Sentiment Analysis of Amazon Products Using Supervised Learning Algorithm

1st Khadija Ilyas
Riphah Institute of Computing and Applied Sciences
Riphah International University,Lahore
ikhadija645@gmail.com

2nd Sania Sagheer
Computer Science Department,
COMSATS University Islamabad
saniasagheer89@gmail.com

3rd Muhammad Furqan Ahmad
Computer Science Department,
University of Lahore
furqanahmad54@outlook.com

***Abstract*—Nowadays the world is becoming more digitalized and e-commerce is dominating by making desired products available within the reach of the customers.So the customers can purchase products from their homes.It is very important to know what people reviews are about products especially for buyers as it creates satisfactions level before purchasing product.A customer needs to go through thousands of reviews to understand and to determine the effectiveness of the product for selection. With the help of machine learning, going through thousands of reviews would be much easier if a model is used to polarize those reviews and learn from it. We used supervised learning method on a large amazon dataset to polarize it and obtained satisfactory accuracy.**

Index Terms—NLTK ,polarity, sentiment analysis, amazon.

1. INTRODUCTION

Data analysis and analytic has enabled to reveal the hidden facts and patterns about the big data. A large volume, variety and velocity of data is called big data. Businesses are having huge amount data on their e-commerce sites, which is exceeding rapidly every day and that results in increasing the capacity of IT departments especially in the fields of business. The data contains other two V's which plays an important role for the businesses which are Veracity and Value. E-commerce data consist of the reviews of the customers which can be used to determine the behavior of the consumers.

Behaviors are considered as the most reliable source to make decisions on the future trends of the data. In the E- commerce dataset could vary from structured to unstructured dataset. The relevant data is filtered and the useful information is fetched from insights of the data. The insights focused on the e-commerce are future trends planning and forecasting, marketing opportunities and getting patterns of consumer behavior etc.

Sentiment Analysis determines computationally whether a piece of word is positive, negative or neutral. It is also known as opinion mining.

A) Why sentiment analysis?

It is used by businesses in the marketing sector to improve their strategies, to understand the feelings of customers about goods or brands, how people react to their advertisements or product releases and why consumers do not purchase such items.The use of sentiment analysis helps in political field to keep track of political views, to detect consistency and incoherence between statements and actions at the level of government.

NLTK is capable of textual tokenization, parsing, stemming, semantic reasoning, classification, tagging and other computing linguistics.NLTK comes with "nltk.sentiment.vaderan" that is built in sentiment analyzer module which can analyze a piece of text and classify the sentences under positive, negative and neutral polarity of sentiments. The advantage of this approach is that sentences containing negated positive words (e.g. "not good", "not happy") will still receive a negative sentence sentiment. Some simpler tools of sentiment analysis will just take the average of the sentiments of the words and would miss subtle details like this.

B) Description of Data

Data has been chosen from Amazon website. Data consist of two portions, training set data and test set data. Both portions have 20,000 reviews about android applications of different mobile companies.Both datasets have different type of reviews and product IDs. For both datasets, each review includes three information; Semantic Value, Product ID, and Review Text. Semantics values ranges from1 to 3. If review is negative,vlaue is 1.If review is neutral,value is 2 and if review is positive,value is 3. In test dataset, we have to filter reviews of nine product IDs that belongs to three companies; three products per company. Based upon the reviews, we have to recommend among three companies ,which company is better to invest in.

2. LITERATURE REVIEW

One fundamental issue in the study of sentiment is the categorization of polarity of sentiment [1-4].The problem is to categorize the text into one particular polarity of sentiment,either positive or negative (or neutral).There are three levels of sentiment polarity categorization based on the scope of text i.e document based,sentence and entity based [5].Analysis of views and opinions[6-12]studies the feelings of people towards certain entities. Internet is a resourceful place for information about sentiments.

In 2018,KTH student[13]found SVM performance for a large volume of data is better than NB in sentiment analysis. Amazon products categories including electronics,house-hold[14]with three different classifiers: Naive Base, SVM and Random Forest. They concluded that Random Forest has usually given results more reliable.They found also that SVM performed better in larger data sets, than Naive Base. Different feature extraction techniques for sentiment analysis are performed in paper [15].

Amazon data set is collected at first and prepossessing is performed for stop words and also to remove special characters. Naive Bayes was used as classifier. They noted that Naive Bayes is generating better phrase level results than single or multi-words.From a user's perspective, people can post their own content through different social media, such as blogs ,forums or online social networking sites.
From the perspective of a researcher, many social media sites are releasing their application programming interfaces , prompting researchers and developers to collect and analyse the data.The sentiment analysis was performed of reviews of Amazon beauty products.

3. METHOD and MATERIALS

The data chosen consist of different types of data from nominal to the text format. In order to extract the useful information out of the data Python development environment has been used to deduce useful results in the form of different graphs patterns. Initially the random data has to be made in a form where it can be used. The cleaning of data has been done and a Data Frame is created. By looking insights to the data following tasks has been done in cleaning the data and making it in a usable form:

i. Loading datasets in .txt format.
ii. Extracting reviews and semantic values.
iii. Creating a DataFrame.
iv. Applying a semantic analyzer.
v. Creating binary labels for supervised learning.

After creating Data Frame and arranging semantics values & reviews in two different columns, sentiment analyzer algorithm has been applied. It gives four useful values for each review; polarity value, positive value, negative value and a neutral value. These four values have been further used as features for creating a binary label.
We have applied five different supervised learning classifiers to train the given training dataset, the results of which are mentioned below:

Table 1. Logistic Regression Classifier Applied on Both Test and Train Datasets

Train Set Accuracy	0.765120463860
Train Set ROC	0.6946836826242029
Train Set Recall	0.78780807103420
Train Set Precision	0.93444504455

Table 2. Neural Networks Classifier Applied on Both Test and Train Datasets

Train Set Accuracy	0.766669999000
Train Set ROC	0.6921625958463619
Train Set Recall	0.80104463437796
Train Set Precision	0.91115311909

Table 3. Random Forrest (Specialized) Classifier applied on both test and train datasets

Train Set Accuracy	0.962061381585
Train Set ROC	0.9677760422072779
Train Set Recall	0.95712705744583
Train Set Precision	0.99324871725

Table 4. K-Nearest Neighbour Classifier Applied on Both Test and Train Datasets

Train Set Accuracy	0.769619114265
Train Set ROC	0.69824294184374
Train Set Recall	0.80124003542958
Train Set Precision	0.91608155549

Table 5.Support Vector Machine (SVM)Classifier applied on both test and train datasets

Train Set Accuracy	0.764420673797
Train Set ROC	0.6969507578619044
Train Set Recall	0.78312307261003
Train Set Precision	0.94295166081

4. RESULTS AND DISCUSSIONS

The first attempt to understand this data is to measure score of the sentiment of each review, positive or negative. Further polarity, positive, negative, neutral count will be focused.

Neutral/Negative/Positive Score: Indicates the potency of these classes between 0 and 1.Polarity Score: Measures the difference between the Positive/Neutral/Negative values, where a positive number closer to 1 indicates overwhelming positive and a negative number closer to -1 indicates overwhelming negativity.From the results, It has been seen that most of the reviews got the positive sentiments same as the binary label. But on the contrast to the binary label, it has also been seen that, each positive review has high polarity value while low negative and positive values.However positive values are bit higher than negative values. The reviews with positive sentiment polarity score have increasing occurrence as the rating goes high. But for negative and neutral polarity score, the binary label is 0. As mentioned above, we calculated accuracy, ROC value, Recall and Precision for each classifier to make sure the best performed classifier is to be used for test dataset.

For test dataset, raw data includes many android application products from various companies. But task given requires to evaluate three companies based on the reviews of three android products from each company and recommend the most suitable company for investment.

So we have first filtered out all reviews related to these nine android products and made data-frame of each company separately. Each data frame includes semantic value and review text itself. We applied the same sentiment analyzer on the test dataset and generated four columns of polarity value, positive value, negative value and neutral value against each review.

Next,We have applied same selected classifier(Random Forrest) to predict binary label for each company. Once we got binary labels for three android products of all three companies, we just calculated the percentage of positive reviews. Since task requires to nominate the best company for investment, it's a non-statistical problem. We only sorted the best company on number of positive reviews ignoring neutral and negative reviews.

With "ADD_1" company has highest percentage of positive reviews for its products, it is recommended as the most suitable company to be invested in.

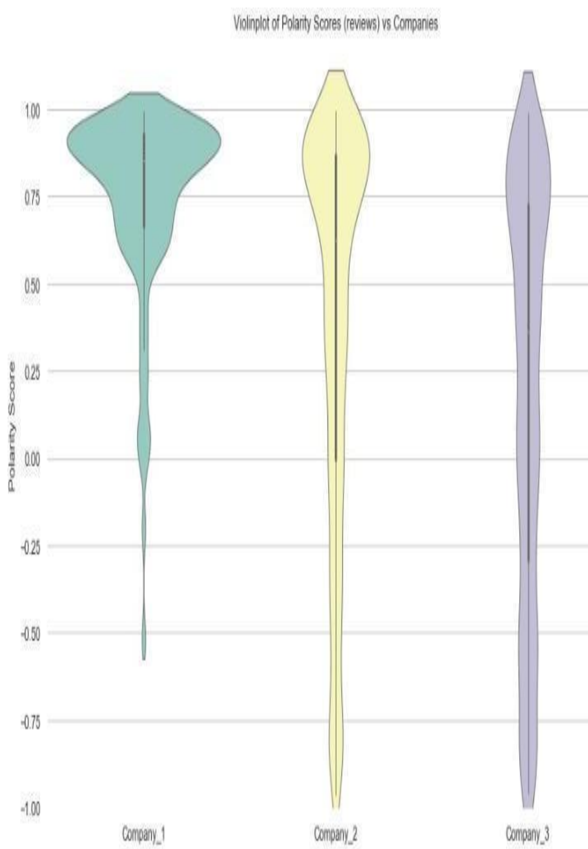

Fig.1 displays violin-plot of polarity(reviews) vs three companies

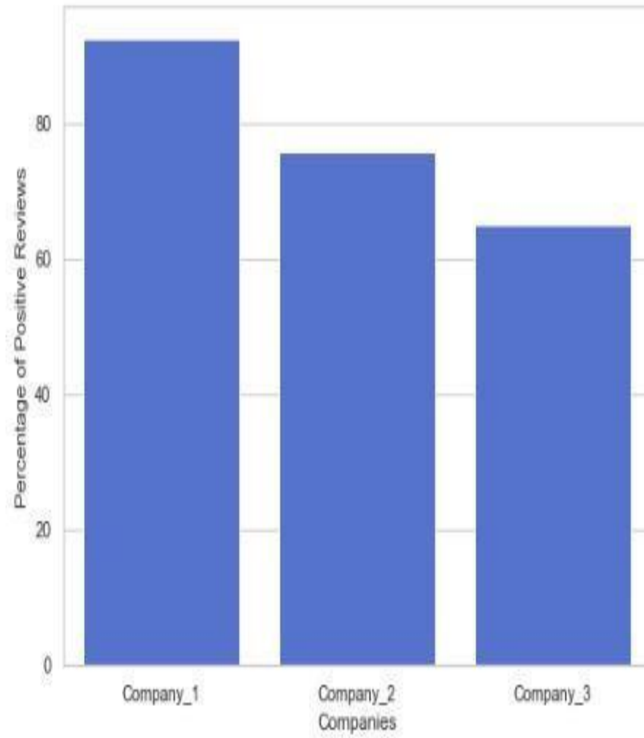

Fig.2 displays percentage of positive reviews of three companies

5. CONCLUSION

With the advancement of technological change, business processes specially e-commerce has been shifted from manual to internet-based solutions in order to increase the recurrent revenue. This paper focuses on the sentiment analysis of sales data of an e-commerce store in order to identify the hidden patterns of purchase history and to predict the scenarios where user preferred to recommend or not recommend the products. We found that "ADD_1" company has three android products that have most positive reviews and are popular. Similarly, if a customer is giving low rating to a product and never recommended a certain item; actually, the customer is having complaints about that particular product. It turns out that, Random Forrest technique has the highest accuracy and precision value in predicting whether the product will have positive review or not based on customers reviews. With Vader Sentiment anyway, it was difficult to understand the exact meaning of the positive, neutral, and negative breakout values. The compound score is useful, but more research is required in future.

REFERENCES

I. Chesley, P., Vincent, B., Xu, L., & Srihari, R. K. (2006). Using verbs and adjectives to automatically classify blog sentiment. *Training*, *580*(263), 233

II. Koehn, P., & Mihalcea, R. (2009, August). Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

III. Xing, C., Crestani, F., & Rauber, A. (Eds.). (2011). *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation: 13th International Conference on Asia-Pacific Digital Libraries, ICADL 2011, Beijing, China, October 24-27, 2011, Proceedings* (Vol. 7008). Springer Science & Business Media.

IV. Ni, C. X., Cresta, F., & Rauber, A. (2011). Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation.

V. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

VI. Barry, S., & Fellbaum, C. (2004). Proceedings of Coling: The 20th International Conference on Computational Linguistics.

VII. Indurkhya, N., & Damerau, F. J. (2010). *Handbook of natural language processing*. Chapman and Hall/CRC.

VIII. Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. WWW'05: Proceedings of the 14th International Conference on World Wide Web.

IX. Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining In: Proceedings of the Seventh conference on International Language Resources and Evaluation. *European Languages Resources Association, Valletta, Malta*.

X. Kantardzic, M. (2002). Data Mining: Concepts, Models, Methods, and Algorithms, ACL'04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. *Wilet-IEEE Express*.

XI. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1-135.

XII. Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis, in ‘CIKM’05: Proceedings of the 14th ACM international conference on Information and knowledge management’.

XIII. Paknejad, S. (2018). Sentiment classification on Amazon reviews using machine learning approaches.

XIV. Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1), 5.

XV. Shaikh, T., & Deshpande, D. (2016). Feature selection methods in sentiment analysis and sentiment classificationof amazon product reviews. *Int J Comput Trends Technol*, *36*(4), 225-230.