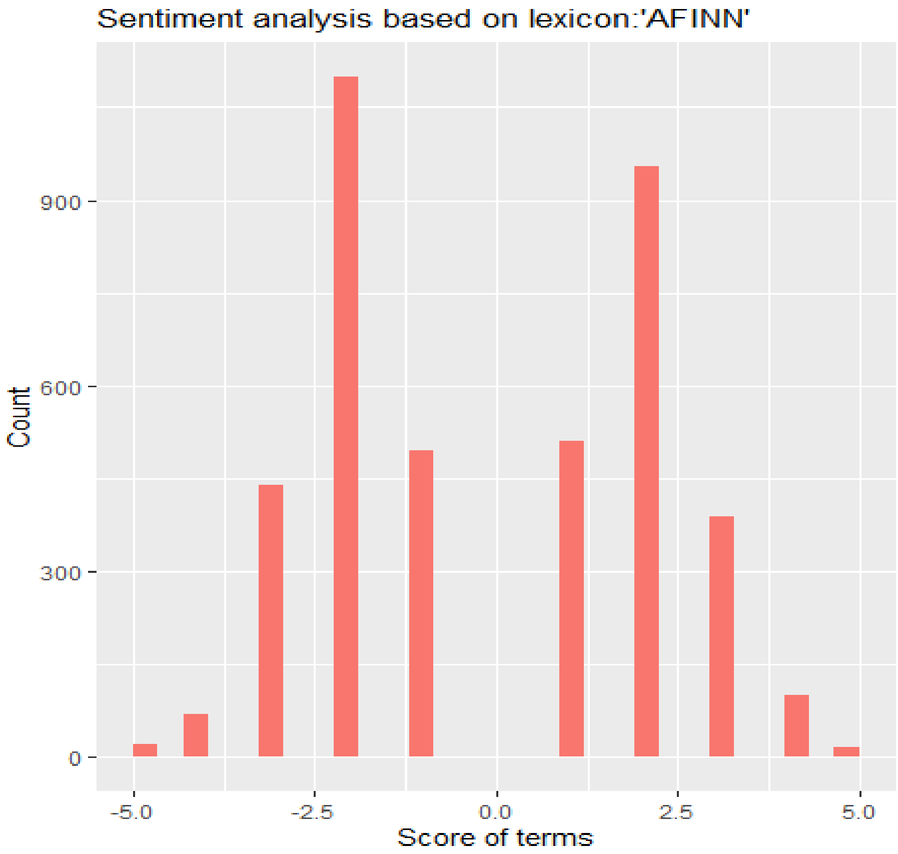

Figure 4.5: Distribution of the terms scores ranges from minus five to plus five.

AFINN terms are divided into four categories:

- Very Negative (rating -5 or -4)
- Negative (rating -3, -2, or -1)
- Positive (rating 1, 2, or 3)
- Very Positive (rating 4 or 5 or 6)

In particular, a score greater than zero indicates positive sentiment, while a score less than zero would mean negative overall emotion. Table 4.4 indicates that the percentage of the total number of the negative terms was 51.8%, while 48.1% corresponds to the number of positive terms.

Table 4.4: Data Sentiment Analysis Results using AFINN lexicon.

Sentiment	Count
negative	2125
positive	1970

Some very positive and negative terms were included in Table 4.5. In contrast to the other two lexicons, Afinn reveals other sentiment terms and I observe that the terms are as emotional as the BING lexicon. Additionally, there are common terms among the three lexicons.

Table 4.5: Very positive and negative terms using AFINN lexicon.

Positive Terms (4, 5)	Negative Terms (-4,-5)
outstanding	hell
superb	rape
amazing	pissed
fun	torture
awesome	fraud
wonderful	damned
heavenly	damn
fabulous	bullshit
miracle	whore

brilliant	

5 Conclusions

After utilizing the technique of sentiment analysis on the Twitter text, in this project, we conclude that both LDA topic modelling and sentiment analysis are powerful exploratory tools when used on a collection of Twitter data. The sentiment analysis was performed using the syuzhet package which contained three sentiment lexicons. The three packages under examination differ from each other. Some observations are: (a) NCR lexicon favors the positive terms in contrast to the other two lexicons. The AFINN and BING lexicons have many negative terms which mean that the overall opinion of users about the music genres is negative. Of course, some words may not refer exclusively to any of the music genres. Thus, in this case, it would be interesting to focus on sentences and not just in words. (b) NRC classifies words inaccurately into emotional categories, such as, human names (c) According to researches, the dictionaries with most words usually obtaining more accurate results as they have greater detection of emotions, in this project I could say that in this case, this is not true. The NRC lexicon despite more terms than the other two lexicons, we found that its results were imprecise, as there were terms that did not fit in the emotion categories, there were the same terms in more than one category and the resulting terms did not show as much emotion as the other two dictionaries. Furthermore, definitely the results from the lexicons would be different in a large dataset, as a larger dataset could mine deeper insights than a small case study. Thus, for a small case study, the lexicon that I would use is the Afinn lexicon.