Analysis of Amazon Cell Phone Reviews Using IBM Cloud

SMARTINTERNZ GUIDED PROJECT

By

Lokesh kumar. P -**18MIS0194**

Rohit kumar.KA-**18MIS0166**

Mugunthan.B-**18MIS0173**

Munagala Lakshmi Srikanta - **18BSW0022**

Index

- ➢ Acknowledgement **2**
- ➢ Introduction **3**
  - Problem Statement
  - Solution
- ➢ Literature Survey 5
- ➢ Experimental Investigations **8**
- ➢ Analysis **9**
  - Hardware and software used
  - Architecture
  - Methodology
- ➢ Scenario **12**
- ➢ Flow Chart **12**
- ➢ Conclusion **13**
- ➢ Limitations and Future Scope **13**
- ➢ References **14**
  - Articles
  - Bibliography

Acknowledgement

This project consumed a certain amount of work, research and dedication. Still, implementation would not have been possible if we did not have the support of many individuals including our mentor as well as our team leader. Therefore, we would like to extend our sincere gratitude to all of them.

First of all, we are thankful to smartinternz and Prof.**Pradeepthi Duggaraju** for her support and for providing necessary guidance concerning the implementation of the project.

Without the superior knowledge and experience of the staff, the project would have lacked in quality of outcomes, and thus her support has been very essential towards this work. We are very grateful to Vellore Institute of Technology for allowing us to register for this course and pursue this project, which helped us gain great knowledge towards our field and gave us quite an experience that will be useful to our work in the future.

Nevertheless, we express our gratitude towards our colleagues for their kind cooperation, constant support and encouragement which helped us in the completion of this project.

Lokesh kumar. P -**18MIS0194**

Rohit kumar.KA-**18MIS0166**

Mugunthan.B-**18MIS0173**

Munagala Lakshmi Srikanta - **18BSW0022**

Introduction

Traditional methods of searching and buying needed a revamp to ensure more security by detecting factors which even the humans may miss. The project is about simplifying hours of search needed by multiple personnel rewinding and forwarding reviews to find a small detail, which often gets also missed most of the time. Having automation in this field will greatly improve the efficiency of the search while decreasing the human error factor. The utility value to the society can be measured by the increase in safety and the decrease in effort which is required to produce that level of safety and hence the overall impact is very high if the project is implemented properly after optimizing and eliminating the errors using higher grade machinery for computation.

Problem statement :

Mobile phones have revolutionized the way we purchase products online, making all the information available at our fingertips. As the access to information becomes easier, more and more consumers will seek product information from other consumers apart from the information provided by the seller. Reviews and ratings submitted by consumers are examples of such of type of information and they have already become an integral part of customer's buying-decision process. The review and ratings platform provided by e-Commerce players creates transparent system for consumers to take informed decision and feel confident about it.

Solution:

It is difficult to read all the feedback for a particular item especially for the popular items with many comments. In this project, we will attempt to understand the factors that contribute to classifying reviews as positive or negative

We will be using Natural language processing to analyze the sentiment (positive or a negative) of the given review.After building our model, we saved the model and bag of count vector file which we will be loading it in flask to predict the output on the UI page. A sample web application is integrated to the model built which is called as flask.Flask is an API of Python that allows us to build up web-applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web-Application. A Web-Application Framework or Web Framework is the collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc. Flask is based on WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine. Then we deployed our project in IBM cloud. IBM Cloud is a suite of cloud computing services from IBM that offers both platform as a service (PaaS) and infrastructure as a service (IaaS).

With IBM Cloud IaaS, organizations can deploy and access virtualized IT resources -- such as compute power, storage and networking -- over the internet. For compute, organizations can choose between bare-metal or virtual servers.

Literature Survey

Sl no	Title	Citation	Author	Methodology and Discussion	Conclusion
1	Bridging the quantitative-qualitative divide: the lexical approach to textual data analysis,	R. Bolden and J. Moscarola, "Bridging the quantitative-qualitative divide: the lexical approach to textual data analysis," Social science computer review, vol. 18, no. 4, pp. 450–460, 2000.	R.Bolden and J.Moscarlo.	It mitigates the problem of huge calculation of word frequencies. First, the computer generates a lexicon of all words (graph forms) present in the text (corpus). By listing these in decreasing order of frequency and applying filters (e.g., removing tool words), we can quickly gain an idea of the main content. Further manipulation of the lexicon permits the grouping of similar terms (manually or automatically), identification of context (e.g., repeated expressions and relative lexicons), and the progressive isolation of terms key to the investigation.	Paper gives an overview of lexical analysis software that offers a means of exploring data in which the time invested can be adjusted according to the objectives of the research, analyses can easily be made between open- and closed-response variables in the study, and a new type of information can be examined: the structural aspects of language

2	Monitoring technological and bibliographical information systems via textual data analysis	Moscarola, J. (1998). Monitoring technological and bibliographical information systems via textual data analysis: The LEXICA system. In Proceedings of the 4th international conference on Current Research Information Systems (CRIS). Luxembourg: European Commission [On-line]. Available: http://www.cordis.lu/ cybercafe/src/moscarola.htm. Moscarola, J., & Bolden, R. (1998).	Moscarola, J., & Bolden, R	Having identified keywords, the lexicon permits a selective return to the original text. Such a return is useful in determining the contextual meaning of words and in ensuring that the researcher remains close to the raw data. The selective return to the text can be done by a variety of means,	This system could integrate the hardware data to text communication developed with node.js. Which sends the real time data to a csv By which it provides the system service in a low cost and easily accessible manner.

3	A neural word embeddings approach for multi-domain sentiment analysis	M. Dragoni and G. Petrucci, "A neural word embeddings approach for multi-domain sentiment analysis," IEEE Transactions on Affective Computing, vol. 8, no. 4, pp. 457–470, 2017.	M. Dragoni and G. Petrucci	It provides a methodology and system to Experiments on multi-domain sentiment classification and cross-domain sentiment classification on 16 different domains to demonstrate the effectiveness and advantages of the proposed model..	This system solves the problem of a novel completely-shared neural model to make use of different training data across all domains. Our model builds domain-aware word embedding to express domain and context information for words, and proposes domain-aware attention mechanisms to focus on more significant words in the text.
4	Comparative study of machine learning approaches for amazon reviews	A. S. Rathor, A. Agarwal, and P. Dimri, "Comparative study of machine learning approaches for amazon reviews," Procedia computer science, vol. 132, pp. 1552–1561, 2018.	A. S. Rathor, A. Agarwal, and P. Dimri	Proposed a model for the machine learning method, SVM can be regarded as the baseline learning method for Amazon reviews as it has highest accuracy. Applying weight to the reviews has increased the accuracy, so we can say that "more precise the data more exact results can be obtained".	The results show that user reviews are very important for the decision making of customers. But the more surprising and interesting part is to bifurcate reviews into positive, negative and neutral. After comparing three machine learning approaches using unigrams features and weighted unigram features in Amazon reviews the accuracy of classifiers are identical

5	Natural language processing machine classifiers," Neural processing letters	J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," Neural processing letters, vol. 9, no. 3, pp. 293–300, 1999.		Neural Processing Letters promotes fast exchange of the current state-of-the art contributions among the artificial neural network on various aspects of artificial neural networks and machine learning systems	Dimensionality reduction, missing data imputation, dictionary learning, data retrieval, feature extraction Classification, regression,clustering for big data and Data pre-processing for multi-task or multi-view data

Experimental Investigations

This project addresses the biggest issue with the present search methods is that they lack a detection feature. Of all the research papers cited in our survey, there are plenty of comparisons on which could be a better search software, which dataset will yield the most effective output, which algorithm can give better accuracy, but the most important limitation of these researches is that they need made their standard input as a program dataset, during a real-world scenario, that's not the case.

But new reviews come in different languages and different expressions. So, combining the detection algorithm, which takes the input ,trains then using NLP not only makes it very accurate (due to the mixture of accuracies of algorithms) but also makes it rather more realistic for implementation on a larger scale.

Analysis

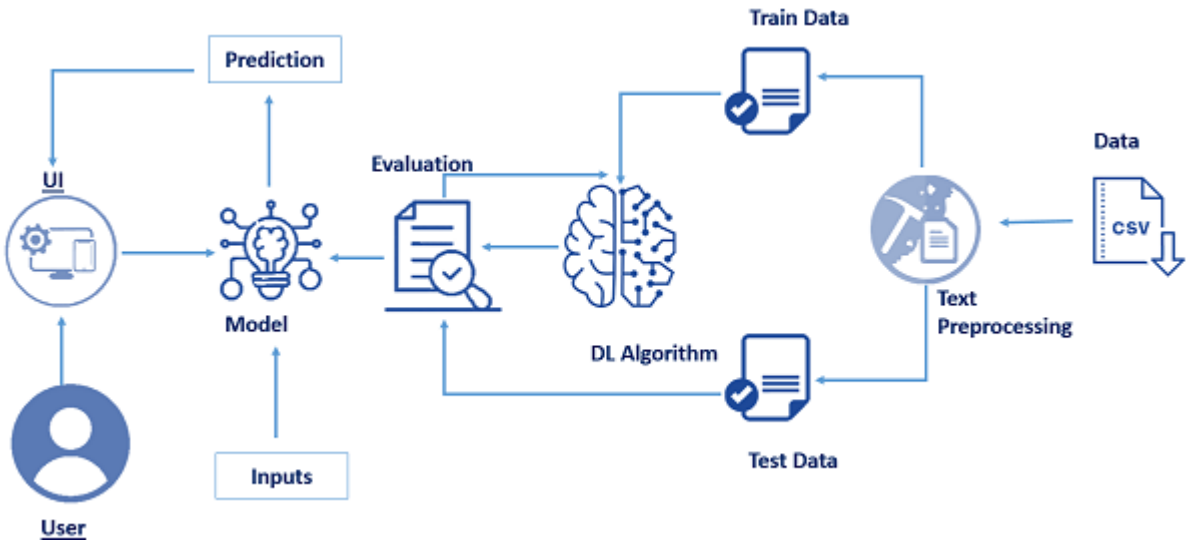
This project contains the following Activities

- Data Collection.
  - Collect the dataset or Create the dataset
- Text Preprocessing.
  - Import the Libraries.
  - Importing the dataset.
  - Remove Punctuations
  - Convert each word into a lower case.
  - Stemming.
  - Splitting Data into Train and Test.
- Model Building
  - Import the model building Libraries
  - Initializing the model
  - Adding Input Layer
  - Adding Hidden Layer
  - Adding Output Layer
  - Configure the Learning Process
  - Training and testing the model
  - Optimize the Model
  - Save the Model
- Application Building
- Create an HTML file
- Build a Python Code

Hardware and Software Specifications

- Spyder IDE
- Tensorflow-Backend for Keras to build neural layers
- Keras- This package is used for training the model and building layers
- Flask-This package is used for building User Interface(UI)
- Anaconda
- Google colab
- IBM Cloud
- GPU NVIDIA GTX 1080 Ti

Architecture



The outputs from the machine learning model are stored in the csv file as collections.

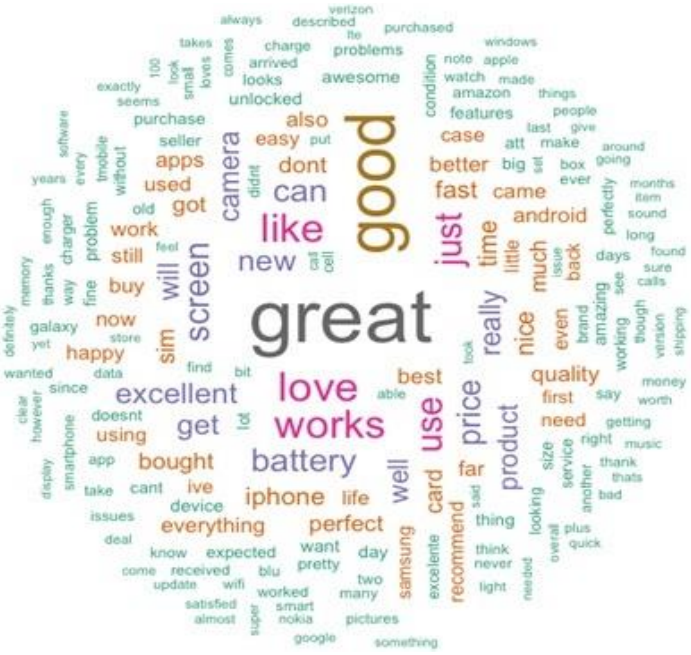
Methodology

Amazon.com is a treasure trove of product reviews and their review system is accessible across all channels presenting reviews in an easy-to-use format. The product reviewer submits a rating on a scale of 1 to 5 and provides own viewpoint according to the whole experience. The mean value is calculated from all the ratings to arrive at the final product rating. Others can also mark yes or no to a review depending on its helpfulness – adding credibility to the review and reviewer. In this study, we analysed more than 400 thousand reviews in the Dataset.

WORD CLOUD

We segregated the reviews according to their ratings – positive reviews (4 or 5 star) and negative reviews (1 or 2 star). In both type of reviews there are certain common words like ‘‘work’’, ‘‘battery’’ and ‘‘screen’’. The most frequently used words in positive reviews are: ‘‘great’’, ‘‘good’’, ‘‘camera’’, ‘‘price’’, ‘‘excellent’’, etc. In case of negative reviews words such as ‘‘return’’, ‘‘back’’, ‘‘problem’’, ‘‘charge’’ are prevalent.

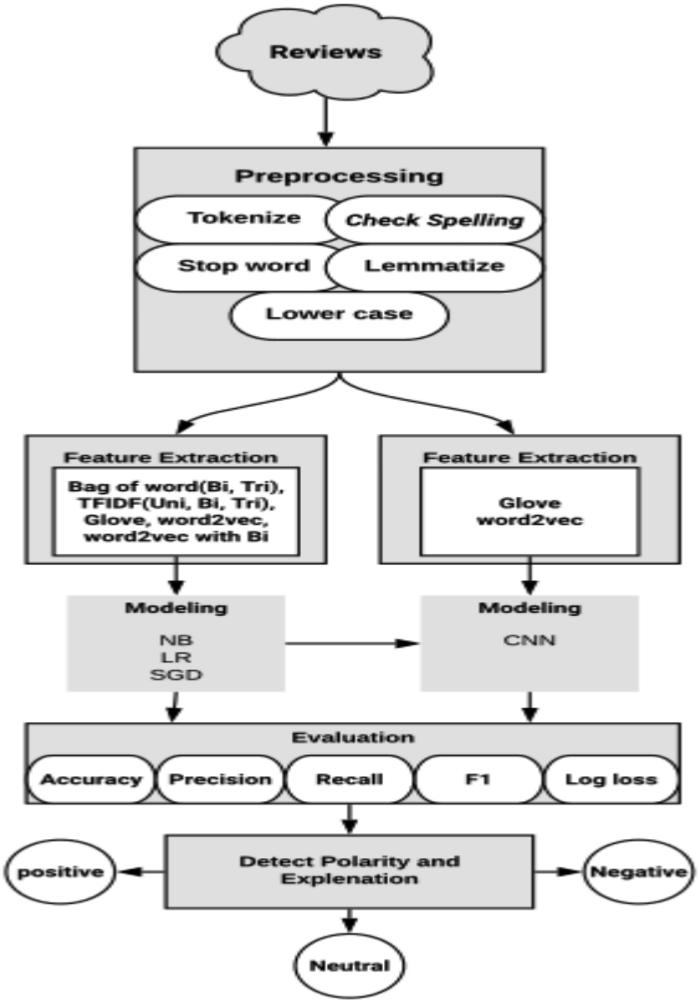
Figure: Word cloud of positive reviews



Scenario

Assume the person was searching for a product at a point . When the analysis program is activated after the detection, it runs the trained data model and if the person is reviewing using the same words which were detected earlier, then the word are replaced removing the barrier between human and computer using NLP.

Flow Chart



Conclusion

Amazon's product review platform shows that most of the reviewers have given 4-star and 3-star ratings to unlocked mobile phones. The average length of the reviews comes close to 230 characters. We also uncovered that lengthier reviews tend to be more helpful and there is a positive correlation between price & rating. Sentiment analysis shows that positive sentiment is prevalent among the reviews and in terms of emotions, 'trust', 'anticipation' and 'joy' have highest scores.

It'd be interesting to perform further analysis based on the brand (example: Samsung vs. Apple). We can also look at building a model to predict the helpfulness of the review and the rating based on the review text. Corpus-based and knowledge-based methods can be used to determine the semantic similarity of review text. There are many more insights to be unveiled from the Amazon reviews.

Limitations and Future Scope

In this study, we implemented four types of algorithms with a variety of feature extraction. Some algorithms that remain to be applied in future work include LSTM, KNN, and Maximum entropy. Then, we will compare the result to the result we performed in this project. Our research has some limitations: NLP is relatively a new topic, and highly advanced; hence, it needs a lot of research to understand the field and how it works. Furthermore, we faced some problems with computer memory causing experiments to be highly time consuming. We also used Google Colab to increase the performance, but it did not give us the expected speed.

References

Articles

- https://www.kdnuggets.com/2017/01/data-mining-amazon-mobile-phone-reviews-interesting-insights.html
- https://www.kaggle.com/keras/NLP
- https://www.aitribune.com/dataset/2018051063
- https://deepai.org/dataset/market-1501
- https://numpy.org/.
- https://en.wikipedia.org/wiki/Matplotlib#:~:text=Pyplot%20is%20a%20Matplotlib%20module,being%20free%20and%20open%2Dsource.
- https://www.geeksforgeeks.org/os-module-python-examples/#:~:text=The%20OS%20module%20in%20python,using%20operating%20system%20dependent%20functionality.
- https://cocodataset.org/#home

Bibliography

[1] D. Li, X. Chen, Z. Zhang and K. Huang, "Learning Deep Context Features over testing and training," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 7398-7407, doi:10.1109/CVPR. 20 17.782

[2] Jie, Zhu & Fei, Wang & Zhi- wei, Zhang & Weinong, Wang & Kang- long, Zhang. (2018),"A Multi-loss Deep Network Based on Pyramid for training data in NLP",. 1552-1557. 10.1109/CAC.201 8.8623067

[3] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, Chi Zhang, " Pattern Recognition of words", Volume 94, 2019, ISSN 0031-3203, https://doi.org/1 0.1016/j.patcog.2019.05.028

[4] Kaiyang Zhou, Tao Xiang, "Keras: Natural language processing in Python",2019, 1910.10093, https://dblp.org/ rec/journals/cor r/abs-1910- 10093.bib

[5] Omar A., Khalid J, "Design and Implementation of an Online Location Based Services using Flask", 2014 International Journal of Computer