A Tool for Sentiment Analysis on Twitter: Film Ratings as an Example

1. **Introduction**

Nowadays, electronic equipment has become an essential part of people’s lives. According to Johnson (2022), an average British spent 4 hours every day on mobile usage in 2021. This brings about the prosperity of social media platforms. As an example, there are more than 400 million monthly active users on Twitter alone (Statista, 2022). With the increasing availability of textual data brought about by social media platforms, the need for sentiment analysis for gaining insights is naturally high.

Sentiment analysis, also known as opinion mining, refers to the computational technique to detect affective polarity (e.g., positive, negative) of a given text. Thus, this technique is often applied to social media platforms to monitor public opinion of a product/service/brand. Also, since people are not always able to express their true feelings, sentiment analysis can be applied to check insight consistency by analysing people’s behavioural data (e.g., tweets on Twitter) that implies their opinions.

This report first proposes a tool for Twitter sentiment analysis based on Naïve Bayes classifier, and then apply the model to film ratings as a case study. In the case study, keywords extraction, topics extraction, and rating comparison are performed based on the sentiment analysis on The Northman, a recent film. Furthermore, a list of film ratings from IMDb, a film website, is compared with the ratings calculated from the sentiment analysis results.

1. **Model**

**2.1 Data**

The dataset used is Sentiment140 (Go, Bhayani and Huang, 2009). The dataset contains 1600000 labelled tweets that are automatically extracted from Twitter. The tweets are labelled as 0 (negative) and 4 (positive) evenly. The balance distribution of labels is quite favourable as imbalanced dataset is likely to result in poor prediction performance, especially on minority labels. While some may argue that including the neutral class can better reflect the real-world situations than otherwise, Pak and Paroubek (2010) found bad performance when including a neutral class in the model based on naïve bayes method. Moreover, in the context of this report, it is either too time-consuming to include a labelled neutral class to the same number as the others, or unfavourable to include a neutral class with a relatively small number, which may bring about the imbalance data problem. Therefore, the labels are left as they originally are.

**2.1.1 Data Preparation**

Data preparation can be divided into two parts—data split and text pre-processing. For data split, since naïve bayes method requires no parameter tuning, the original dataset is divided into train set and test set, where the test set is 20% of the original dataset.

For text pre-processing, tokenisation, word regularisation, and word length control are applied with punctuation removed. Specifically, tweets are first converted to a list of lists with tokenised lower-case words. Then, each word is further normalised with ‘-’ and ‘.’ removed so that terminologies are regularised (e.g., U.S. converted to US). Moreover, words with a single letter are removed. Finally, the cleaned tweets are stored in a list with each word joined by white space. The reason why stop words are not removed is that TF-IDF representation is used, which is discussed in more details in session 2.2.

* + 1. **Text Representation**

Typical text representations are bag of words and word embeddings. While it is generally accepted that word embeddings excel bag of words in a way that they reflect the distance between words in the real world by the distance between the corresponding words in the trained vector space, they are often used in neural networks rather than in naïve bayes methods as naïve bayes assumes that word order does not affect probability computation.

The text representation used in this report is term frequency-inverse document frequency (TF-IDF). TF-IDF is calculated by multiplying term frequency and inverse document frequency. While the former is naturally a word’s frequency, the latter is computed by log (n/df), where df is the number of times the word appears in a document. In this way, words that are easily seen in many documents are penalised as less important. For the same reason, stop words are penalised heavily as they appear in almost every document, which justifies not removing stop words in the data preparation step. Therefore, it is clear that TF-IDF is more favourable than one-hot coding representation, as the former representation contains more information.

* 1. **Model and Results**

**2.2.1 Model based on Binomial Naïve Bayes**

The model is built based on binomial naïve bayes method using ‘BernoulliNB’ from sci-kit learn. The mathematical representation of Bayes Theorem is:

P (A|B) = P (B|A) \* P (A) / P (B) (1)

where A and B are events, P (A|B) is the probability of A given B is true (also known as posterior probability), P (B|A) is the probability of B given A is true, and P (A) and P (B) are independent probability of events A and B, where P (A) is also known as prior probability.

In the context of text classification, the formula can be reframed into:

P (yi | x1, x2, …, xn) = P (x1, x2, …, xn | yi) \* P (yi) / P (x1, x2, …, xn) (2), where y is label and x is word.

Naïve Bayes method further assumes that each word is independent from each other so that (2) is transformed to:

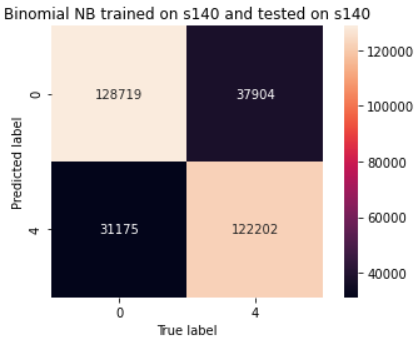
P (yi | x1, x2, …, xn) = P (x1|yi) \* P (x2|yi) … P (xn|yi) \* P (yi) (3). In this way, the probability of labels is calculated and the one with the largest probability is predicted.

* + 1. **Results**

Overall, the model receives the same performance on accuracy, precision, recall, and f1-score at 77% correct rate as presented in table 1 and figure 1. As indicated by table 2, the result implies a satisfying performance of the model with no sign of overfitting considering the same result from the four metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-score |
| 0 (negative) |  | 77% | 81% | 79% |
| 4 (positive) |  | 80% | 76% | 78% |
| Overall (Macro) | 78% | 78% | 78% | 78% |

*Table 1. Test Results*



*Figure 1. Heatmap*

|  |  |
| --- | --- |
| Metrics | Formula |
| Accuracy | (TP+TN)/(TP+TN+FP+FN) |
| Precision | TP/(TP+FP) |
| Recall | TP/TP+FN |
| F1-score | 2\*Precision\*Recall/(Precision+Recall) |
| Notes | TP=True Positive, FP=False Positive, TN=True Negative, FN=False Negative |

*Table 2. Metrics Formula*

**3 Case Study**

Overall, this session is divided into two parts. First, a sentiment analysis on a recent film, The Northman, is conducted to find the public opinion towards it. The tweets are further analysed for keywords, topics extraction and rating comparison. The second part concerns multiple sentiment analyses on a list of popular films from IMDb, a film website with film ratings rated by users. A comparison is then conducted between the ratings from the website and the ratings computed from the detected sentiment. The features of the films with contradictory findings are further analysed to investigate the potential reasons behind it.

**3.1 Sentiment Analysis on The Northman**

The Northman is a recent action film. According to IMDb (2022), The Northman ranks in the top 5 of the most popular films. It also has a fairly high rating score at 7.9 out of 10 rated by 29K IMDb users.

**3.1.1 Data and Methodology**

The data is retrieved via Twitter API v2 with essential access. Therefore, the tweets used in this study are tweeted within the past 7 days and the number of tweets for each request is 100.

In this study, the function for getting sentiment is defined using the saved naïve bayes model from the previous session and the 100 tweets retrieved by Twitter API. Tweets are processed the same way as the training set with stop words removed before being fed into the naïve bayes model for consistency. The ‘get\_keywords’ function is designed to take the output data frame of ‘get\_sentiment’ function, and a required sentiment input. The function selects the required tweets and sentiment corresponding to the sentiment input, and compute TF-IDF score for each word based accordingly. Keywords are the ones with top scores. Same as ‘get\_keywords’, the ‘get\_topics’ function also takes the output from ‘get\_sentiment’ and the required sentiment as inputs. The difference is that it utilises Latent Dirichlet Allocation (LDA). LDA is an unsupervised algorithm that generates the possible topics of given text. It suits this study in the way that it is a bag of words model. Finally, ‘get\_rating’ function takes the output sentiment data frame and computes the rating based on the percentage of positive predicted labels out of all predicted labels.

**3.1.2 Findings and Discussion**

|  |  |
| --- | --- |
| **Keywords (Negative)** | **TF-IDF** |
| fully | 0.201581 |
| icy | 0.201581 |
| feared | 0.201581 |
| generic | 0.201581 |
| hamlet | 0.201581 |
| seen | 0.201581 |
| hate | 0.201581 |
| hours | 0.201581 |
| land | 0.201581 |
| drunk | 0.201581 |

*Table 3. Keywords for Tweets with Negative Sentiment*

|  |
| --- |
| haaaaard thenorthman tonight scotchnoblemen eastern standard stream team th time |
| https art says fuck a8tlrhi37v thenorthman rogelio carlos nic diaz |
| theater watching thenorthman northman https haaaaard feeling vikings story tomorrow |
| thenorthman northman https feeling vikings haaaaard story tomorrow watch want |
| review kinda watch want tomorrow que thenorthman y6fsxthjut robert revenge |

*Table 4. Top 5 Topics for Tweets with Negative Sentiment*

|  |  |
| --- | --- |
| **Keywords** | **TF-IDF** |
| shoul | 0.353559 |
| far | 0.353559 |
| obama\_llama\_fur | 0.353559 |
| favorite | 0.322282 |
| year | 0.322282 |

*Table 5. Keywords for Tweets with Positive Sentiment*

|  |
| --- |
| https thenorthman screen film rt big playing seen theaters star |
| watch https thenorthman en el 2022 sobre way gd3niccghi rt |
| thenorthman rt eggers bloody robert https tonight going northman saw |
| scene incest hfr4wvubox close read roxana\_hadadi diabolical thenorthman https big |
| thank thenorthman eggers rt robert costumes cinema toxic loud certified |

*Table 6. Top 5 Topics for Tweets with Positive Sentiment*

|  |  |
| --- | --- |
| Rating on IMDb | 7.9 |
| Rating based on Sentiment Analysis | 9.2 |

*Table 7. Rating Comparison for The Northman*

The findings of this study are presented in table 3 to 7. As can be seen from the tables, the results for keywords and topics are not particularly interesting. Not only are there non-English words, urls and unregularized words, but also the remaining keywords and topics do not provide much information. N-grams might help with this situation in terms of future work.

However, the rating calculated from sentiment analysis is consistent with the high rating on IMDb. 9.2 in the rating from sentiment analysis means that 92 out of 100 tweets on The Northman are predicted positive. In this sense, it is likely that The Northman is a rather interesting film. The result also indicates that IMDb could be a credible website, which, however, cannot be determined until further investigation.

**3.2 Comparison of Ratings: Intended vs Behavioural**

Like IMDb, most ratings for items on websites are rated by users. However, people do not always know themselves as they think they do, and the understanding of the same object may not be consistent across different people. In this sense, behavioural data (e.g., tweets, likes, retweets) can supplement intended ratings as well as help with consistency check.

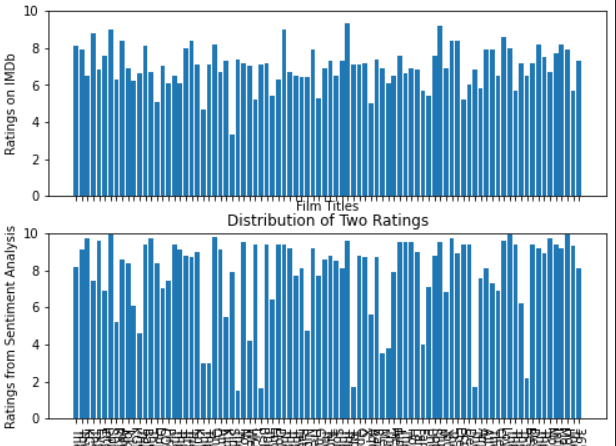
**3.2.1 Data, Methodology and Analysis**

As mentioned in previous sessions, this part of study compares the ratings of 88 films on IMDb website, and those computed based on the proportion of positive classes out of total classes for each film. Therefore, the data used comes from two sources—webpage information on IMDb ‘Most Popular Movies’ website crawled by ‘BeautifulSoup’, and public tweets acquired from Twitter API v2 with essential access. The number of tweets acquired from each query is 100.

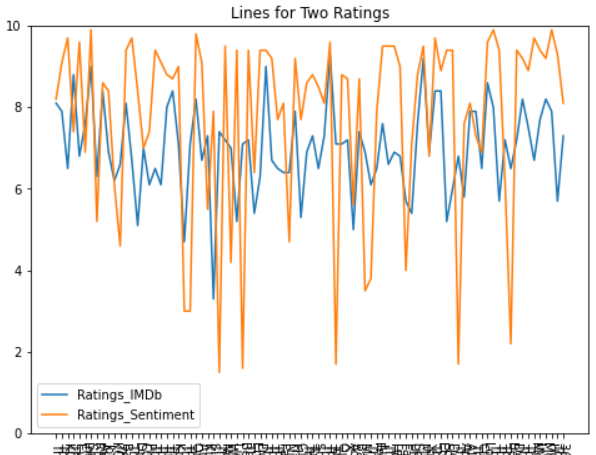
Specifically, the crawled information regards film titles, film ratings, and film released years. After dropping missing values, removing specific punctuations, and dropping films unsuitable for API query (e.g., films with ‘or’ or ‘and’ in titles), the information of 88 films are stored in 3 lists in such a way that their indices correspond with each other. Ratings from sentiment analysis are generated by sending film titles to ‘get\_sentiment’, the output data frame of which is then sent to ‘get\_rating’. The computed ratings are stored in a list corresponding to the indices of the list of film titles.

For analysis, a bar chart and a line plot are first presented to show the trends of the two ratings. It should be noted that the x axis is the film titles so the plot should be interpreted categorically. The adoption of line chart is purely for illustration convenience. Then, pairs of ratings are stored in a list, which is subsequently converted to pairs of sentiment by replacing ratings smaller than or equal to 5.5 with 0 (negative), and the else with 4 (positive). After that, films with contradictory sentiment are presented with their released years, followed by a bar chart on contradictions by years. The films in years with contradiction rate no less than 50% are investigated manually.

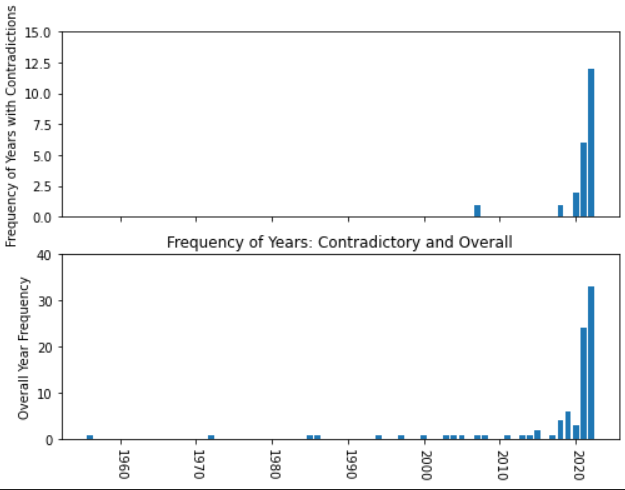
**3.2.2 Findings and Discussion**



*Figure 2. Ratings Comparison Bar Chart*

**

*Figure 3. Ratings Comparison Line Chart*

**

*Figure 4. No. of Contradictory Film Sentiment by Year*

Overall, the ratings generated from sentiment analysis are consistent with the ratings from IMDb, which indicates that IMDb is a reliable film website. While there is an upward trend in contradictory findings when the publish year gets recent, this pattern agrees with the overall year frequency. The likely reason for the skewness is that the tweets used in this study are created within the last 7 days. Naturally, recent films are more often discussed than the old ones.

Based on analysis on contradiction rate (contradiction/overall per year), 365 Days, Cleaner, and The Night House are further investigated manually on IMDb. It appears that Cleaner and The Night House share the genres of both mystery and thriller while 365 Days is classified as drama and romance. This result is somewhat interesting as thriller, mystery, and crime as genres are often regarded similar in the psychological effect on the audience. The result could suggest that films arousing such psychological states might result in polarity in viewer sentiment and reviews. However, it should be emphasised that the datapoints are too few to reach a concrete conclusion, and a thorough investigation on features of more films is encouraged.

**4 Conclusion**

In conclusion, this project trains a binomial naïve bayes model with 78% test accuracy, and applies it to film tweets using Twitter API. While the keywords and topics extracted are not particularly interesting, there is fairly high consistency between the ratings on IMDb and the ratings calculated from tweets, indicating the reliability of IMDb. Despite an upward trend in contradictory sentiment when year increases, the phenomenon is consistent with the skewed data in years. Moreover, further investigation into the contradiction rate suggests that films in thriller and mystery genres might receive polarity in viewer sentiment, which is not very concrete due to the limited datapoints.

**References**

Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford, 1*(12), p.2009.

IMDb (2022) *Most Popular Movies*. Available at: https://www.imdb.com/chart/moviemeter/?ref\_=tt\_ov\_pop (Accessed: April 2022)

Johnson, J. (2022) *Average daily mobile usage in the United Kingdom from 2019 to 2021.* Available at: https://www.statista.com/statistics/1285042/uk-daily-time-spent-mobile-usage/ (Accessed: April 2022)

Pak, A. and Paroubek, P., 2010, July. Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 436-439).

Statista (2022) *Most popular social networks worldwide as of January 2022, ranked by number of monthly active users.* Available at: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (Accessed: April 2022)