Twitter-Based Sentiment Analysis of Movies Using Text Mining and Natural Language Processing Techniques

Dr. Rohan Verma[1]*, Dr. Priya Nair[1]
[1]Indian Institute of Technology Delhi, Department of Computer Science and Medical Data Analytics, New Delhi, India

ABSTRACT

A movie review is an article describing the movie and giving critics on it that allows the viewer or anyone to know and understand whether it is really worth to watch that movie or not. The practice of using analytics, using audience sentiment, to measure movie's success is not a new phenomenon. The present work is based on the major project work carried out as a part of Bachelor of Engineering programme and is intended to demonstrate gauging audience sentiments about the movie "Alita: Battle Angel". Text Analytics using Twitter data is used to calculate sentiments using two approaches, namely, "tm" package and "syuzhet" package.

KEYWORDS: text analytics, sentiment analysis, tm package, syuzhet package, twitter.

1. INTRODUCTION

Information has become the life line of the present-day business environment. Business uses the information to identify, understand and fulfil the requirements of its stakeholders e.g. customers, employees etc. The World Wide Web is a big storehouse of information available in various forms like chats, emails, survey results, tweets, call center data, phone transcripts, online customer reviews, blogs, recorded interactions, articles and other documents. The information so available is in raw form and not easy to understand. It can be made meaningful by using a text analysis tool.

Text Analytics, also known as Text Analysis or Text Mining, is a process of converting unstructured or raw text data into simple, meaningful and understandable form to provide an insight into the requirements of stakeholders and draw an inference from it. The unstructured data can be analysed with the help of software that can identify concepts, patterns, trends, keywords and other attributes in the data and helps in finding valuable business insights in corporate documents, customer emails, call center logs, survey comments, social network posts, medical records and other sources of text-based data. Document retrieval, spam filtering, chatbots, sentiment analysis, financial forecasting etc. are some common applications of text analytics. The practice of using analytics to gauge audience sentiment to measure the success of a movie is not a new phenomenon. Social media can be effectively used to measure audience sentiment about a movie. The views and opinions expressed by the people on social media platforms and digital media like Twitter, Face book etc. can be collected and used to gauge their sentiments.

Sentiment analysis is used to understand sentiment of customers i.e. their feelings, attitude, emotions, appraisals or opinions towards certain objects, entities or events expressed in text. Natural language processing, statistics, and text analysis is used to extract, and identify the sentiment of text into positive, negative, or neutral categories. Sentiment Analysis identifies the phrases in a text that bears some sentiment. The writer may speak about some objective facts or subjective opinions. It is necessary to distinguish between the two. Sentiment analysis finds the subject towards whom the sentiment is directed. A text may contain many entities but it is necessary to find the entity towards which the sentiment is directed. Sentiments are classified as objective (facts), positive (denotes a state of happiness, bliss or satisfaction on part of the writer) or negative (denotes a state of sorrow, dejection or disappointment on part of the writer). The sentiments can be given a score based on their degree of positivity, negativity or objectivity [4].

The decision making process is affected by the thought of people. People orally share attitudes, opinions, or reactions about businesses, products, or services with other people. People rely on families, friends, and others in their social network. This is known as Word of Mouth communication and it relies on social networking and mutual trust [4]. In addition, a person, willing to buy a product, can start by reading the online reviews, opinions and ideas written by other people [3]. This is where Sentiment Analysis comes into play. Growing availability of opinion rich resources like online review sites, feedbacks, blogs, social networking sites have made this "decision-making process" easier. It is an established fact that consumer voices affect shaping voices of other consumers. Sentiment Analysis thus finds its use in Consumer Market for Product reviews, marketing for knowing consumer attitudes and trends, Social Media for finding general opinion about recent trending topics, movies to find whether a recently released movie is a hit [4].

Pang-Lee et al. [2] broadly classified the applications of Sentiment Analysis into the following categories:

i. Applications to Review-Related Websites
- Movie Reviews, Product Reviews etc.

ii. Applications as a Sub-Component Technology
- Detecting antagonistic, heated language in mails, spam detection, context sensitive information detection etc.

iii. Applications in Business and Government Intelligence
- Knowing Consumer attitudes and trends

iv. Applications across Different Domains
- Knowing public opinions for political leaders or their notions about rules and regulations in place etc.

The present work is based on the major project work carried out as a part of Bachelor of Engineering programme and is intended to demonstrate gauging audience sentiments about the movie "Alita: Battle Angel". Text analytics is used for sentiment analysis of the movie 'Alita: Battle Angel', released in India on 8th February 2019. Alita: Battle Angel is a 2019 American cyberpunk action film based on Yukito Kishiro's manga series Gunnm, also known as Battle Angel Alita. It is directed by Robert Rodriguez, written by James Cameron and Laeta Kalogridis and produced by Cameron and Jon Landau. Rosa Salazar stars as Alita, a cyborg, with supporting roles portrayed by Christoph Waltz, Jennifer Connelly, Mahershala Ali, Ed Skrein, Jackie Earle Haley and Keean Johnson.

2. METHOD

The method used for sentiment analysis broadly consists of following steps:

i. Getting data for analysis: Gets Tweets from Twitter
ii. Data Pre-processing to clean the data
iii. Perform a Sentiment Analysis.
iv. Creating a Word Cloud of positive and negative words.
v. Making a Prediction Algorithm for Movie Success.

Data for the Analysis

The data required for analysis has been retrieved from tweets on Twitter. The data is publicly available and can be collected scraping through the website or by using a special interface for programmers that Twitter provides, called API. The 'twitteR' package in R language is used to collect tweets/retweets about the movie 'Alita: Battle Angel'. The tweets are extracted, saved in a 'csv' file, and then read into R using the 'readr' package.

Data Pre-processing

Data pre-processing is a process that transforms raw data into an understandable format. The raw data is often incomplete, inconsistent and lacks certain attributes of interest or trends. The raw text data is filtered through several cleaning phases to get it transformed into a tabular format for analysis.

At first a matrix comprising of documents and terms is created. A document can be considered as a matrix in which each row represents product description and each column represents terms. Terms refers to each word in the description. Usually, the number of documents in the matrix equals to number of rows in the given data. In the next step Text cleaning is performed in following ways:

- Remove words - If the data is extracted using web scraping, it is essential to remove html tags.
- Remove stop words - Stop words are a set of words which helps in sentence construction and don't have any real information. Words such as a, an, the, they, where etc. are categorized as stop words.
- Convert to lower case - To maintain standardization across all text and get rid of case differences and convert the entire text to lower case.
- Remove punctuation - Punctuation is removed as they don't deliver any information.
- Remove number - Similarly, numerical figures are removed from text
- Remove whitespaces - Then, used spaces in the text are removed.
- Stemming & Lemmatization - Finally, the terms are converted into their root form to capture the intent of terms precisely. For example: Words like playing, played, plays gets converted to the root word 'play'.

Perform Sentiment Analysis

Two approaches, namely, "tm" package and "syuzhet" package are used to calculate the sentiment scores:

Approach I - the "tm"package

The 'tm' package is a framework for text mining applications within R. It works on d Text Mining 'Bag of Words' Principle. Bag of Words approach of handling texts is very simple. It just counts the frequency of each word appearance in the text and uses these counts as the independent variables. This simple approach is very effective and it's used as a baseline in text analytics projects and for natural language processing. The sentiment is calculated using "calculate_sentiment function". This function loads text and calculates sentiment of each sentence.

Table 1 and 2 below shows one page each of the frequency of words in the text classified as positive and negative respectively. The same has been generated using the data table function.

Table 1 Frequency of words in text classified as positive

Show 10 entries Search:

	text	sentiment	freq_up
action	action	Positive	249
actual	actual	Positive	77
adapt	adapt	Positive	139
ahead	ahead	Positive	6
allow	allow	Positive	1
ambassador	ambassador	Positive	2
angel	angel	Positive	2471
appeal	appeal	Positive	1
art	art	Positive	71
artist	artist	Positive	3

Showing 1 to 10 of 322 entries Previous 1 2 3 4 5 … 33 Next

Table 2 Frequency of words in text classified as negative

Show 10 entries Search:

	text	sentiment	freq_up
acid	acid	Negative	1
afraid	afraid	Negative	4
alien	alien	Negative	1
anger	anger	Negative	1
annoy	annoy	Negative	6
anxious	anxious	Negative	1
argument	argument	Negative	2
ass	ass	Negative	21
avoid	avoid	Negative	3
await	await	Negative	3

Showing 1 to 10 of 300 entries Previous 1 2 3 4 5 … 30 Next

The Wordcloud of positive and negative words is shown in Figure 1 and 2.

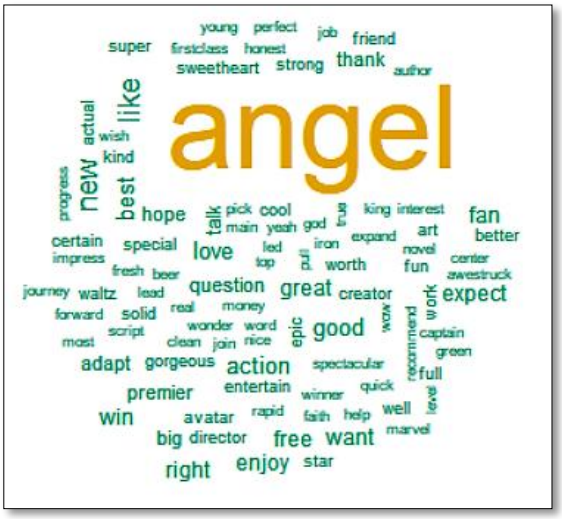

Figure 1 Worldcloud of positive words

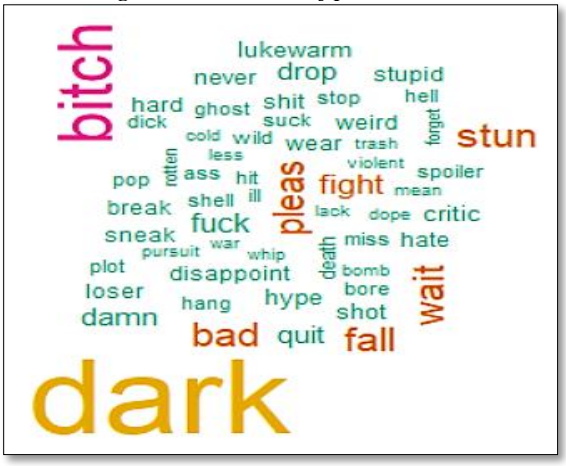

Figure 2 Wordcloud of negative words

The review is considered positive if it contains more positive than negative terms, and negative if the number of negative terms is greater than the number of positive terms. A review is neutral if it contains an equal number of positive or negative terms [1]. Based on above analysis the ratio of positive to negative words is 9105/2899 = 3.14 which indicates that the movie has been received well by the audience.

Approach II - the "Syuzhet" package:
Syuzhet is R package for the extraction of sentiment and sentiment-based plot arcs from text. The name "Syuzhet" comes from the Russian Formalists Victor Shklovsky and Vladimir Propp who divided narrative into two components, the "fabula" and the "syuzhet." Syuzhet refers to the "device" or technique of a narrative whereas fabula is the chronological order of events. Syuzhet, therefore, is concerned with the manner in which the elements of the story (fabula) are organized (syuzhet). It uses four sentiment lexicons afinn, bing, nrc, and of course the default syuzhet itself for the analysis.

After processing the data the 'get_nrc_sentiment' function is used to extract sentiments from the tweets. The output is a data frame where each row represents a sentence from the original file. The columns include one for each emotion type as well as a positive or negative valence. The ten columns are as follows: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative" and "positive". The ten columns are shown as bar chart in Figure 3.

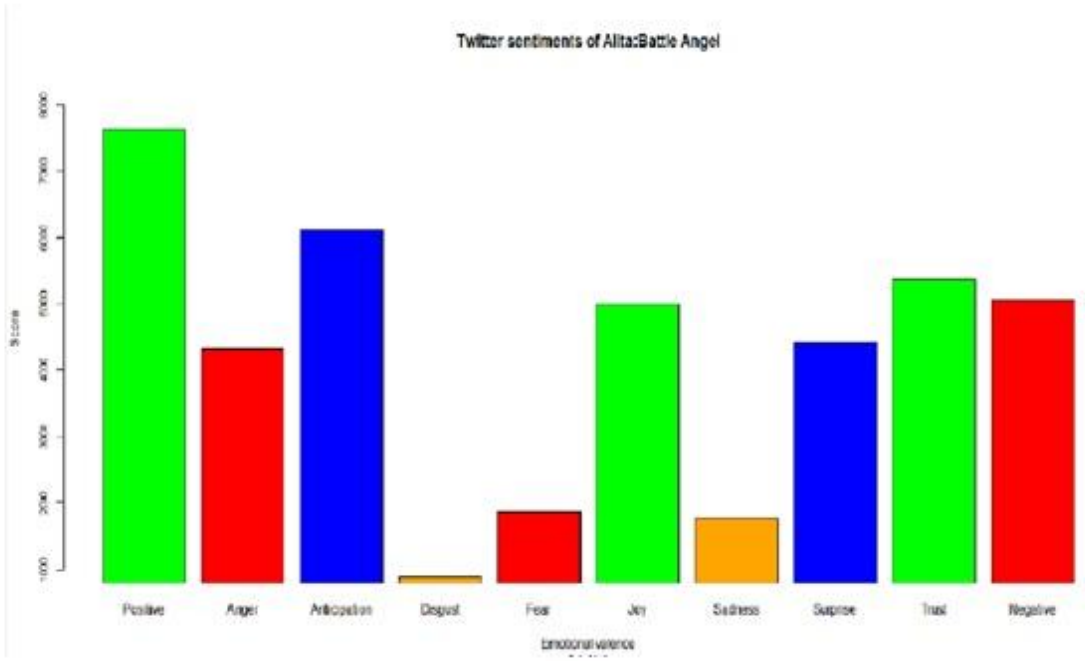

Figure 3 Bar chart of emotions

Looking at the bar chart and the sum of these emotions, we can see that the positive sentiments ('positive', 'joy', 'trust') comfortably outscore the negative emotions ('negative', 'disgust', 'anger'). Thus it can be said that the audience has received the movie positively.

3. RESULTS AND DISCUSSION

i. Approach I – the "tm" package: the ratio of positive to negative words is 9105/2899 = 3.14 which indicates that the movie has been received well by the audience.

ii. Approach II – "Syuzhet" package: the sum of positive sentiments outscore the negative sentiments indicating that the audience response towards the movie is positive.

4. CONCLUSION

Both the approaches the "tm"package and the Syuzhet package suggests that the movie ''Alita: Battle Angel" has been well received by the audiences, as measured by the Positive Tweet to Negative Tweet ratio as well as by the visual representation of various emotions.

REFERENCES

[1] Alistair Kennedy and Diana Inkpen, Sentiment classification of movie and product revi ews using contextual valence shifters, Computational Intelligence, 22(2):110–125, 2006.

[2] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and trends in information retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.
[3] A.B. Pawar, M.A. Jawale and D.N. Kyatanavar Fundamentals of Sentiment Analysis: Concepts and Methodology, Sentiment Analysis and Ontology Engineering, W. Pedrycz and S.-M. Chen (eds.), Studies in Computational Intelligence 639, DOI 10.1007/978-3-319-30319-2_2.
[4] Subhabrata Mukherjee and Dr. Pushpak Bhattacharyya, Literature survey on Sentiment Analysis, IIT Bombay (2012).
[5] Measuring Audience Sentiments about Movies using Twitter and Text Analytics, www.analyticsvidhya.com, 2017