An Analytical Approach to Customer Sentiments Using NLP Techniques and Building a Brand Recommender Based on Popularity Score

** Sudesna Baruah*
*** Subhabaha Pal*

Abstract

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence that is concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

This field has a great importance in many areas like retail stores and websites, social networking sites like Twitter, Facebook etc.; for understanding the various views or sentiments of the public or the consumers or end users, regarding a particular product or topic.

Keywords: Natural Language Processing, Sentiment Analysis, Topic Modeling, LDA, VADER

I. INTRODUCTION

The retail sector is seeing a continual and phenomenal transformation over the last few years, all thanks to the NLP technology. It changed (and is changing) the rules of customer engagement with brands, but with many competitors and the ever-growing retail market, customers are demanding more and better. Getting them to notice your brand in such situations is a greater challenge.

Customer engagement is very crucial for any retail business, be it E-commerce or a regular brick-and-mortar business. Technology comes to the rescue during the times of launching a new product, enhancing the brand image, connecting with the customers, etc. Companies are taking full advantage of the latest technologies, like Artificial Intelligence to gain a better understanding of the customers and to drive personalized engagement.

Retailers are now slowly realizing the value that data mining can bring to their business as it provides useful insights about their operations and customers. Unstructured data, like emails, feedbacks, social media, etc., also provides rich insights to any business. Hence, Natural Language Processing becomes an obvious choice to analyze the data and gain insights, from both voice and text-based data inputs.

II. NATURAL LANGUAGE PROCESSING (NLP)

NLP is an AI and Machine Learning technology that enables machines to process huge amounts of structured and unstructured customer data.

It involves speech and text-based communication. It is a technology behind chatbots, virtual assistants, online translation services and many more.

Imagine a customer who wants to buy trek shoes that are in red colour. He/she will search on an e-Commerce website by entering a search query as 'red trek shoes' or 'red shoes for trekking'. A search engine that is not

Manuscript Received: November 11, 2019; Revised: November 19, 2019; Accepted: November 25, 2019. Date of Publication: December 5, 2019.
* S. Baruah is Analyst at Tata Consultancy Services Ltd., Electronic City, Phase II, Bengaluru – 560 100, Karnataka, India.
(email: freespiritsudesna@gmail.com)
** S. Pal is Senior Faculty, Data Science and Machine Learning with Manipal ProLearn (Manipal Academy of Higher Education – South Bangalore Campus), 3rd Floor, Salarpuria Symphony, 7, Service Road, Pragathi Nagar, Electronics City Post, Bengaluru – 560 100, India.
(email: subhabaha@gmail.com)
DOI : 10.17010/ijcs/2019/v4/i6/150421

powered by NLP would show the images of red-coloured regular shoes or trek shoes in any colour as it cannot accurately interpret the language. The search engine, without NLP, won't be able to understand the context and the exact intent of the user. This traditional search process is continuously trained and improved with NLP and provides the customer with the images of red trek shoe products.

A. Statement of the Problem

The target of a retail website is to make profits by making preferred products available online to the customers in order to retain their loyalty. Loyalty is retained by not only making the products available, but by offering suitable discounts during certain period of the year or regular discounts to the regular customers for any purchase.

In order to get this loyalty, the retail website needs to sell those products more, which the customers have liked in the past and may buy those products again in future. For example, if a customer has purchased a Samsung Galaxy mobile, has liked it, and has mentioned about her likes and preferences about the phone in the comment section of the site, and likewise, if there are many other customers giving positive feedback about Samsung Galaxy phones, then as a retailer, I should make an effort to advertise to sell Samsung Galaxy phones more. Also, I should see that the phones are sold at comparatively cheaper rates, with appropriate price endings. I should be giving festive offers like discounts, free accessories or gifts etc.

The next thing is to identify which features of the phone the customers have liked the most. So, based on the feature preference, a report can be sent to the phone manufacturers to improve or include those features in their mobile phones for mutual benefit of both the parties, or I can sell Samsung phones which offer the best of such features.

It is very important to know the positivity or negativity of the sentiments of the customer. Therefore, a Sentiment Analysis shall be the topmost priority. Based on the polarity of the sentiments, information like which mobile brand and model has been the most loved one, which features have been liked the most, which brand or model is hated & why can be found out.

A barrier in these investigations could be language. As the sales are across the globe, the reviews are provided in different languages on the website.

It is not possible to know or understand all the languages in which the comments are posted. Therefore, language identification and machine translation to a known language would be helpful.

B. Objectives

The aims of this project are :

- To find out the sentiment polarity of the comments or reviews posted
- To identify the features or sub-features of the mobile phones the customers have commented on
- To build a brand recommender to give mobile phone suggestions to the customers of the retail website
- To identify and translate an unknown language (in which a comment has been posted) to a known language like English

The project will be useful for the retailers to know the customer requirements and customize their site accordingly to sell the desired products with the preferred features. The project targets achieving customer satisfaction and loyalty towards the retail website in terms of mobile phone purchase. This is done by applying suitable NLP techniques & Popularity Score calculation.

C. Project Goal and Scope

The project aims at successfully understanding the requirements or likes and dislikes of the customers and customizing the website accordingly. The ultimate goal is to sell products to customers as per their choice and demand, and make the same available for sale at reasonable prices along with complementary low cost accessories (if necessary).

The scope of such a project work is getting customer loyalty and patronage towards the retail website, which in return, shall yield benefits to the retailer.

The project work will focus on performing various NLP techniques on comments or reviews posted by customers on the retail website in order to improve sales through the website.

D. System Overview

This project involves the implementation of VADER Sentiment Analysis, Topic Modelling, Brand Recommendation, Text Language Identification, and Machine Translation on the reviews provided by customers. These techniques shall help in understanding the needs and preferences of customers.

Using this information, the website should be presented in such a way that the customers find it convenient and preferable to buy a product of their choice

from that site only, rather than any other retail website. It can be said that the project aims to help retailers build a customer centric website for their loyalty and patronage.

If customers like Samsung Galaxy phones more, the retailer should try to sell Samsung Galaxy phones more. If customers have liked certain features, the retailer should sell products that offer best of such features.

E. Natural Language Processing Techniques

The NLP Techniques involved in this project work are:

- VADER Sentiment Analysis
- Topic Modelling
- Text Language Identification &
- Machine Translation

Also, a Brand Recommender based on Popularity Score has been designed.

F. Literature Review

In the past few years we have seen that text analysis has emerged to be a very important technique to analyze or understand customers'- sentiments, aspects or features highlighted etc. As it is important to understand customers' preferences and sentiments, it was necessary to come up with NLP techniques like Sentiment Analysis, Topic Modelling, Machine translation etc. Such techniques alone can help identification of these requirements by the retailer.

There are many such NLP techniques which can be applied. Some of the commonly used ones have been considered for this project work. A brand recommender can help a retailer offer brands and models on the website which are popular among the customers or are the most preferred. Section III covers literature review.

III. LITERATURE REVIEW

A. VADER : A Parsimonious Rule-Based Model For Sentiment Analysis of Social Media Text [1]

The inherent nature of social media content poses serious challenges to practical applications of sentiment analysis. We present VADER, a simple rule-based model for general sentiment analysis, and compare its effectiveness to eleven typical state-of-practice benchmarks including LIWC, ANEW, the General Inquirer, SentiWordNet, and machine learning oriented techniques relying on Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms. Using a combination of qualitative and quantitative methods, we first construct and empirically validate a gold-standard list of lexical features (along with their associated sentiment intensity measures) which are specifically attuned to sentiment in microblog-like contexts. We then combine these lexical features with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. Interestingly, using our parsimonious rule-based model to assess the sentiment of tweets, we find that VADER outperforms individual human raters (F1 Classification Accuracy = 0.96 and 0.84 respectively), and generalizes more favourably across contexts than any of our benchmarks.

B. Sentiment Analysis of Student Evaluations of Teaching

In [2] the researchers used a sentiment analysis tool, VADER (Valence Aware Dictionary and sEntiment Reasoner), to analyze Student Evaluations of Teaching (SET) of a single course from three different sources: official evaluations, forum comments from another course, and an unofficial “reviews” site maintained by students. They compared the positive and negative valences of these sites; identified frequently-used key words in SET comments and determined the impact on positivity/negativity of comments that included them; and determined positive/negative values by question on the official course SET comments. Many universities use similar questions, which may make this research useful for those analyzing comments at other institutions. Previous published studies of sentiment analysis in SET settings are few.

C. Utilizing Microblog Data in a Topic Modelling Framework for Scientific Articles' Recommendation [3]

Researchers are actively turning to Twitter in an attempt to network with other researchers, and stay updated with respect to various scientific breakthroughs. Young and novice researchers have also found Twitter as a valuable source of information in terms of staying up-to-date with various developments in their field of research. In this paper, we present an approach to utilize this valuable information source within a topic modelling framework to suggest scientific articles of interest to novice researchers.

The approach in addition to producing effective recommendations for scientific articles alleviates the

cold-start problem and is a step towards elimination of the gap between Twitter and science.

D. A Text Mining Research Based on LDA Topic Modelling

A large volume of digital text information is generated every day. Effectively searching, managing, and exploring text data has become an important task. In this paper, the researchers [4] first represented an introduction to text mining and a probabilistic topic model Latent Dirichlet allocation. Then two experiments are proposed, namely, Wikipedia articles and users' tweets topic modelling. The former builds up a document topic model, aiming to a topic perspective solution on searching, exploring, and recommending articles. The latter one sets up a user topic model providing a full research and analysis over Twitter users' interest. The experiment including data collection, data pre-processing, and model training, which is fully documented and commented. Furthermore, the conclusion and application of this paper could be a useful computation tool for social and business research.

E. Guest Editors' Introduction : Machine Learning and Natural Language [5]

The application of machine learning techniques to natural language processing (NLP) has increased dramatically in recent years under the name of \corpus-based,"\statistical," or \empirical" methods. However, most of this research has been conducted outside the traditional machine learning research community. This special issue attempts to bridge this divide by assembling an interesting variety of recent research papers on various aspects of natural language learning (many from authors who do not generally publish on the traditional machine learning literature), and presenting them to the readers of Machine Learning.

IV. FEASIBILITY STUDY AND REQUIREMENT ANALYSIS

A. Feasibility Study

While natural language processing isn't a new science, the technology is rapidly advancing, thanks to an increased interest in human-to-machine communications, plus an availability of big data, powerful computing and enhanced algorithms. As a human, you may speak and write in English, Spanish or Chinese. However, a computer's native language known as machine code or machine language is largely incomprehensible for most people. At a device's lowest levels, communication occurs not with words but with millions of zeros and ones that produce logical actions.

B. Large Volumes of Textual Data

Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment, and determine which parts are important.

Today's machines can analyse more language-based data than humans without fatigue and in a consistent, unbiased way. Considering the staggering amount of unstructured data that is generated every day from medical records to social media, automation will be critical to fully analyse text and speech data efficiently.

C. Structuring a Highly Unstructured Data Source

Human language is astoundingly complex and diverse. We express ourselves in infinite ways, both verbally and in writing. Not only are there hundreds of languages and dialects, but within each language there is a unique set of grammar and syntax rules, terms and slang. When we write, we often misspell or abbreviate words, or omit punctuation. When we speak, we have regional accents, we mumble, stutter, and borrow terms from other languages.

While supervised and unsupervised learning, and specifically deep learning are now widely used for modelling human language, there is also a need for syntactic and semantic understanding, and domain expertise that are not necessarily present in these machine learning approaches. NLP is important because it helps resolve ambiguity in language and adds useful numeric structure to the data for many downstream applications, such as speech recognition or text analytics.

The proposed system might not produce accurate results for the long term. Some variations or changes might be necessary in the codes to yield accurate results in future on the basis of the data fed to it.

This might happen due to variation in the data extracted and fed. However, the framework shall be the same, more or less, with some minor changes in code.

Also, in future, there might be a possibility that a

better technique shall yield better results, outperforming the current system. Many parameters may affect it on the way due to which long term usage might not be feasible.

D. Requirement Analysis

After extensive analysis of the problems, we are familiarized with the requirement that the current system needs. The requirement that the system needs are listed below:

- The system should be able to analyse the sentiment polarity of the reviews or comments posted.
- It should be able to procure the words/topics/documents which majority of customers have laid more emphasis on.
- Should be able to identify a language correctly in which a comment has been posted and should also be able to translate the same to another language of choice, correctly.

V. DATASET UNDERSTANDING

The data is a collection of customer reviews for all mobile phones purchased from retail website. The column 'title' is a summary or heading for a product review, and 'body' is a detailed description of the reviews. Both the columns have been considered for performing the necessary analysis.

Each review has been provided against a model of a brand. A popularity score calculated using the data of the ratings provided can help in knowing the preferred brands and their respective models. The columns 'rating', 'TotalReviews', 'brand', 'title' have also been considered for performing the necessary analysis.

The reviews are given in various languages apart from English, such as German, Spanish etc. Use of Machine Translator, can effectively help in retrieving such comments in a known language.

VI. SYSTEM DESIGN AND ARCHITECTURE

A. Solution Approach

The CRISP DM methodology has been followed. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. We do not claim any ownership over it. We did not invent it. We are however, evangelists of its powerful practicality, its flexibility, and its usefulness when using analytics to solve thorny business issues. It is the golden thread that runs through almost every client engagement.

This model is an idealised sequence of events. In practice many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions. The model does not try to capture all possible routes through the data mining process.

(i) Business Understanding : Understand business problems for business perspectives.

(ii) Data Understanding : Collect data, perform NLP techniques,and all necessary actions to meet business requirements from the data.

(iii) Data Preparation : Perform various activities to derive insights from the raw data & preparing a final dataset of it. The final dataset is used for modelling. This step includes data cleaning & data transformation.

(iv) Modelling : Select and apply various modelling techniques. Focus is on choosing a model which can best predict subscription to term deposit based on campaign.

(v) Evaluation : Thoroughly evaluate the model using statistical techniques and ensure that business objectives are met.

(vi) Deployment : Generate insights and recommendation from the data and model that can be presented to the business.

B. Architecture

The input to our system is extracted from the retail website on yearly basis to analyse the sentiments of its customers as a yearly report, and calculate or formulate a statistical report. The report shall include details like overall polarity of customer sentiments for mobile phones sold through the retail website, net positive and net negative scores, highlighting features that are liked or disliked by customers, understanding the comments posted by customers. Also, translating the comments posted in an unknown language to a known language.

VII. MODELLING

A. Sentiment Analysis

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It is also known as opinion mining, deriving the opinion or attitude of a speaker.

VADER (Valence Aware Dictionary and Sentiment

Reasoner) [1] is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of a sentiment lexicon which is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative. VADER not only tells about the Positivity and Negativity score, but it also tells us about how positive or negative a sentiment is.

Businesses today are heavily dependent on data. Majority of this data however, is unstructured text coming from sources like emails, chats, social media, surveys, articles, and documents. Micro-blogging content coming from Twitter and Facebook poses serious challenges, not only because of the amount of data involved, but also because of the kind of language used in them to express sentiments, i.e., short forms, memes, and emoticons.

Sifting through huge volumes of this text data is difficult as well as time-consuming. Also, it requires a great deal of expertise and resources to analyze all of that. In short, not an easy task.

Sentiment Analysis is also useful for practitioners and researchers, especially in fields like sociology, marketing, advertising, psychology, economics, and political science, which rely a lot on human-computer interaction data. Sentiment Analysis enables companies to make sense out of data by being able to automate this entire process! Thus, they are able to elicit vital insights from a vast unstructured dataset without having to manually indulge with it.

Though it may seem easy, Sentiment Analysis is actually a tricky subject. There are various reasons for the understanding that emotions through text are not always easy. Sometimes, even humans can get misled, so expecting 100% accuracy from a computer is asking for the moon! A text may contain multiple sentiments all at once. For instance,

"The intent behind the movie was great, but it could have been better".

This sentence consists of two polarities, that is, positive as well as negative. So how do we conclude whether a review was positive or negative? Computers aren't too comfortable in comprehending figurative speech. Figurative language uses words in a way that deviates from their conventionally accepted definitions in order to convey a more complicated meaning or heightened effect. Use of similes, metaphors, hyperboles etc. qualify for a figurative speech. Let us understand it better with an example:

The best I can say about the movie is that it was interesting."

Here, the word 'interesting' does not necessarily convey positive sentiment and can be confusing for algorithms. Heavy use of emoticons and slangs with sentiment values in social media texts like that of Twitter and Facebook also makes text analysis difficult. For example a " :)" denotes a smiley and generally refers to positive sentiment while ": (" denotes a negative sentiment on the other hand. Also, acronyms like "LOL", "OMG" and commonly used slangs like "Nah", "meh", "giggly" etc. are also strong indicators of some sort of sentiment in a sentence.

These are few of the problems encountered not only with sentiment analysis but with NLP as a whole. In fact, these are some of the Open-ended problems of the Natural Language Processing field.

VADER has been found to be quite successful when dealing with social media texts, NY Times editorials, movie reviews, and product reviews. This is because VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is. It is fully open-source under the MIT License.

Advantages of using VADER :

VADER has a lot of advantages over traditional methods of Sentiment Analysis. These are:

- It works exceedingly well on social media type text, yet readily generalizes to multiple domains.
- It doesn't require any training data but is constructed from a general, valence-based, human-curated gold standard sentiment lexicon
- It is fast enough to be used online with streaming data, and
- It does not severely suffer from a speed-performance trade-off.

B. Topic Modelling

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modelling[2] is a frequently used text-mining tool for discovery of hidden semantic structures in a text body.

Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and

"meow" will appear in documents about cats, and "the" and "is" will appear equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words. The "topics" produced by topic modelling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.

Topic models are also referred to as probabilistic topic models, which refer to statistical algorithms for discovering the latent semantic structures of an extensive text body. In the age of information, the amount of written material we encounter each day is simply beyond our processing capacity. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies.

Originally developed as a text-mining tool, topic models have been used to detect instructive structures in data such as genetic information, images, and networks. They also have applications in other fields such as bioinformatics.

In this brief introduction, we focus on a widely used text-mining method: Latent Dirichlet Allocation (LDA). LDA gained popularity due to its ease of use, flexibility, and interpretable results. First, we briefly explain the basics of the algorithm for all non-technical readers. In the second part of the post, we show what LDA can achieve with a sufficiently large data set.

Latent Dirichlet Allocation (LDA) : Text data is high-dimensional. In its most basic but comprehensible to computers form, it is often represented as a bag-of-words (BOW) matrix, in which each row is a document and each column contains a count of how often a word occurs in the documents. These matrices are transformable by linear algebra methods to discover the hidden (latent and lower-dimensional) structure in it.

Topic modelling assumes that documents such as news articles contain various distinguishable topics. As an example, a news article covering the Cambridge Analytica scandal may contain the following topics: social media, politics and tech regulations, with the following relations: 60% social media, 30% politics, and 10% tech regulations.

The other assumption is that topics contain characteristic vocabularies, for example, the social media topic is described by the words Facebook, Twitter etc.

LDA has been proposed by Blei et al. (2003) on the basis onfBayesian statistics. The method's name provides its key foundations [3]. Latent comes from the assumption that documents contain latent topics that we do not know a priori. Allocation shows that we allocate words to topics, and topics to documents. Dirichlet is a multinomial likelihood distribution: it provides the joint distribution of any number of outcomes. As an example, Dirichlet distribution can describe the occurrences of observed species in a safari. In LDA, it describes the distribution of topics in documents, and the distribution of words in topics.

The basic mechanism behind topic modelling methods is simple: assuming that documents can be described by a limited number of topics, we try to recreate our texts from a combination of topics that consist of characteristic words. More precisely, we aim at recreating our BOW word-document matrix with the combination of two matrices: the matrix containing the Dirichlet distribution of topics in documents (topic-document matrix), and the matrix containing the words in topics (word-topic matrix).

The construction of the final matrices is achieved by a process called Gibbs sampling. The idea behind Gibbs sampling is to introduce changes into the two matrices word-by-word: change the topic allocation of a selected word in a document, and evaluate if this change improves the decomposition of our document. Repeating the steps of the Gibbs sampling in all documents provides the final matrices that provide the best description of the sample.

C. Machine Translation

Machine Translation (MT) is the task of automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language. While machine translation is one of the oldest subfields of artificial intelligence research, the recent shift towards large-scale empirical techniques has led to very significant improvements in translation quality. The Stanford Machine Translation group's research interests lie in techniques that utilize both statistical methods and deep linguistic analysis.

Determining the appropriate weights for a translation system's decoding model is usually performed using Minimum Error Rate Training (MERT), a procedure that optimizes the system's performance on an automated measure of translation quality.

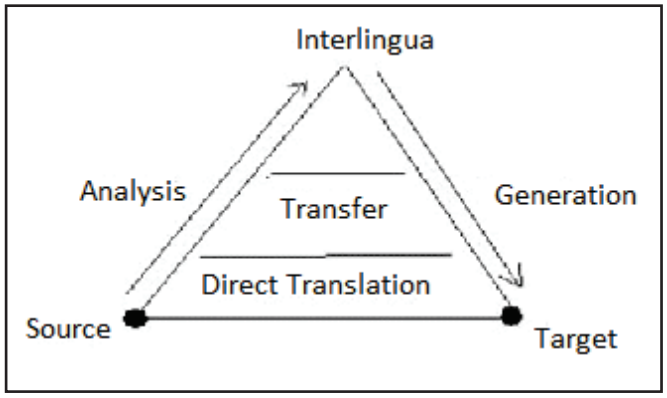

Fig. 1. Translation Approaches

We are continuing to investigate the feasibility and effectiveness of training to evaluation metrics that perform a deeper semantic and syntactic analysis of the translations being evaluated.

Machine translation systems that produce translations between only two particular languages are called bilingual systems and those that produce translations for any given pair of languages are called multilingual systems. Multilingual systems may be either uni-directional or bi-directional. Multilingual systems are preferred to be bi-directional and bi-lingual as they havethe ability to translate from any given language to any other given language and vice versa. Fig.1. shows the Translation approaches.

(1) Direct Machine Translation Approach : Direct translation approach is the oldest and less popular approach. Machine translation systems that use this approach are capable of translating a language, called source language (SL) directly to another language called target language (TL).

The analysis of SL texts is oriented to only one TL. Direct translation systems are basically bilingual and uni-directional. Direct translation approach needs only a little syntactic and semantic analysis. SL analysis is oriented specifically to the production of representations appropriate for one particular TL.

(2) Interlingua Approach: Interlingua approach intends to translate SL texts to that of more than one language. Translation is from SL to an intermediate form called interlingua (IL) and then from IL to TL. Interlingua may be artificial one or auxiliary language like Esperanto with universal vocabulary. Interlingua approach requires complete resolution of all ambiguities in the SL text.

(3) Transfer Approach : Unlike interlingua approach, transfer approach has three stages involved. In the first stage, SL texts are converted into abstract SL-oriented representations. In the second stage, SL-oriented representations are converted into equivalent TL-oriented representations. Final texts are generated in the third stage.

In transfer approach complete resolution of ambiguities of SL text is not required, but only the ambiguities inherent in the language itself are tackled. Three types of dictionaries are required: SL dictionaries, TL dictionaries, and a bilingual transfer dictionary. Transfer systems have separate grammars for SL analysis, TL analysis and for the transformation of SL structures into equivalent TL forms.

(4) Recommendation System : How does YouTube know what videos you'll watch? How does Google always seem to know what news you'll read? It uses a Machine Learning technique called Recommender Systems.

Practically, recommender systems encompass a class of techniques and algorithms which are able to suggest "relevant" items to users. Ideally, the suggested items are as relevant to the user as possible, so that the user can engage with those items: YouTube videos, news articles, online products, and so on.Items are ranked according to their relevance, and the most relevant ones are shown to the user.

Relevance is something that the recommender system must determine and is mainly based on historical data. If you have recently watched YouTube videos about elephants, then YouTube is going to start showing you a lot of elephant videos with similar titles and themes!

Recommender systems are generally divided into two main categories: collaborative filtering and content-based systems. Fig. 2 shows a recommender system.

(5) Collaborative Filtering Systems : Collaborative filtering methods for recommender systems are methods that are solely based on past interactions between users and target items.

Thus, the input to a collaborative filtering system will be all historical data of user interactions with target items. This data is typically stored in a matrix where the rows are the users, and the columns are the items.

The core idea behind such systems is that historical data of users should be enough to make a prediction,that is, we don't need anything more than historical data, no extra push from the user, no presently trending

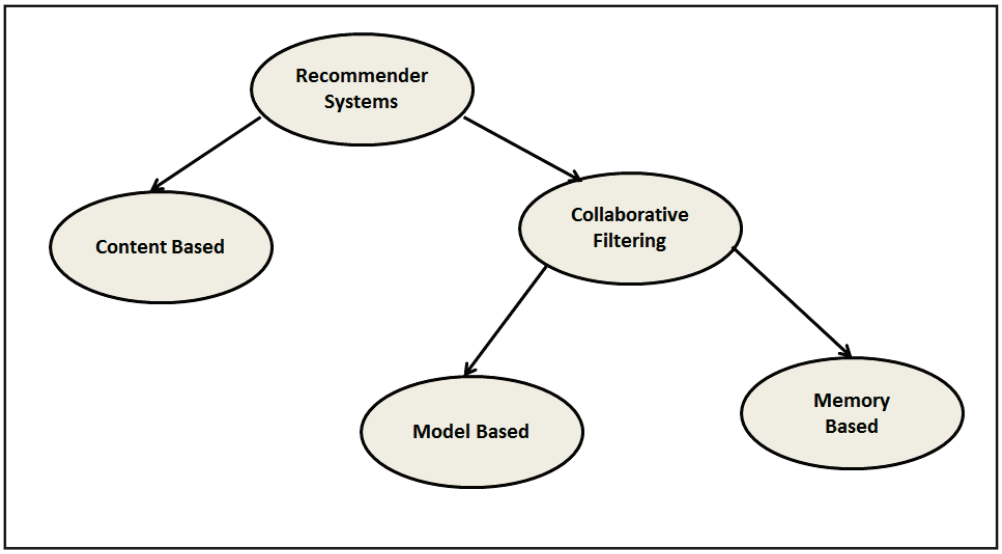

Fig. 2. Recommender System

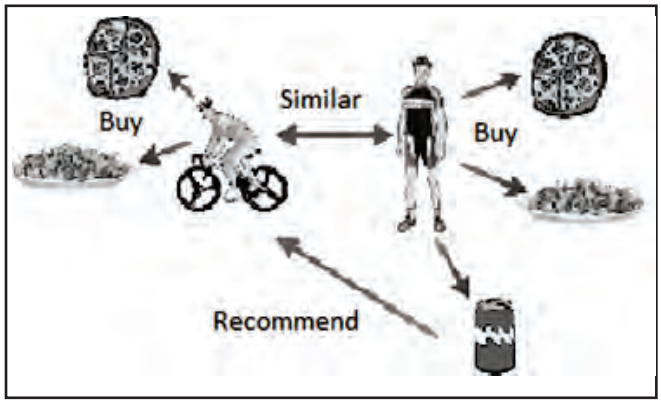

Fig. 3. Collaborative Filtering

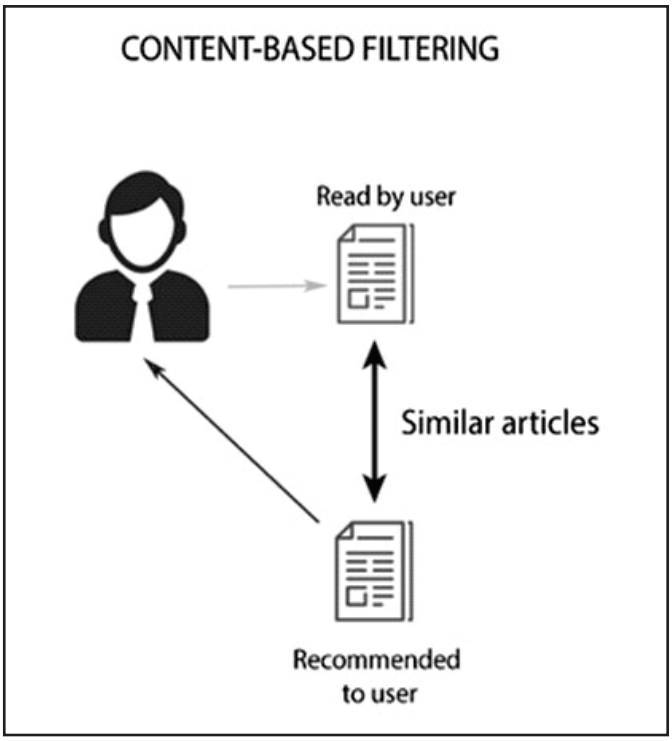

Fig. 4. Content Based Filtering

```
Sentiment_mapping = {1: "High Negative", 2: "Negative", 3: "Neutral", 4: "Positive", 5: "High Positive"}

map_sentiment = lambda val : digitize (val, [-1, -.8, -.25, .25, .8])
```

Fig. 5. Popularity Score

information, etc. Fig. 3. shows a collaborative filtering system.

(6) Content-based Systems : In contrast to collaborative filtering, content-based approaches will use additional information about the user and / or items to make predictions.

For example, a content-based system (Fig. 4) might consider the age, sex, occupation, and other personal user factors when it makes predictions.

When you sign up for many online websites and services, they ask you to (optionally) give your date of birth, gender, and ethnicity! It is just more data for their system to make better predictions.

Thus, content-based methods are more similar to classical machine learning, in the sense that we will build features based on user and item data and use them to help us make predictions. Our system inpus are the features of the user and the features of the item. Our system output is the prediction of whether or not the user would like or dislike an item.

In this project, we propose an algorithm to predict social popularity (i.e., the ratings and total numbers of ratings) of content on Amazon Retail services using only text annotations. Instead of analysing image/video content, we try to estimate social popularity by a combination of ratings provided by customers on the website, and total number of ratings given. Since our proposed algorithm uses text annotations instead of image/video features, its computational cost is small. As a result, we can estimate social popularity more efficiently than previously proposed methods. Our experiments involved using more than 80,000 reviews (text) on Amazon retail website and the results showed a high correlation between actual social popularity and the determination thereof using our algorithm.

VIII. IMPLEMENTATIONS ON DATASET

A. VADER Sentiment Analysis

As VADER takes a considerably longer time to work on a large dataset (82,000 rows approximately), the given data has been split into 8 documents comprising of 10,000 rows (approximately) each. Taking one split document at a time, the data has been read and the 'title' column has been considered for sentiment analysis.

The polarity score of each comment present in the 'title' column has been calculated using the Sentiment Intensity Analyser package and VADER algorithm. The polarity scores are bound within maximum and minimum limits, and accordingly assigned sentiment category where *'1' indicates 'High Negative', '2' indicates 'Negative', '3' indicates 'Neutral', '4' indicates 'Positive' , and '5' indicates 'High Positive'* (Fig. 5).

A calculation has been done to identify the number of positive and number of negative comments posted. It has been found that majority of comments posted in the website are positive, and customers have appreciated Samsung phones the most.

Also, interestingly, the most unwanted phone brand is Samsung as well.

Marketing Strategies :

• As a retailer, the website should work on making Samsung phones more available for customers to buy, in order to gain their loyalty and patronage to brand and the website.These phones should include all the recent releases of the mobile brand Samsung that have been sold the highest number of times.

• The next thing is to sell those phones at a slight lower price than the market price or competitor price. Say, for example, if the market price of a product or the price of the product in another website is ₹1,312, then a competitor retail website should be selling the same for ₹1,299 (say). Use of price ending strategy shall help in increasing the sale of the phones from our retail website.

Also, some cheaper accessories or gift hampers or cashback using certain payment gateways can be offered along with handsets to attract more. customers.

B. Topic Modelling

The whole file has been read and the column 'body' has been considered to perform Topic Modelling algorithm & Latent Dirichlet Allocation (LDA).

• All necessary data pre-processing steps have been performed, that is, Data Read, Removing Capitals,and Removing Stopwords.

• A DTM is created using TF-IDF Vectorizer. The DTM is fit into an LDA model.

• The topics and documents are taken and put in a dataframe. A count of the dominating topics is taken thereafter.

• Top twenty words for each topic has been found out. T he same is displayed in a dataframe.

• The same can be helpful in understanding which features customers have laid more emphasis on.

Marketing Strategies:

- Using the above information, the retail website can focus on selling those phones whose features have been talked about. Say, if there has been mention of 'poor battery life', then the retailer should sell phones that have longer battery life, or sell those Samsung Galaxy phones that have better battery longevity.
- Phones having good battery life, good camera pixel quality, and latest Android version should be sold apart from Apple phones to get the patronage of customers.
- At the same, there has to be a check on the market price of such phones to sell them at a comparatively lower price.
- The retailer should be able to advertise the availability of such phones at a lower price in various media like newspaper, street hoardings etc.

C. Language Identification and Machine Translation

The comments posted in the website are provided in various languages spoken across the world. It is not possible for a data analysts to know all the languages. Therefore, a language identification algorithm can help in identifying the language in which a comment has been written (text) and posted on the website. A Machine translation algorithm is helpful in converting the text written into an unknown language to a text written in a known language.

The comments posted on Amazon Retail website are in various languages spoken across the world,such as, English (en), Somali (so), Slovak (sk), Italian (it), German (de), French (fr) etc. Implementing a Language Identification code using '*langdetect'* package has helped in identifying the unknown language, whether it is German or French or any other (Fig. 6).

```
[en:0.9999962860888639]
[so:0.9999986675629264]
[sk:0.9999936933138593]
[nl:0.5714280808412637,
[ca:0.4285694384286877]
```

Fig. 6. Language Identification Code

To take a deeper dive into the subject matter, for an effective analysis, the text could be translated from non-English languages to English. Here this has been done using the '*Translator'* package of '*Googletrans'* library (Fig. 7).

The implementation is comparatively easier than many of the popular libraries or packages. Once implemented, the translated data can be exported to an excel sheet and the previous analysis done can be carried out again on the translated data. The analysis would be more appropriate and precise compared to the analysis carried out without language identification and machine translation.

In this work, a portion of the data has been taken which bears the comments in non- English or unknown language (German). The text written in German has been translated to known languages like English, Hindi, and Bengali to check the effectiveness of the code and see if it conveys the same meaning in each language or not.

The translation is found to be 90% effective (approximately), with some deviations, as it has not been

```
from googletrans import Translator
Translator = translator (service_urls = [
  'translate.google.com',
  'translate.google.co.kr',
])
```

Fig. 7. Translation to English using Googletrans Library

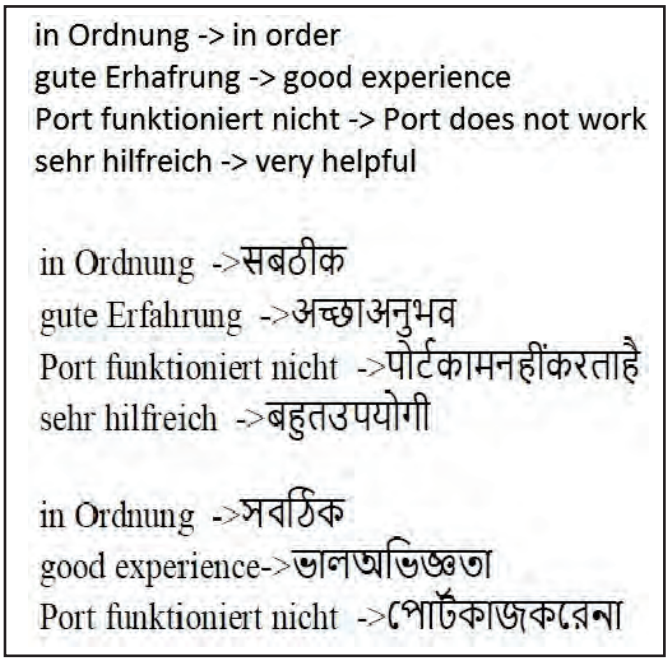

Fig. 8. Translation from German into English, Hindi and Bengali

Top 10 Samsung Models (random)

```
1 top_ models = top18 [(top10.brand - - 'Samsung')]
2 top_models.head (10)
```

	brand	title_x rating-	
71888	Samsung	Samsung Galaxy S9+ Plus Verizon + GSM Unlocked...	5
82663	Samsung	Samsung Galaxy A80 (128GB, 86B RAM) 6.7" DIspi.,,	5
82144	Samsung	Samsung Galaxy S7 Edge G935P SM-G935P 32GB - S...	5
82338	Samsung	Samsung Galaxy J3 V 3rd Gen SM-J337V Eclipse 2...	5
73253	Samsung	Samsung Galaxy Note 9, Verizon, 512GB, Ocean B...	5
69715	Samsung	Samsung Galaxy S8 Plus G955U 64GB Phone - Spri...	5
53212	Samsung	Total Wireless Samsung Galaxy S8+ 4G LTE Prepa...	5
53211	Samsung	Total Wireless Samsung Galaxy S8+ 4G LTB Prepa...	5
53493	Samsung	Samsung Galaxy S8 + 64G8 Phone- 6.2" display -...	5
T1S73	Samsung	Samsung Galaxy S9 + Plus Verizon + GSM Unlocked...	5

Fig. 9. Top 10 Samsung Models (random)

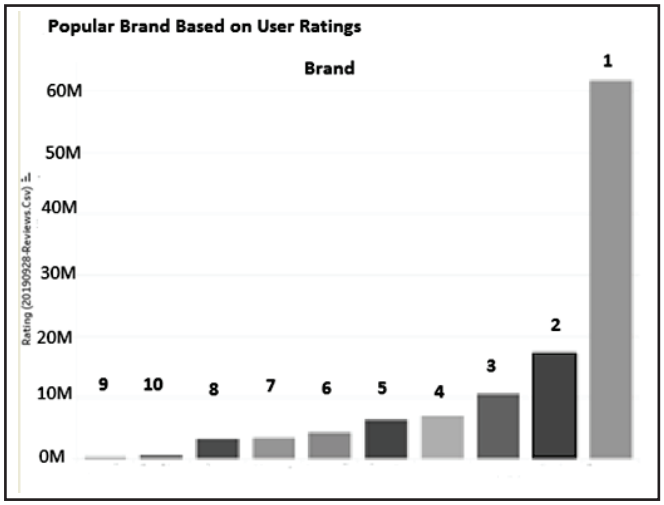

Fig. 10. Popular Brand Based on User Rating

able to translate each comment correctly (conveying the exact message or meaning) in a few languages like Hindi, Bengali, etc. Example: 'in Ordnung' in German means 'in order' in English.

in Ordnung -> in order
guteErfahrung -> good experience
Port funktioniertnicht -> Port does not work
sehrhilfreich -> very helpful

The same has been converted to Hindi & Bengali as '*Sabthik*' and '*Sob thik*' respectively (all okay in English for both). However, instead if we feed 'in Order' for translation to Hindi & Bengali, it yields 'Kram me' & 'Krome' (in order or in a sequence in English for both). (Fig. 8).

So, we can say there is a slight deviation from the original comment or original meaning upon translation. This could also lead to some difference in analysis.

Legend for fig. 10 to 13:

1. Samsung
2. Apple
3. Motorola
4. Nokia
5. Google
6. Huawei
7. Xiaomi
8. Sony
9. ASUS
10. OnePlus

1) Brand Recommendation Based on Popularity Score and Rating

Based on the available data from the dataset, a product of the Customer Ratings and Total Reviewshas been taken. This product is considered as a Popularity Score for identifying the most popular brand. This finding is of immense importance as the retailer can use this information, to recommend a popular brand to its customers.The brand that shows the highest popularity score is Samsung & the one having the lowest popularity score is ASUS.

Also, based on the Customer Ratings, the popular brand and phone model can be identified. The brands that show the highest rating are 'Samsung', 'Google', 'Sony', 'Motorola', 'Apple', 'Xiaomi',and 'ASUS ' (in descending order).` The lowest ratings are given to 'Nokia', 'OnePlus' and 'HUAWEI' (in ascending order). The highest ratings 4 & 5 are given to Samsung for the highest number of times, that is, 44576 times.

There is a slight difference in both the findings, but for the highest popularity score & ratings, Samsung has been the winner. A list of ten Samsung mobile phone models (random) has been obtained by filtering a rating of 5 & brand as 'Samsung', to see which models of Samsung

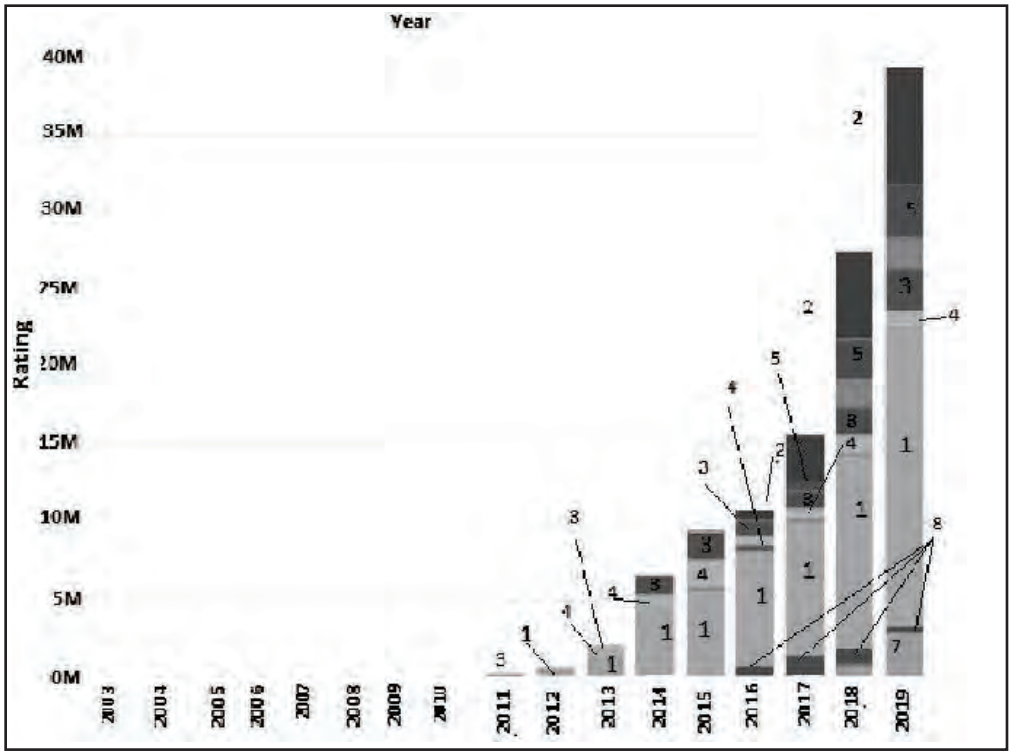

Fig. 11. Popular Brand Based on User Rating Over the Years

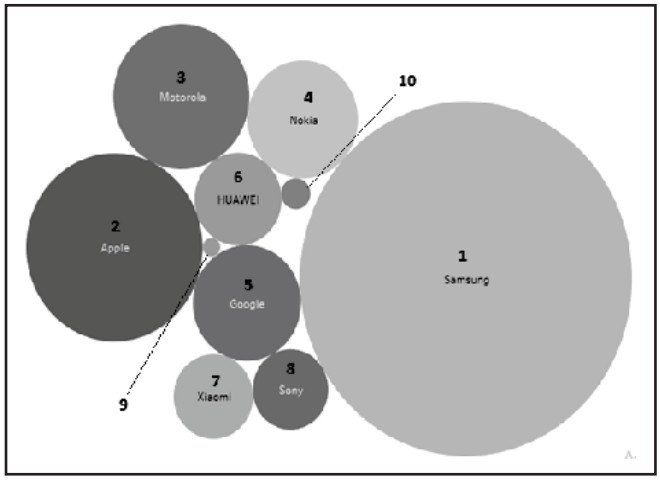

Fig. 12. Popular Brand Based on Popularity Score

has been in demand recently. Samsung phones have dominated the markets of this retail website since 2012. Some popular Samsung models are shown in Fig. 9.

Marketing Strategies :

The top brands based on user ratings and the respective models can be availed in the retail website, along with the most preferred brand on the basis of popularity score.

This strategy, if adopted, shall drastically improve the current sales figure, as all customer preferred brands will be available for sale and as mentioned in the previous strategy, should be sold at a slightly lower price, compared to the competitor websites. Fig. 6 shows popular brand based on user ratings. Fig. 7 shows popular brands based on user ratings over the years. Fig. 8 shows popular brand based on popularity score. Fig. 10. Popularity of brand based on helpful votes and total reviews.

IX. LIMITATIONS AND FUTURE ENHANCEMENTS

A.Limitations

Despite obtaining useful analysis, there will always be certain limitations in the existing analysis or the technology employed to do so. There will always be a scope for improvement or enhancement in the future. As limitations, we can mention the following points for this project work:

1) While using **VADER Sentiment Analysis**, we are not able to consider the whole dataset (82,000 rows) at a time

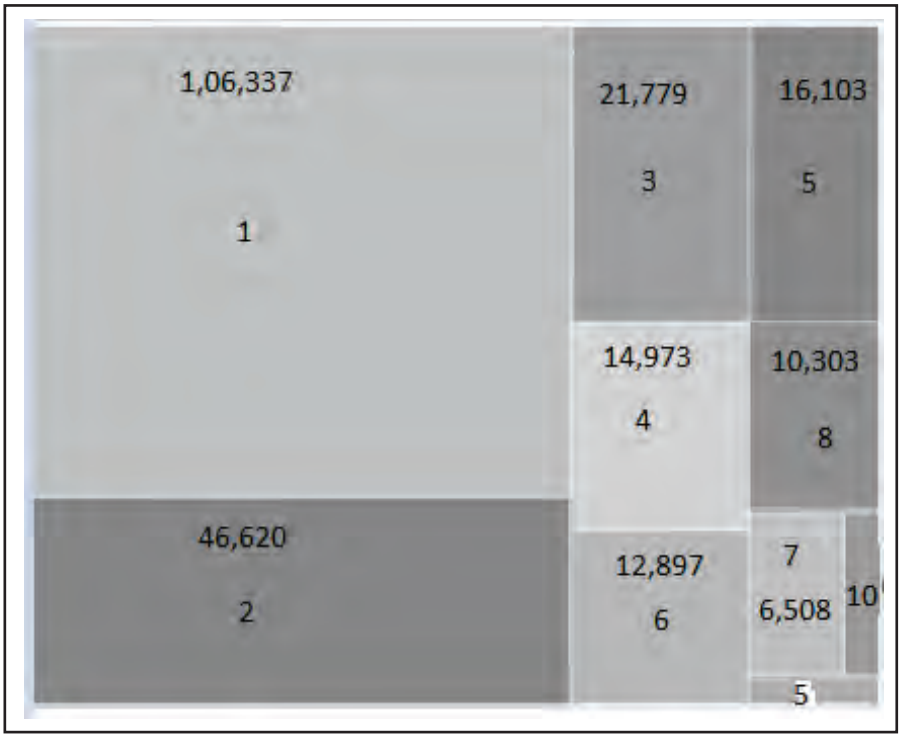

Fig. 13. Popularity of Brand Based on Helpful Votes & Total Reviews

as VADER takes a long time to work on a huge dataset and sometimes may throw an error after running for a long time, like session timeout or any error thrown as a result of being employed on a huge dataset.

To implement VADER, the dataset had to be split into 8 documents, comprising of 10-12,000 rows, and each document had to be run separately to see the VADER output completely. This leads to some imperfections in the analysis. Hence, there have to be some improvements in the current VADER algorithm for making it robust for implementing on a large dataset.The output of each had to be summed up manually for taking the count of net positive and net negative comments.

2) Another limitation of VADER is that it can be used only when the comments are posted in correct English. Therefore, anything other than English has been left undetected or ignored.

So, another Language Identification & Machine Translation code has to be run on each split file to convert all comments to correct English, before running the VADER Algorithm.

3) Topic Modelling has proved to be immensely useful in identifying the features of a mobile phone, about which the users or customers have commented on. Using the DTM & LDA, the features had been featured on the basis of document term matrix (array), in such a way that the features could be categorised for a particular topic. For example: terms *like 'lens', 'pixel', 'good', 'photography', 'quality', 'pictures'* etc., could be categorised under '***Mobile phone Camera'***.

Also, at the same time, certain terms are such that these cannot be categorised into a category. This could be treated as a limitation of a Topic Modelling algorithm using LDA.

4) Again, it is not possible to understand which mobile phone brand or model features have been talked about. This could be added as another limitation for this dataset and project work, employing Topic Modelling & LDA code.

5) Language Identification has been immensely useful in procuring a list of known and unknown languages in which a comment has been posted. The only limitation here is that it can identify only those languages which the algorithm has been trained in. Any other language, apart

from the languages in which it is trained, will be ignored or might be identified as some other similar language.

6) Also, at times it tends to display or identify, only that language which has been fed or provided to the code in majority. For example, if comments are posted mostly, in English, the code might show the language(s) identified as English 'en' alone, even if there are comments posted in non-English languages. In such cases, the code has to be manipulated or tweaked.

7) Machine Translation has been of immense importance here as VADER has used the machine translated dataset to work on sentiment polarity of the comments. Any non-English comment has been translated to English for this project work. Also, few non-English (German) comments are extracted from the dataset and fed to the machine translation code to check its effectiveness.

8) The comments are translated into English, Hindi, and Bengali, and a comparative study has been done to see if the exact meaning has been conveyed in all the three known languages. The code has shown nearly 90-95% effectiveness in translation, with some minor *misunderstanding* of sentiments.

9) This can be treated as an inefficiency of the machine translation code. This, in a larger dataset, shall be more inefficient, leading to inaccurate analysis of the data. If identification of sentiments or understanding sentiment polarity is incorrect, analysis like VADER Sentiment Analysis will be inefficient or inaccurate.

B. Future Enhancements

There is vast scope for **future enhancements** of the existing technologies used here. Say, in case of **VADER**, use of a huge dataset hampers the execution time of the code or algorithm. Hence, we can say, there is a scope of improvement or enhancement of the code to work on a large dataset. At the same time, VADER, works only when the text is written in proper English. Any other text written in non-English or improper English (slangs or short forms) is ignored. Of course, there are other Sentiment Analysis methods, but VADER being simple is widely used. Hence, a change or enhancement in the code will be a great step towards Sentiment Analysis.

Topic Modelling using LDA outputs a document term matrix, displaying the frequently occurring words in each document, but it fails to be more informative like the features of which brand or model are being talked about. So, in future, there are chances of enhancement for topic modelling output to be more informative.

Machine Translation has a drawback of being unable to translate with the correct message and meaning in certain languages. So definitely, there are chances of improvement in the algorithm for accurate translation.

Brand Recommender using popularity score might not show the correct statistics always. Hence, some more exercise or workarounds are needed to verify the same, like the popular brand is found out using the user ratings as well. A comparative study is done to understand, whether the analysis of the Brand Recommender is correct or not. Also, to determine the popular models of the mobile phone brand is possible, through user ratings only (for this dataset).

X. CONCLUSION

This work aims at analysing the Amazon Cell Phone Reviews data collected over the years till September, 2019. As the dataset is based on customer reviews, hence, the analysis has been performed using various NLP techniques, which could be employed in the dataset (based on the information provided through the datasets). Considering the limitations of the dataset, the possible techniques have been chosen and implemented. The techniques have yielded several insights into the dataset like the kind of sentiment expressed, positive, negative, or neutral; cell phone features which have been talked about mostly; languages in which the comments have been posted, and the language in which the comments have been posted mostly; what the customers are saying by translating the comments posted in unknown languages to known languages; which brand is the most preferred or voted, and the respective models as well.

These insights and findings shall be useful for the retailer (Amazon) to understand the customer requirements. On the basis of these findings, they could customize their site to sell the desired products at reasonable prices with attractive offers. *This will help the retailer in gaining customer loyalty, and shall help in increasing the sale of mobile phones in the site.* This work is yet another example of proving the power and efficiency of Natural Language Processing Techniques and their use in the current world. It is a step forward in proving the importance of NLP in the world of Data & Data Science.

REFERENCES

[1] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," Georgia Inst. of Technol., Atlanta, GA, 2014. [Online]. Available: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

[2] H. Newman and D. Joyner, "Sentiment analysis of student evaluations of teaching," Georgia Inst. of Technol., Atlanta, GA, USA, 2018. Doi: https://doi.org/10.1007/978-3-319-93846-2_45

[3] A. Younus, M. A. Qureshi, P. Manchanda, C. O'Riordan, and G. Pasi, "Utilizing microblog data in a topic modelling framework for scientific articles' recommendation," Aiello L. M., McFarland D. (eds.) *Social Informatics. SocInfo 2014. Lecture Notes in Comput. Sci.*, vol. 8851. Springer, Cham., 2014. Doi: https://doi.org/10.1007/978-3-319-13734-6_28

[4] Z. Tong and H. Zhang, "A text mining research based on LDA topic modelling," *Jodrey School of Comput. Sci., Acadia University,* Wolfville, NS, Canada, 2016. doi: 10.5121/csit.2016.60616

[5] C. Cardie and R. J. Mooney, "Guest editors' introduction: machine learning and natural language," *Mach. Learning*, vol. *34*, pp. 5-9, 1999. Doi: https://doi.org/10.1023/A:1007580931600

About the Authors

Sudesna Baruah has been working as an Analyst at Tata Consultancy Services Ltd., Electronic City, Bangalore for over 3 years. She has a B.Tech. degree in Electronics and Communication Engineering from Sikkim Manipal Institute of Technology, Sikkim.

She is also a Data Science enthusiast and has a Post Graduate Diploma in Data Science from Manipal ProLearn, Bengaluru, Karnataka.

Dr. Subhabaha Pal is a seasoned Data Scientist and Academician with over 16 years of experience working in varied fields of Information Science and Analytics. He had been nominated as the Top 20 Data Science and Machine Learning Academicians in India in 2018 by Analytics India Magazine. He completed Ph.D. from the University of Calcutta. He has taught Data Science at well-renowned institutions like Manipal University, T. A. Pai Management Institute, and International Institute of Digital Technologies among others. He had worked in senior software related roles in organizations like Kuwait Petroleum Corporation and Manipal Global. He has around 40 research papers in the field of Data Science and Analytics and three books in renowned publications to his credit. He has delivered many data science projects to different SMEs and has worked in various domains.