Voice of Customer in Auto Industry

Shewatha Arul
PGDM, Business Analytics, REVA University
shewatha.ba05@reva.edu.in

Sudeep Matthew
MBA, Business Analytics, REVA University
sudeep.ba05@reva.edu.in

Anand Limbare
PGDM, Business Analytics, REVA University
anand.ba05@reva.edu.in

Saumyadip Sarkar
MBA, Business Analytics, REVA University
saumyadip.ba05@reva.edu.in

Sneha Tiwari
MBA, Business Analytics, REVA University
sneha.ba05@reva.edu.in

Abstract

It is common practice in any industry to ask their customer to review the products and the associated services. Sometimes reviews come in form of discussion in various discussion forum or web portal. The number of customer reviews can be in hundreds and thousands for popular products. This makes it difficult for the industry to evaluate them in order to get overall feedback and customer satisfaction level.

Identifying customer requirements by analyzing opinions, comments and reviews have been a prime focus of many industries. Sentiment analysis using product review data [8], Opinion Mining and Sentiment Analysis [9], Product weakness finder [10], Sentiment Analysis: A Multi-Faceted Problem [11], Comparative Experiments on Sentiment Classification for Online Product Reviews [12], which form the basis of your work. Providing recommendations to the car industry based on insights extracted from mining web portal will make our project one of its kind as car industry has so far not been covered under any of the available studies.

This project focuses on Indian Auto industry, which can benefit immensely by identifying the features that customers are expecting or not expecting in a vehicle. This project aims to provide this recommendation by analyzing the structured and unstructured text data available in Team BHP portal. This would help car manufacturers design their future product better and upgrade existing products to meet the needs of customers.

We start by extracting text data from TeamBhp.com as is. We then clean the data by removing punctuations and stop words. The cleaned text data is tokenized by splitting in each word and perform lemmatization and stemming on the tokenized data and the results are visualized in word cloud and word frequency table. We then convert the data representations using bag of words (BOW) technique. And we move into experimental modeling using latent dirchlet allocation (LDA) to build a better model. Finally we start extracting the review sentiments by using textblob based unsupervised algorithm and score the data for customer likeliness and unlikeness.

Keywords: Voice of Customers, Natural Language Processing, Bag Of Words and LDA

I. INTRODUCTION

A widely used method to know about customer feedback is Voice of Customer (VOC). However, in the era of digitization, VOC is often available in the form of reviews and comments in various discussion forums or web portal. This is often unstructured in nature and can be humongous for popular products. This makes the task of identifying overall customer feedback and satisfaction level extremely difficult.

In this work a framework for the development of VoC template from unstructured data is proposed. VoC in general and a specific case study for demonstrating the efficacy of the proposed framework are presented.

This project focuses on **MG Hector – a newly launched vehicle in India**. It aims to provide the recommendation to its manufacturer by analyzing the unstructured text data available in Team BHP portal.

The outcome of this study would give us a list of most talked about features that customers liked or disliked about a particular car.

II. RELATED WORK

The Voice of the Customer (VoC) is a process of capturing customer requirement's which produces a detailed set of customer wants and needs organized into a hierarchical structure, and then prioritized in terms of relative importance and satisfaction with current alternatives" [1]. In the olden days, Voice of Customer (VoC) was done through phone calls by large call centers. But with rapid adoption of mobile technologies, VOC can now be captured more efficiently, quickly and less expensively. The timelines have been cut down from months to hours, costs have been trimmed to almost half. Since everyone has access to a cell phone, market research companies can work using phone calls, texts, and emails all to one device [2]. And then we have online forums, blogs, and reviews, tweets readily available through social media channels. Identifying customer requirements through VOC by analyzing opinions, comments and reviews have been a prime focus of many industries. This is usually done through the technique called “Text Mining”.

Text mining is a technique to detect information that might not get recognized when extracted from various text-based sources. Here, an unstructured datasets like text documents, emails, and web files can be handled using text mining technique. In many instances the text mining methodology is repeated until the required information is mined [3].

According to a study in [4] text analytics generally refers to a process of extracting the interested information from a unstructured text which is further used to study word frequency distributions also called lexical analysis, pattern recognition, information extraction, tagging/ annotation, and data mining techniques including link and association analysis, visualization and predictive analytics [5].

A key element in text mining is the linking together the extracted information to form some new facts or new hypotheses for further exploration by means of experimentation and the final aim is to discover the unidentified information [6].

Recently Text Mining has quickly evolved by adding methods able to classify documents according to their latent topic or to infer about the “sentiment” of customers or the users of social networks. The boost to these approaches have gone along with the evolution of both the computational efficiency of the algorithms necessary to analyze textual data and the technology needed to store information [7].

A whole lot of work has been done on Sentiment analysis using product review data [8], Opinion Mining and Sentiment Analysis [9], Product weakness finder [10], Sentiment Analysis: A Multi-Faceted Problem [11], Comparative Experiments on Sentiment Classification for Online Product Reviews [12], which form the basis of our work.

However, Sentiment Analysis remains a very challenging proposition since our knowledge of the problem and solution are very limited since it deals with Natural Language Processing (NLP). Also relying too much on Machine Learning Algorithm produce no human understandable results [13]. Since the comment section is used for detecting sentiment and using lexical resources to capture information about the informal/casual language used in web forums, usage of sentiment lexicon proved useful to determine positive/negative/neutral sentiment [14]. Akilesh et.al have performed “TextBlob” by studying the existing frequency of the word and how often they appear in a positive/negative/neutral context and decide the polarity [15].

The exhaustive literature survey indicates the lack of effective frameworks/templates for carrying out the VoC analysis from unstructured data. In this work we propose a framework/template to perform the VoC on unstructured data and demonstrate the outcomes with a case study in a specific sector.

This project focuses on Indian Auto industry, one of the fastest growing market in the world, contributing to a large share in Indian economy [16]. India is poised to be world’s third-largest passenger-vehicle market by 2021. Currently, the auto industry contributes more than 7% to India’s GDP. According to the Automotive Mission Plan 2026, the government and industry set a target to triple industry revenues, to $300 billion, and expand exports sevenfold, to $80 billion. To meet these aims, it is estimated that the sector could contribute more than 60 million additional direct and indirect jobs, and the result could be improved manufacturing competitiveness and reduced emissions” [17].

As a result many new entrants are vying for a slice in such a huge market. This will increase competition, reduce product lifecycle and give customer more choice. Now to move and adapt fast and gain upper hand in this dynamic and challenging market, it’s very important how well an industry understands and responds to customer needs.

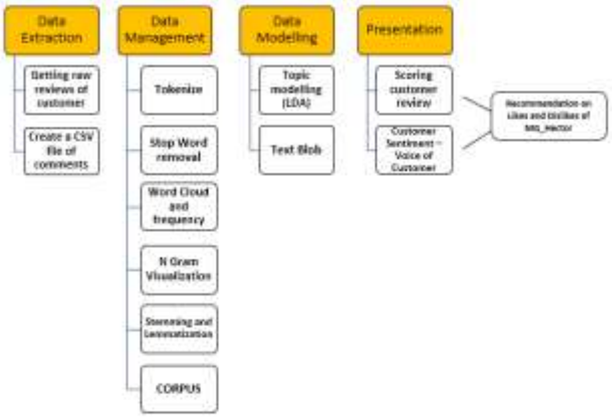
This is where our study can be of immense importance as it seeks to identify the features that customers are expecting or not expecting by analyzing the structured and unstructured text data available in open source social networks (OSN) like Team BHP portal which is known for its unbiased reviews and comments [18]. This would help car manufacturers design their future product better and upgrade existing products to meet the needs of customers.

III. METHODOLOGY

A. Framework

The Voice of the Customer is a product development technique that is used to collect feedback from consumers. Since customer experience is the major differentiating factor against rival companies or competitors, we collect data from a relevant website and do analysis on the product’s review and get back with a sentiment analysis with respect to the product.

Fig.1: Flow diagram of the project on VoC



B. Data Collection

The data for the study is extracted from website TEAM-BHP.com with the help of web scrapping tools. The content of the data set will have the VOCs (Voice of Customers) gathered through feedback and reviews on the TEAM-BHP.com subsequently.

The data collected contains 1624 comments, which is quite huge for performing manual analyses in a systematic or a productive way. The reviews are checked for quality and preprocessed as discussed in the next section.

C. Text Pre-Processing and Data exploration

After collection of data we have started with pre processing and data exploring by following below methods

- Data labeling - As we don’t have any labeled data we classified the sentiments by using unsupervised approach to find the sentiments of each customer review.
- Data Cleaning – also known as pre-processing procedure includes Stop word removal using available libraries and Custom Stop word removal. The kind of words to be removed will typically include words that do not have much semantic value (e.g. the, it, a, etc). Additional, specialist words also need to be removed.
- Data exploration- It is usually done to identify the most common repeated words and uncommon words and visualize in word cloud and word frequency.
- Stemming and Lemmatization – Stemming reduces the word to its derived form. It just is an equal to or smaller form of the word. Lemmatization involves resolving words to their original form.
- N-gram –The N-Gram tokenizer breaks the given sentence down into set of individual words whenever it comes across one of a list of specified characters or words. Then it emits N-grams of each word of the specified length.

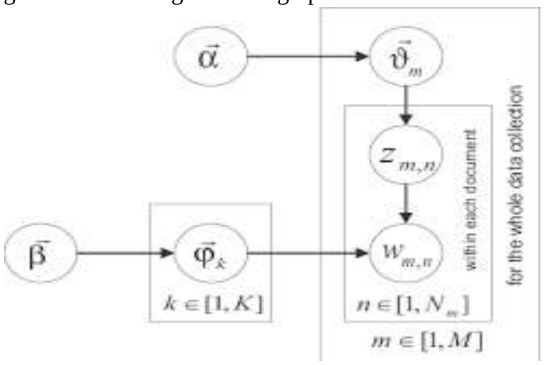
D. Unsupervised approach using TextBlob

TextBlob is a classifier using multiple algorithms [20]. The input to text blob will be a processed string and the output will be a sentiment score between -1 for negative to +1 for positive and zero for the neutral.

E. LDA (Latent Dirichlet Allocation)

LDA Approach is an unsupervised learning algorithm to identify the topics that are hidden in a documents. Here we are trying to extract the topic relating to MG hector – a newly launched car in the Indian market. LDA was used to extract the important topics from the customer reviews.LDA is closely related to the probabilistic latent semantic analysis (pLSA) by Hofmann [14], a probabilistic formulation of LSA. However, it has been pointed out that LDA is more complete than pLSA in such a way that it follows a full generation process for document collection [12, 16]. Models like pLSA, LDA, and their variants have more successful applications in document & topic modeling [12, 16], dimensionality reduction for text categorization [12], collaborative filtering [15], ad hoc IR [17], entity resolution [18]. Here we are using LDA to identify hidden topics from 1623 customer reviews and topic can be used to get feelings about customer review. Also this provides an understanding what is most important concerns which customers are facing.

Fig. 2: Diagram for LDA: a generative graphical model

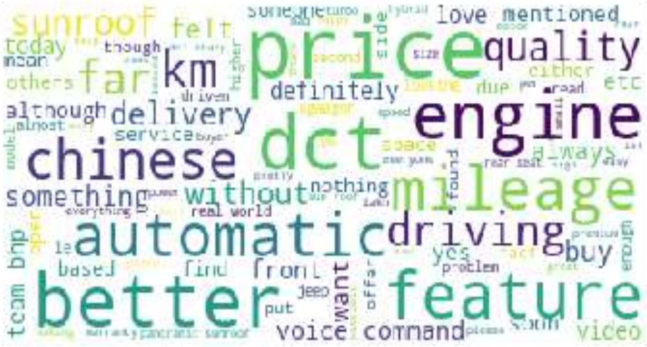


IV. EXPERIMENT

A. Visualising in Word Cloud-

Most Frequent Words are visualized in Word cloud

Fig. 3: Word Cloud



B. Ngrams visualization

Once we created cleaned Corpus, we proceeded with N grams for visualizing the data. One can think of an N-gram as the sequence of N words, by that notion, a 2-gram (or bigram) is a two-word sequence of words like “please turn”, “turn your”, or ”your homework”, and a 3-gram (or trigram) is a three-word sequence of words like “please turn your”, or “turn your homework”.

C. Visualize top N Uni-grams, Bi-grams, Tri-grams

The below images were frequency distribution for Uni-gram, Bi-gram and Trigram

Fig.4:Uni-gram

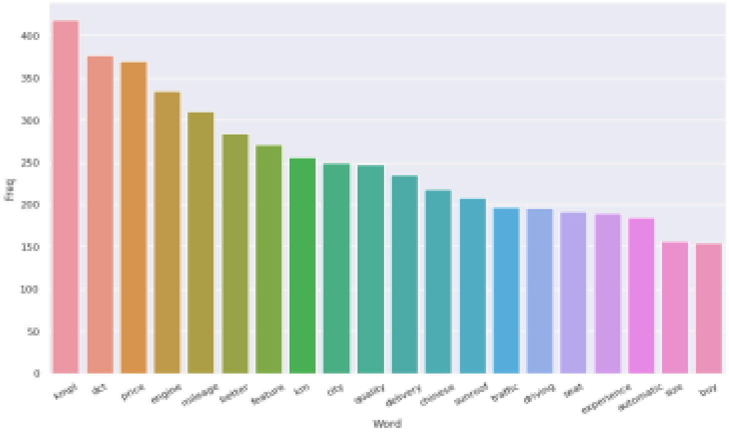


Fig. 5: Bi-Gram

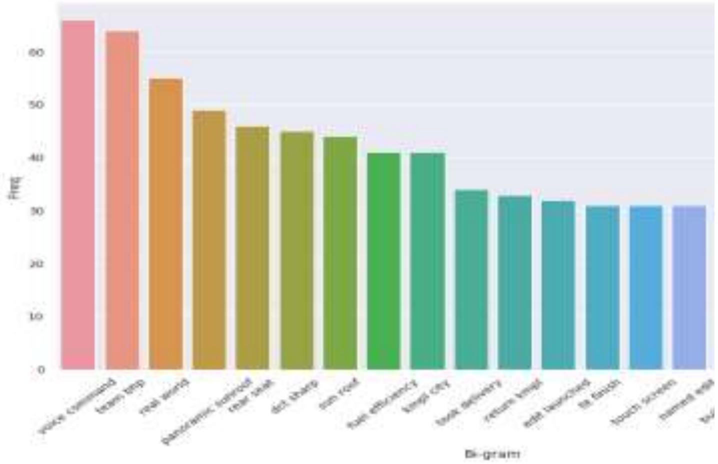
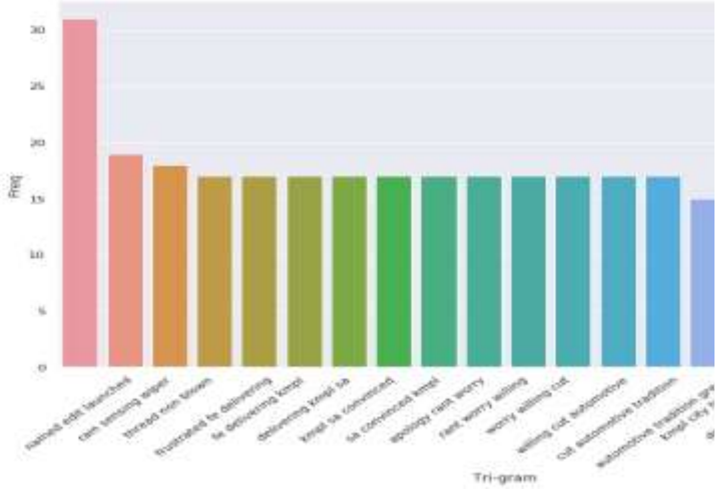


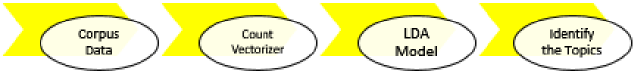
Fig. 6: Tri-gram



A. LDA (Latent Dirichlet Allocation)

LDA Approach is an unsupervised learning algorithm to identify the topics that are hidden in a document. Here we are trying to extract the topic which related to MG hector car that is newly launched into the market. LDA was able to extract the important topics from the customer reviews. We started by creating a count vectorizer for our cleaned corpus data. And this count vectorizer is the input for our for LDA topic model. In this LDA model we were able to identify five topics and each review is classified into this topic by LDA model. In below FIG 7 explains the working flow for the topic model

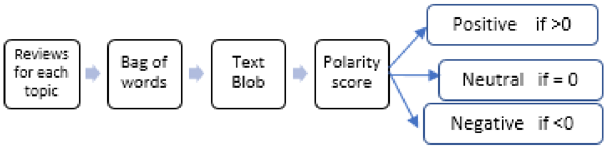
Fig. 7: Topic Model Flow



B. Unsupervised Textblob Approach to Find Sentiment Score

We have identified 5 major topics using LDA about the MG hector from the customer review. Here we are identifying customer feelings about each topic by performing a sentiment analysis. As per the model flow given in FIG 8 below, we have five topics and their reviews. This has been converted to a bag of words which is an input feature for our unsupervised Textblob approach. We identified sentiment polarity for the customer review in each topic. Using these scores we will be able to identify the customer feelings. We classified the customer review by score > 0 as “Positive” and score < 0 as negative and equal to 0 as Neutral. In this paper we are focusing only positive and negative sentiments hence we could able to identify customer feelings.

Fig. 8: Sentiment Analysis Model Flow



V. RESULT

A. Topic Modeling Result

TABLE I . Frequent Words in Each Topic

Topic 1 Most Frequent 10 Words	Topic 2 Most Frequent 10 Words	Topic 3 Most Frequent 10 Words	Topic 4 Most Frequent 10 Words	Topic 5 Most Frequent 10 Words
engine	well	suv	light	petrol
road	vehicle	mid	panel	india
much	friend	time	issue	get
time	road	drive	plastic	like
well	issue	traffic	attachment	engine
price	guy	even	gap	vehicle
good	http	highway	people	year
Like	screen	experience	vehicle	even
Dct	car	india	seat	car
Mg	get	city	quality	price
drive	like	hector	like	diesel
petrol	delivery	km	sunroof	month
diesel	booking	mg	look	booking
seltos	hector	mileage	hector	hector
hector	mg	kmpl	harrier	mg

We were able to find out the five important topics by analyzing frequent repeating words in each topic by referring Table 1 and the customer reading customer review manually. Here are the topic results listed below.

First Topic: First Topic is mainly talking about the engine, road, much, well and all other words showing that Topic will be: **Mg Hector Engine performing well**

Second Topic: Words such as vehicle, road, issue etc tells us that Topic will be: **MG Hector issues and mostly related delivery delay**

Third Topic: Words such as SUV, mid, time, drive, traffic, highway tells us that Topic mostly related to **MG hector drive in city and Traffic also fuel efficiency.**

Fourth Topic: light, panel, issue, plastic, seat, quality and these words suggests that mostly topic is related to **MG hector accessories and quality of each items.**

Fifth Topic: Petrol, India, hector get The Topic mostly related to review **Comparison of Petrol and Diesel car and Also Name MG hector.**

B. Result of Sentiment Analysis

By performing sentiment analysis, we obtained positive and negative reviews. Table II illustrate the counts. Our key focus lies on Positive and Negative reviews. Neutral comments are ignored.

TABLE II. Sentiment Counts in Each Topic

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Positive	397	194	177	165	278
Negative	77	48	26	50	49

We are listing out important positive and negative findings about MG hector in below results

The comments such as "**No body roll, steering handling, awesome drive smooth noise"** which points out the fact that the **performance of the car is good.** Also, customers talk about **"excellent pricing hope jeep something reduce price".** Here customers talk about excellent pricing of MG while comparing with competitors like Jeep. Next comment is about customer positive feelings about the "**road presence better than Tata and Mahindra".** As MG delivery time high, **"MG providing redeem points for service** "which is good for the customer who are in waiting list**. MG's British** Heritage is positive factor for the customers who wish to buy the product. "**Panoramic sunroof mg hector** "is another positive point of MG. Moreover "**infinity jbl audio and internet connectivity** "are positive features which customer mostly feel good about MG hector.

"Delay in delivery "is the main negative factor about the Mg Hector**. "Breakdown issues** "are mentioned in comments which is also negative. Also "**fuel efficiency shockingly poor in city** "is an important factor customers are talking about in the reviews. Moreover customer feels unsafe about the car **"sunroof"** considering probable theft attack which can happen breaking the glass.

Also Indian customer wishes to have Hindi voice command system inside the car. Addition of multiple local language voice command system is recommended.

VI. CONCLUSION AND FUTURE WORK

This paper focuses on voice of customer as key decision making information in the business. We have tried to extract hidden topics to identify what customers are talking about MG hector. This helped us to find out overview of customers' discussion points related to MG hector. Later we used unsupervised techniques to find out customer sentiments. By finding customer opinion or sentiments we are able to identify what are the most talked about positive features like "Avoiding body roll, handling, drive, looks, redeem offers, British heritage, sunroof, price and internet connectivity".

In future, enabling Indian local languages voice command systems in the car would make a positive impact in customers' sentiments. Moreover, we are able provide recommendation to MG hector which can impact customer positively.

The scope for further work includes - N gram model instead of words to identify topics and sentiments and to achieve a better result. And we can extract hidden features by doing LSA and K means clusters. Also, we can use word embedding using word2vec to include the context.

REFERENCES

[1] *Steven P. Gaskin, Abbie Griffin,John R. Hauser, Gerald M. Katz, Robert L. Klein, "VOICE OF THE CUSTOMER", https://www.mit.edu/~hauser/Papers/Gaskin_Griffin_Hauser_et_al%20VOC%20Encyclopedia%202011.pdf*

[2] *https://www.driveresearch.com/single-post/2017/12/05/The-What-How-and-Why-of-Voice-of-Customer-VoC-Market-Research*

[3] *H Cui, V Mittal, M Datar, "Comparative experiments on sentiment classification for online product reviews", - AAAI, 2006 - new.aaai.org*

[4] *Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan, "Using Text Mining Techniques for Extracting Information from Research Articles", - https://link.springer.com/chapter/10.1007/978-3-319-67056-0_18*

[5] *NiharRanjan, Abhishek Gupta, IshwariDhumale, PayalGogawale and *RugvedGramopadhye, "A SURVEY ON TEXT ANALYTICS AND CLASSIFICATION TECHNIQUES FOR TEXT DOCUMENTS", - https://www.journalijdr.com/survey-text-analytics-and-classification-techniques-text-documents*

[6] *Marti Hearst , " What Is Text Mining? ", - http://people.ischool.berkeley.edu/~hearst/text-mining.html*

[7] *Diego Zappa, MattiaBorrelli, Gian Paolo Clemente, Nino Savelli , "Text Mining In Insurance: From Unstructured Data To Meaning" - https://www.variancejournal.org/articlespress/articles/Text_Mining-Zappa-Borrelli-Clemente-Savelli.pdf*

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

[9] *Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.*

[10] *Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Systems with Applications, 39(11), 10283-10291.*

[11] *Liu, B. (2010). Sentiment analysis: A multi-faceted problem. IEEE Intelligent Systems, 25(3), 76-80.*

[12] *Cui, H., Mittal, V., &Datar, M. (2006, July). Comparative experiments on sentiment classification for online product reviews. In AAAI (Vol. 6, No. 1265-1270, p. 30).*

[13] *Bing Liu, Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, 2010*

[14] *EfthymiosKouloumpis, Theresa Wilson, Johanna Moore, Twitter Sentiment Analysis:The Good the Bad and the OMG!, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.*

[15] *Akhilesh Kumar Singh, Deepak Kumar Gupta and RajMohanSingh, Sentiment Analysis of Twitter User Data on Punjab Legislative Assembly Election,2017, I.J.Modern Education and Computer Science,2017, 9, 60-68,2017*

[16] *https://www.grantthornton.in/globalassets/1.-member-firms/india/assets/pdfs/indian_auto_industry_2.0.pdf*

[17] *http://www.forbesindia.com/blog/business-strategy/indian-automotive-industry-the-road-ahead/*

[18] *https://www.team-bhp.com/aboutus/overview*

[19] *RemiLebret and Ronan Collobert N-GRAM-BASED LOW-DIMENSIONAL REPRESENTATION FOR DOCUMENT CLASSIFICATION. A Paper based N gram based Low Dimensional Representation 2015*

[20] *D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. JMLR, 3:993–1022, 2003*

[21] *S. Deerwester, G. Furnas, and T. Landauer. Indexing by latent semantic analysis. Journal of the American Society for Info. Science, 41(6):391–407, 1990.*

[22] *T. Hofmann. Probabilistic LSA. Proc. UAI, 1999.*

[23] *T. Hofmann. Latent semantic models for collaborative filtering. ACM TOIS, 22(1):89–115, 2004*

[24] *T. Griffiths and M. Steyvers. Finding scientific topics. The National Academy of Sciences, 101:5228–5235, 2004.*

[25] *X. Wei and W. Croft. LDA-based document models for ad-hoc retrieval. Proc. ACM SIGIR, 2006.*

[26] *I. Bhattacharya and L. Getoor. A latent Dirichlet model for unsupervised entity resolution. Proc. SIAM SDM, 2006*

[27] *Ali Hasan , Sana Moin , Ahmad Karim and ShahaboddinShamshirband Machine Learning-Based Sentiment Analysis for Twitter Accounts*