Recommendation System Based on Text Analysis

Voggu Suman Venkata Sai, Yuvraj Singh Champawat, B.K Tripathy

Abstract*: *Recommendation Systems have gathered a lot of attention from the research community following the introduction of internet. Internet provided platform for development and deployment of web, mobile and desktop-based applications. The overall penetration of the internet has increased across the globe over the last two decades which in turn provides more customers for the tech companies. In this project, we are mostly focussing on E-commerce companies like Amazon. It is never easy to find a product that has all the features you need. We have developed a model that can be used to assist the customers in choosing the best product available that has features as specified by the customer. This model will list down all the products that have that feature. Additionally, it will also provide a feedback to the manufacturer of the product regarding the features that did not impress most of the customers. The manufacturer can work on these features and improve them when launching upgraded versions or new products in the same category. The core idea of this project is to analyse the product reviews given by existing customer to assist a new customer in choosing the best product having the feature as specified by the customer.

Index Terms*: *Community detection, Social media, Recommendation system, E-commerce, Text processing, Stemming, Lemmatization, Reviews

I. INTRODUCTION

The web has flourished rather exponentially following the creation of World Wide Web (WWW) in 1989 by Tim-Berners-Lee. It gave rise to a new set of companies that generated income by deploying their Applications on the web. These are now known as "Tech-giants". It includes Google, PayPal, Microsoft, Amazon and many more. Most of these companies have developed algorithms to assist customers in buying their products. These algorithms are known as "Recommendation Systems". The reason why they use such algorithms varies from company to company. Some companies like YouTube, Facebook use it to improve user experiences for customer retention while others like Amazon, eBay use it to make the customer to spend more money by suggesting related products. The main idea behind deploying such algorithms is mostly to retain more customers.

In this project, we have focussed our study on E-commerce website Amazon. Here, we will use the dataset that consists of data collected by using web-scrapping on customer reviews for three products. Most of the data collected was noisy and had missing values. Noise in data can be referred to as any meaningless information that unnecessarily increases the skewness and corruption in data. We pre-processing of this data. Pre-processing basically refers to the techniques applied to convert the data into a form that can be understood by the computer. It becomes very difficult for a system to understand and interpret noisy data correctly. Therefore, it is very important to handle or remove the noise from the data as it may adversely affect the results of data analysis and skew the conclusions. There are many different types of noises that can be present in text for e.g. missing values, incomplete, outliers etc. We can use a number of techniques such as binning, clustering or regression to handle noisy data.

	ratings_name	Reviews
0	2.0 out of 5 stars	This is a really worst phone
1	2.0 out of 5 stars	Facing problem with battery drainage
2	2.0 out of 5 stars	Battery backup is okay, picture quality is bri...
3	1.0 out of 5 stars	Not at all a good phone . Battery lags
4	2.0 out of 5 stars	worst picture and sound quality.processor is v...

Fig 1: The dataset

In this project, we are applying the concept of Natural Language Processing (NLP). NLP is a part of deep learning that mainly focuses on training models on textual data. NLP can be defined as a sub-field of machine learning that focuses on text processing to understand or infer what the text means. We know that machines will never understand the words directly but the idea is to understand the relation between words. Python provides support for NLP inform of NLTK library. It is a very wide range of applications like text-classification, named-entity-recognition, parts of speech tagging, next word predictions (as seen in google, Qwitky keyboard in smartphones). Almost 80 percent of data available on the internet is in text format. Images, audios videos etc occupy the remaining share. It is worth noting that approximately 2.5 EB of data is generated every day. It is being generated as we speak. With such a huge amount of information, it becomes increasingly difficult to process this data. Text processing can be defined as techniques that are applied to convert the data into a form that is understood by the system. One of the techniques that can be applied to pre-process text is stop-words removal. The dataset we collected had customer reviews for three different products. These reviews are simple text and so had a lot of stop-words in them which needed to be removed for further processing. We tokenized each of these reviews and checked if they were present in Stop-words package of nltk.corpus library. We removed all the tokens that present in Stop-words. At the end of this process, we formed a new list consisting only stopwords-free reviews.

Revised Manuscript Received on July 09, 2019

Voggu Suman Venkata Sai, SCOPE, Vellore Institute of Technology, Vellore, India.

Yuvraj Singh Champawat, SCOPE, Vellore Institute of Technology, Vellore, India.

B K Tripathy, SITE, Vellore Institute of Technology, Vellore, India

Another method that can be used for text processing is Text normalization. Text Normalization can be defined as a method to prepare words, sentences, text and documents for further processing. There are many techniques that can be applied in order to normalize the text but we have focussed on the following two: Stemming and lemmatization.

There are a number of different languages a person can speak. Most of these languages contains some words that are derived from another word. The language that consists of different versions of same word is known as inflected language. In grammar, we use different versions of the same word with slight variations by adding some prefix, infix or suffixes to express different grammatical categories such as gender, tense, mood, voice and count. For e.g.

(sit, sat, sitting, seated) -> sit.

As shown in the above example, sit, sat, sitting, seated are inflections or variation of the same root 'sit'. All of them have the same meaning but are used differently based on the context of text. Stemming and Lemmatization are two text normalization methods that can be used to derive the root words from the inflected words. These methods when used can help us reduce the size of the document which makes it easier for further processing and analysis. Stemming can be defined as the process of removing inflections in the word in order to derive the stem, base or root form even if that root form in itself is not a valid word in that language. In Stemming, the program used to derive a stem is known as stemmer. Many stemmers have been proposed some as early as 1960s. There are two most commonly used stemmers – Porter Stemmer and Lancaster Stemmer. Porter Stemmer finds the stem of the given word by following a set of five rules that are applied in phases. Lancaster Stemmer on the other hand stores a set of 120 rules stored in a table that is indexed by the last character of the inflected word in each iteration. Each rule specifies a replacement or deletion. If no such rule is found, the stemmer terminates. Both of the stemmers perform well but Porter Stemmer is fast hence used in Information Retrieval System. Lancaster Stemmer is slow because the large number of iterations but more efficient.

Lemmatization on the other hand can be defined as the process of removing inflections in the word in order to derive the lemma, stem, base or root form such that lemma is a valid word in that language. We can get clear picture by referring the following examples:

Stemming

Friendship->Friendship
Destabilized->Destabil
Troubling->Troubl

Lemmatization

Friendships -> Friend
Destabilized-> Dest
Troubling -> Trouble

There is always a dilemma about which one to use for deriving the root word – stemming or lemmatization. It depends completely on the application. In our project we have used lemmatization as we need a lemma that is a valid word in the English Language.

```
In [26]: for review in newBadReviews:
             tokens = nltk.word_tokenize(review)
             for word in tokens:
                 if word in stop_words:
                     tokens.remove(word)
                 if word in punctuations:
                     tokens.remove(word)

             for word in tokens:
                 str=str + " "+(wordnet_lemmatizer.lemmatize(word,pos="v"))
             badReviewClean.append(str)
             str=" "
```

```
In [16]: badReviewClean

Out[16]: [' face problem battery drainage',
          ' battery backup okay picture quality bright and sound quality moderate',
          ' at a good phone battery lag',
          ' worst picture sound quality.processor very slow',
          ' lookwise phone good battery back also nice',
          ' good camera quality battery almost last a day good display fingerprint
          ' battery back could better',
          ' heat problem not very excellent product',
          ' mobile slow hang',
          ' useless phone hang of time',
```

Fig 2: Cleaned Data After Removal Of Stopwords And Application Of Lemmatization

II. LITERATURE SURVEY

There are quite a few research papers related to the work. In this section, we provide the details of the work done in these papers.

In [1], the authors have emphasized on the need of identifying communities in the research fields. They have analysed the data and identified that "Review of Modern Physics" has the greatest number of citations. Here they have tried to identify the most active and influential nodes. For this they have applied sociometric analysis.

In [2], the authors have nicely explained that Twitter is one of the most important tools for dissemination of information related to scholarly articles. In this paper they have established that Twitter is being mostly used by non-academic users to discover information and develop connections with scholars to gain access to their scholarly materials.

In [3], the authors have tried to investigate if the removal of stopwords negatively affects the classification of tweets in sentimental analysis on twitter data. They have applied six different methods to identify stopwords from equal number of datasets. They concluded that using a static list of stopwords to identify stopwords from six datasets lead to wrong classification of tweets in sentimental analysis. Generating a list of stopwords dynamically lead to high performance classification with reduction in data sparsity.

In [4], the author has done a comparative study on different stemming algorithms. It is decide that the algorithms can be classified into two types – rule based and linguistic based. If one algorithm outperforms other in one area, the latter may perform well in some other area. Furthur, It is concluded that the problems of over stemming and under stemming can be solved by considering the semantics of words and parts of speech of the language.

In [5], the authors have given an introduction to stemming and lemmatization. Here, they have explained the concepts practically by executing the stemming and lemmatization algorithms in python programming language. They have explained the Porter and Lancaster stemming. The writers have mentioned that Porter Stemmer is fast compared to Lancaster Stemmer.

They have used NLP to perform stemming and lemmatization.

In [6], the authors have provided a study on topic modelling. Topic Modelling provides a way for analysing unclassified text efficiently. A topic contains a group of words known as clusters that occurs frequently together A Topic Model can be defined as a unsupervised analysis of unclassified topics across various text documents. These topics which occur are abstract in nature. Similarly, there is a possibility of having multiple topics in an individual document. The paper provides two categories that come under topic modelling. First one discusses the area of methods of Topic Modelling, which has four methods. These methods are as follows Correlated Topic Model (CTM), Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA)and Probabilistic Latent Semantic Analysis (PLSA). The second category is known as Topic Evolution Model, which considers an important factor time.

III. METHODOLOGY

In this project, we have proposed a model that analyses the customer reviews in order to find best features in all the given product. These features will be stored in a list. Anytime a prospective customer searches for product having a particular feature, this model will list down all the products that were highly rated for that feature by the existing customers of those products. We have tested our model for customer reviews of Samsung, Micromax and Jio smartphones.

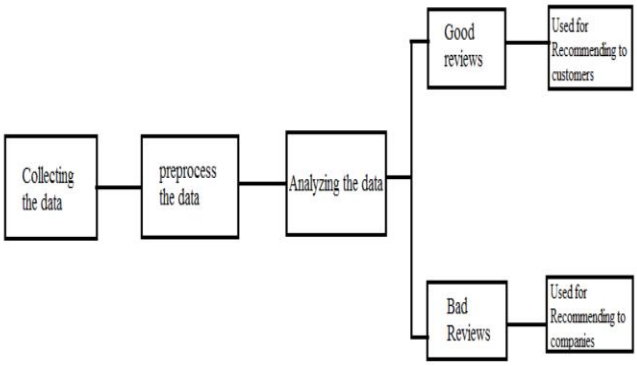

Fig 3: The Methodology

The algorithm of the model is as shown below:

1. The input data, which is the customer ratings and reviews of products (i.e. customer reviews of Samsung, Micromax and Jio smartphones) using web scraping or Amazon API are collected and stored it in their respective .csv files.
2. The required packages and csv file are imported into a list *data* in the Python Notebook. Each of the .csv files have two attributes *Ratings* and *Reviews.*
3. Creation of the lists – *goodReviews* and *badReviews* from *data*. *goodReviews* contains all the reviews whose corresponding ratings are either greater than 3 out of 5. All other rated reviews occupy the *badReviews* list.
4. The stopwords from both of the above-mentioned lists are removed. This is done by tokenizing all the reviews in a list and ensuring that none of these tokens are present in nltk.corpus.Stopwords.word(English). If present, that token is removed from the list.
5. Lemmatization is performed on both of the above-mentioned lists after stopwords removal. This is done by tokenizing the reviews and feeding each of these tokens to nltk.stem.WordNetLemmatizer. Each of these outputs is added to form new lists *goodReviewsClean* and *badReviewsClean* generated from the tokens of *goodReviews* and *badReviews* respectively.
6. Next, the occurrence of tokens in their respective lists to generate a count for each of the tokens. We generate two dictionaries – *goodReviewsDict* and *badReviewsDict* that will store the {token : count} as its elements for all tokens from *goodReviewsClean* and *badReviewsClean respectively.*
7. For each element in *goodReviewsDict*, its count is compared with a threshold value. If the count is greater than the threshold, then we add the token in a new list *finalGoodReviews.* Same procedure is applied for the elements of *badReviewsDict to generate finalBadReviews.*
8. Steps 2 to 7 are repeated for customer reviews all other products stored in their respective .csv files.

Table 1: The Table Depicts The Dataset Collected Using The Reviews Of Products In E-Commerce Websites

A	B	C	D
Poducts	goodFeature1	goodFeature2	goodFeature3
Samsung	display	camera	radio
Micromax	cheap	sound	ROM
OnePlus	processor	camera	sound
Levis	cheap	longetivity	quality
Raymond	quality	formal	indian
Manyavar	traditional	quality	indian
Parle-Agro	cream-biscuits	chips	indian
Patanjali	chipa	cream-biscuits	indian
Britannia	chips	cream-biscuits	indian

IV. RESULTS

If the constraint of customer is camera then the recommended mobiles are

```
The smartphones that have been voted for best Camera are:
1. Samsung
```

For the product Micromax Canvas2 phone the feature that is suggested to upgrade is

```
The recommendations to the product for Micromax Canvas 2 mobile is
Camera
Sound
```

V. CONCLUSION

In this project we have proposed a model that can be used for recommendation of high-rated products based on query of prospective customers. We are analysing customer reviews to find out the best and worst features of three products – Micromax, Samsung and Jio smartphones. This information can be used recommend products based on user queries. Recommended list will be genuine as it has been generated by analysing customer experiences with that product. The information about bad features can be used by manufacturers to improve their future releases or products. This model has been tested on static data. In future, we plan to simulate the model on real-time data.

REFERENCES

1. B S. Khan· M A. Niazi: Network Community Detection, Published in Arxiv,2017
2. R. Fischhoff, S. R. Sundaresan, J. Cordingley, H. M. Larkin, M.-J. Sellier, and D. I. Rubenstein. Social relationships and reproductive state influence leadership roles in movements of plains zebra (equus burchellii). Animal Behaviour, 2006. Submitted.
2. Saif, Hassan; Fernández, Miriam; He, Yulan and Alani, Harith (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings., pp. 810–817.
3. Jivani, Anjali. (2011). A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl. 2. 1930-1938.
4. https://www.datacamp.com/community/tutorials/stemming-lemmatization-python
5. Rubayyi Alghamdi, Khalid Alfalqi, "A Survey of Topic Modeling in Text Mining", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 6, No. 1, 2015
6. K. Carley, M. Prietula, and editors. Computational Organization Theory. Lawrence Erlbaum associates, Hillsdale, NJ, 2001.
7. E. Dahlhaus, D. Johnson, C. Papadimitriou, P. Seymour, and M. Yannakakis. The complexity of multiterminal cuts. SIAM J. Comput., 23:864–894, 1994.
8. Davis, B. B. Gardner, and M. R. Gardner. Deep South. The U. of Chicago Press, Chicago, IL, 1941.
8. Moosavi, S.A., Jalali, M., Misaghian, N. et al. Knowl Inf Syst (2017) 51: 159. https://doi.org/10.1007/s10115-016-0970-8
9. Sergioli, Giuseppe & Santucci, Enrica & Didaci, Luca & Miszczak, JarosÅ‚aw & Giuntini, Roberto. (2017). A quantum-inspired version of the nearest mean classifier. Soft Computing. 10.1007/s00500-016-2478-2.
10. Buccio, E.D., Li, Q., Melucci, M., & Tiwari, P. (2018). Binary Classification Model Inspired from Quantum Detection Theory. ICTIR.
11. Tiwari, Prayag & Melucci, Massimo. (2018). Multi-class Classification Model Inspired by Quantum Detection Theory.
12. S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti "General optimization technique for high-quality community detection in complex networks" Phys. Rev. E 90, 012811 2014.

AUTHORS PROFILE

Voggu Suman Venkata Sai received his B.Tech degree in Computer Science and Technology from Koneru Lakshmaiah Educational Foundation. He is pursuing his M.Tech in Computer Science and Engineering with specialization in Bigdata Analytics from Vellore Institute of Technology. His research interests are Community identification, bigdata streaming and machine learning applications

Yuvraj Singh Champawat received his B.E degree in Information Technology from Gujarat Technological University. He is pursuing his M.Tech in Computer Science and Engineering with specialization in Bigdata Analytics from Vellore Institute of Technology. His research interests are Community identification, web mining applications.

Dr.B K Tripathy is a Senior Professor in SCOPE, VIT, Vellore, India. He has published more than 550 technical papers in international journals, conference proceedings and edited research volumes. He has supervised 48 candidates for research degrees. He has two published books, 6 research volumes, monographs and guest edited some research journals. Dr. Tripathy is a senior member of several professional bodies including IEEE, ACM, IRSS and CSI. His current research interest includes Fuzzy Sets and Systems, Rough Set theory, Data Clustering, Social Network Analysis, Neighbourhood Systems, Soft Sets, Social Internet of Things, Big Data Analytics, Multiset theory and List theory.