Sentiment Analysis of Book Reviews From Amazon Data

Shandro Chakraborty, Zarin Tasnim, Bishakha Dhar

Dept. of Computer Science and Engineering

BRAC University

Dhaka, Bangladesh

shandrone@gmail.com; susmitahq@gmail.com; bishakhadhar.riya95@gmail.com

***Abstract*- This paper proposes a new model of sentiment analysis for book reviews. In the proposed model, book rating will be generated by doing sentiment analysis on public opinion. As Twitter is one of the most popular micro-blogging sites, for public opinion we collected data from Twitter. Before fitting the algorithm we pre-processed the gathered data to a supervised form. In the model Naïve Bayes algorithm was trained on a training set and to validate the performance, the algorithm was tested to get better accuracy. After that, using these algorithm sentiment analysis was done on the data to get polarity. To validate the model we generated rating from calculating polarity and plotted in a graph where results obtaining from all the algorithm were shown.**

Keywords: Supervised Machine Learning, Sentiment Analysis, Naive Bayes, Lemmatizing, Stop-word, Stemming

I. INTRODUCTION

In this age of technology and internet people more often uses the internet and other technologies to help themselves in various way. People who love reading book often don't have much free time to go through all the internet review sites find to select a book. Even then the rating given in any site is usually done by professional journalists and reviewers with whom many people often disagree with. Thus this decision making process becomes more challenging. In this project we'll build a model that generates book ratings based on public opinion which will give users of the model accurate reflection of the public's opinion about any particular book. In this project we'll build a model that generates book ratings based on public opinion which will give users of the model accurate reflection of the public's opinion about any particular book.

The rest of the paper is organized by as follows. Section II describes related works performed by other projecters in this field; Section III described the proposed method to perform the experiment and Section IV describes the primary results and discussion obtained from the experiments, Section V presents the conclusion and suggested future work.

II. BACKGROUND ANALYSIS

A. *Sentiment Analysis*

Sentiment analysis is growing area of research in the field of natural language processing (NLP) [5]. A fair amount of research focused on how sentiments are expressed in various categories such as online reviews, news article etc. along with how sentiments are expressed given the informal language of social media and micro-blogs [9]. Research on Twitter data has found that Twitter data has an impressive predictive power that ranges from stock market to movie performance [1]. Minqing Hu and Bing Liu in 2004 proposed a model they aimed to mine and summarize online opinions in reviews, blogs and forums [2]. Their work in sentiment analysis on online reviews was one the first research that was done in the respective field and this research set the motion for future research on sentiment analysis on online opinion summarizing. For opinion summarization the focused on quantitative aspect of the opinions. E. Junque de Fortuny, T De Smedt, D Martens in 2010 proposed a model to scrape relevant text from websites and perform sentiment analysis on the scraped data. Their subject of sentiment analysis was Belgian elections. The corpus used for processing were gathered from online versions of all Flemish newspaper. A web crawler was used, each adjective was manually given a polarity score for sentiment analysis [3]. Efthymios Kouloumpis, Theresa Wilson , Johanna Moore in 2011 proposed a model that used the features that capture information about the informal language used in micro blogging to analyze the features for detecting the sentiment of Twitter messages. This research also evaluated the usefulness of existing lexical resources [4]. Like above research, we discussed there are a number of improved model for sentiment analysis have been developed so far.

A. Naïve Bayes

Naïve Bayes is a classification technique that is based on the Bayes' theorem. Initially a Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature [6]. Even though Naïve Bayes is relatively simple algorithm, it is known to outperform some of the sophisticated [classifier [6]. The mathematical representation of Naïve Bayes equation is show in figure no 1

$$P(c|x) = \frac{P(x \vee c)P(c)}{P(x)} \tag{1}$$

III. METHODOLOGY

A. Collection of Data

In this project we used Amazon product data's book reviews [10]. This dataset consisted of 47300 user reviews of video games. The columns in the dataset are consisted of reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, reviewTime. Where reviewerID is the unique ID of the reviewer, asin is the ID of the product in this case the video game, reviewerName is the username of the person who wrote the review, helpful is the helpfulness of the given review, overall represents the rating of the video game given by the user on a scale of 1 to 5, summary is the column consisting review summary, reviewText represents the review written by the user, unixReviewTime is the unix time of the review and reviewTime is the raw time of the review.

For twitter data we scraped twitter data using a script which allows us to scrap tweet given a name, date range etc. As official Twitter API for collecting tweets only allows to download only 7 days of data, therefore we had to use a scrapper for collecting data from several months ago.

B. Data Pre-Processing

1. Lemmatization

Lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended parts of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document [7].

2. General Cleaning

Next we have dropped all the blank rows from the data table. After that we have taken care of all the hyperlinks from the data. Next we removed the twitter tags. These tags bears no sentiments. That's why we have eliminated these. Next we have removed all the non ASCII characters. These non ASCII characters has no influence on sentiment analysis. Finally we have removed all the apostrophes.

3. Stop Words

Stop words are words which are filtered out before processing of natural language data. Some examples of stop words are: "a", "and", "but", "how", "or", and "what.". These words have no sentiment values. So we removed these stop words.

C. Training and Testing Data

One of the important parts of Machine Learning model is to split the dataset into two parts, training and testing set. The reason we split our dataset into these two parts is to test the accuracy of the algorithm [8]. Here the algorithm will be trained using the labeled data and tested on the test set which helps to see the accuracy of the algorithm that are used in the model. This accuracy will give us a better idea about how correct the generated rating is based on the algorithm. In the project our feature set number was 1000 and we got an accuracy of 83% where the training and testing set splitting ratio is 80:20.

E. Sentiment Analysis on Twitter

We have used Scikit learn library and NLTK platform in Python for sentiment analysis. For sentiment analysis on Twitter we used the Naïve Bayes classifier built in to SK learn library. Amazon book review dataset was used for training the [classifier and then applied on the pre-processed Tweet data to generate sentiment polarity.

F. Rating Generation

For generating rating we took the polarity values given by the classifier. Then we calculated the rating based on the polarity values, where we took the polarity of the sentiments and doing arithmetic mean and scaling the values on scale of 1 to 5 to show the rating.

IV. RESULT AND DISCUSSION

Using Naïve Byes classifier we classified the tweets and using the polarity ratings were generated for each attributes.

In Fig. 1 we show the number of good and bad reviews based on our sentiment analysis of the book “The Great Gatsby”.

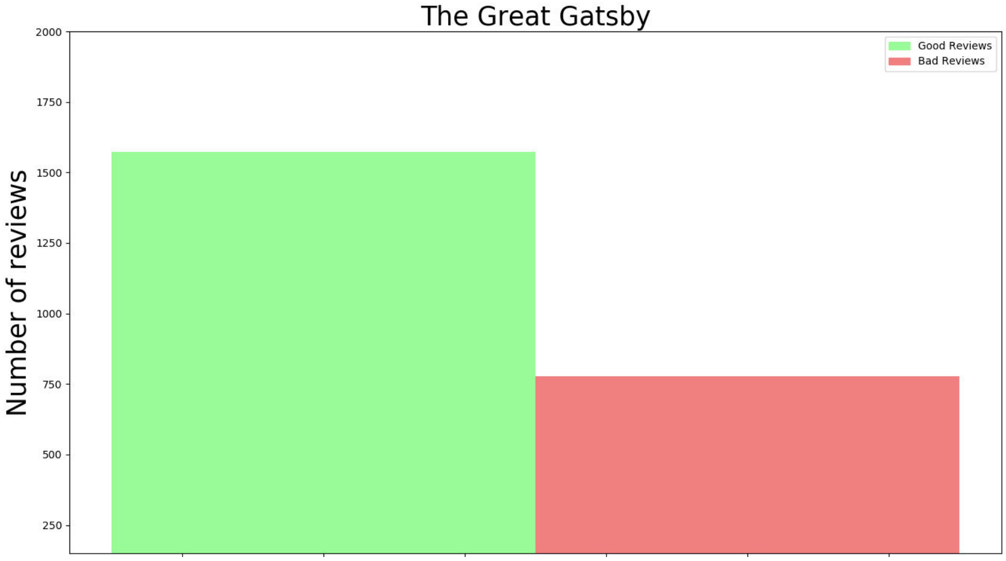

Fig. 1: The Great Gatsby Polarity

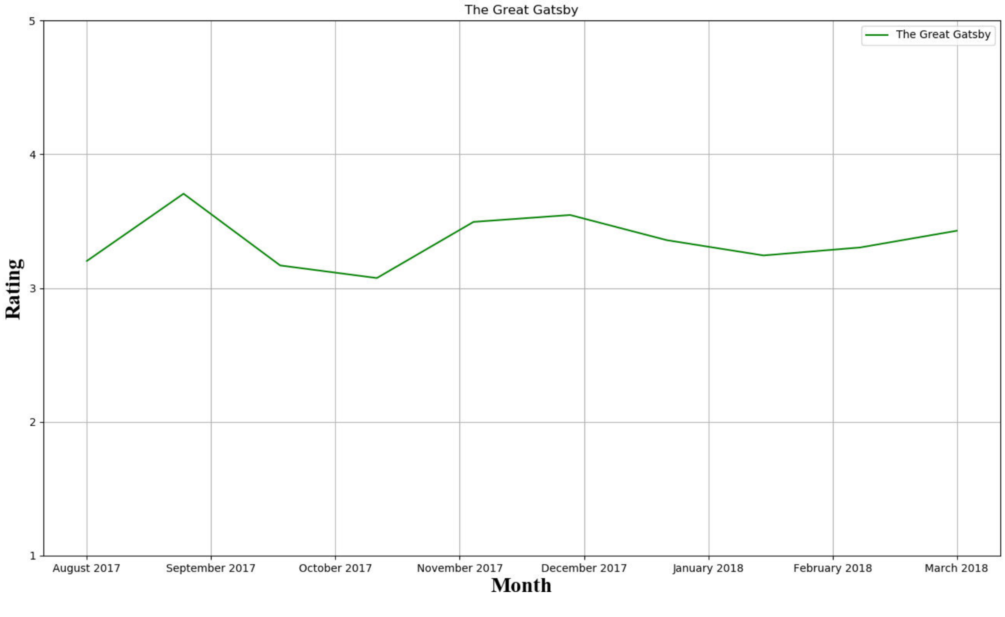

Fig. 2: The Great Gatsby Rating vs. time graph

Fig.2 is a time vs. rating graph for the book “The Great Gatsby”. In the graph x-axis is the time and y-axis is the generated rating. We averaged the polarities of the data from one month and input the result with respect to the month. This graph shows that ratings generated by users’ opinion are varying from month to month.

In Fig. 3 we show the number of good and bad reviews based on our sentiment analysis of each attribute for Little Fires Everywhere.

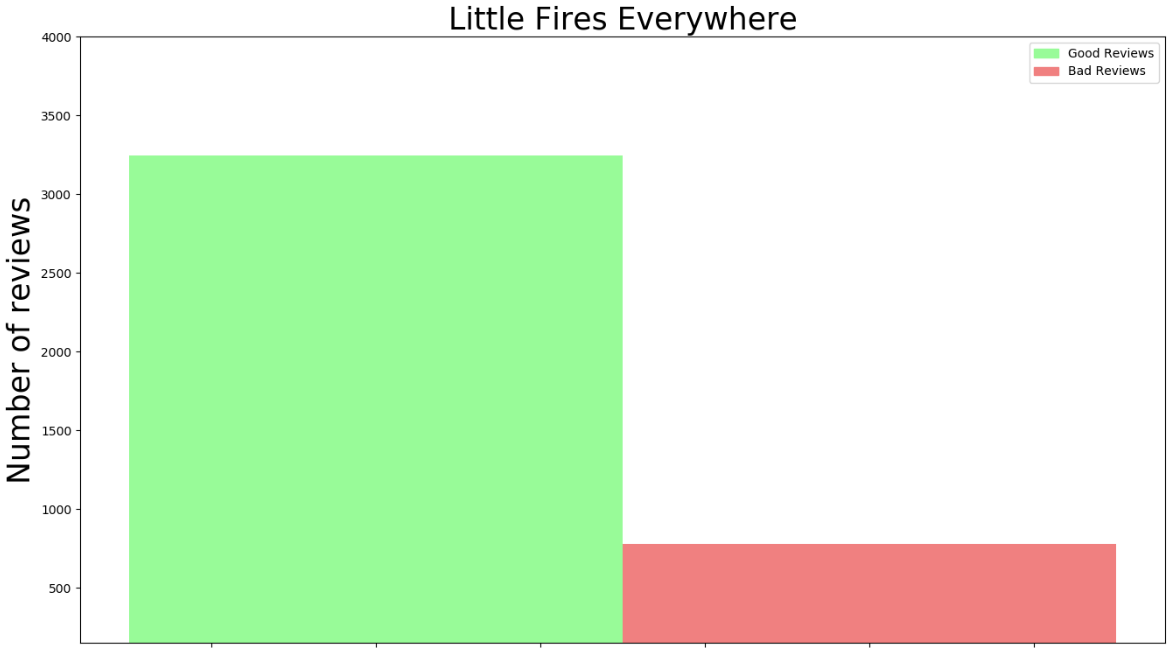

Fig. 3: Little Fires Everywhere Polarity

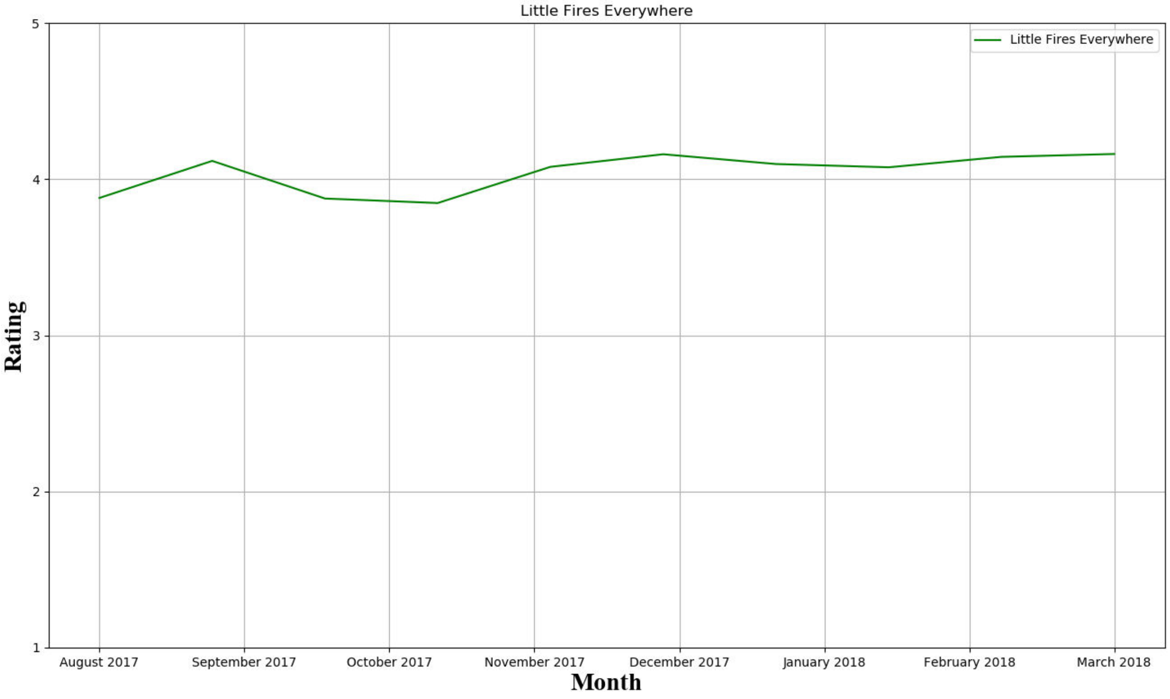

Fig. 4: Little Fires Everywhere vs. time graph

Fig.4 is a time vs. rating graph for the book “Little Fires Everywhere”. In the graph x-axis is the time and y-axis is the generated rating. We averaged the polarities of the data from one month and input the result with respect to the month. This graph shows that ratings generated by users’ opinion are varying from month to month.

V. CONCLUSION AND FUTURE WORK

In this project, we have proposed a model to generate rating of books based on public opinion. Rating generated by our proposed model reflects public's opinion more accurately than a general rating on a book reviewing website. This model will allow users to analyze numbers of books along with their attributes based on their rating and choose the best one depending on user's preferred genre and time. In the future we want to improve this model and bring to a stage where it not only works flawlessly with book reviews but works with other art form reviews for example movies, TV shows, games. Along with that we also want to provide a way to make our model more user friendly.

REFERENCES

1. S. K. Khatri and A. Srivastava, "Using sentimental analysis in prediction of stock market investment," 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016.
2. M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 04, 2004.
3. E. J. D. Fortuny, T. D. Smedt, D. Martens, and W. Daelemans, "Media coverage in times of political crisis: A text mining approach," Expert Systems with Applications, vol. 39, no. 14, pp, 2012.
4. A. Dalmia, M. Gupta, and V. Varma, "IIIT-H at SemEval 2015: Twitter Sentiment Analysis – The Good, the Bad and the Neutral!," Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015.
5. B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012.
6. O. Abdelwahab, M. Bahgat, C. J. Lowrance, and A. Elmaghraby, "Effect of training set size on SVM and Naïve Bayes for Twitter sentiment analysis," 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2015.
7. M. Mhatre, D. Phondekar, P. Kadam, A. Chawathe, and K. Ghag, "Dimensionality reduction for sentiment analysis using pre-processing techniques," 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017.
8. J. Prusa, T. M. Khoshgoftaar, and N. Seliya, "The Effect of Dataset Size on Training Tweet Sentiment [classifier," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015.
9. K. L. S. Kumar, J. Desai, and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016.
10. McAuley, J. (n.d.). Amazon product data. Retrieved March 24, 2018, from http://jmcauley.ucsd.edu/data/amazon/