Optimization Of Social Media Comments To Improve Customer Journey Using Machine Learning

Mr. Tejas Sanjay Chougule
Department of Information Technology
Shah and Anchor Kutchhi Engineering College
Mumbai, India.
tejas.chougule1994@gmail.com

Ms. Swati Nadkarni
Department of Information Technology
Shah and Anchor Kutchhi Engineering College
Mumbai, India.
swati.nadkarni@sakec.ac.in

Dr. Bhavesh Patel
Principal
Shah and Anchor Kutchhi Engineering College
Mumbai, India.
bhavesh.patel@sakec.ac.in

2020 IEEE Bombay Section Signature Conference (IBSSC) | 978-1-7281-8993-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/IBSSC51096.2020.9332188

***Abstract*— The marketing is carried out by using various social media strategies and platforms like Facebook, Instagram, Twitter, Pinterest, LinkedIn, YouTube, etc. these are various platforms that are used for marketing. The audience satisfaction delivers many benefits like loyalty, an increase of being referral, less likely to churn, repeat purchase, buying the product at a premium price. The objective of this project is to analyze customer comments in order to extract product details, issue type, sentiments/emotion using topic modeling which will also showcase keywords fall under particular topics to improve customer satisfaction scores. The customer journey can be analyzed to understand the needs and requirements of customer which when post purchasable of the product also help in understanding customer fulfilment ratio. This project not only helps to understand that promotion through social media is better than the traditional approach but also helps to understand the adaptation of new social media x strategies and their promotion along with customer satisfaction.**

Keywords—Social media, Customer satisfaction, Reviews extraction, Topic modeling.

I. INTRODUCTION

The marketing strategies includes methods like radio and TV ads, hoardings, business cards and in numerous other comparable ways where Internet or online networking sites were not utilized for promotional activities as in business owner publicize their services or products on print media[1][2]. Traditional marketing strategies had constrained customer reach-ability and the extent of driving customers' purchasing patterns. Today's business depends on customers and on their demands. Nowadays people believe more on social media for buying any products, goods, electronic devices rather than traditional methods of blindly purchasing things without gathering knowledge of it. The Web is a huge space to express and share individual opinions, influencing any aspect of life, with implications for marketing and communication alike[3]. Reviews and ratings on the Internet are increasing their importance in the evaluation of products and services by potential customers[2][4][5].

Social Media are influencing consumer's preferences by shaping their attitudes and behaviors. The influence of social media on people purchasing has grown over the years. Social media activities play a vital role in keeping track of customer journey, sentiment towards products, campaigning impact and success, also to analyze top and best influencers related with respect to brands, products or campaigns[6]. Product Matching or items purchased by customers based upon their interest had given rise to both customer information and product availability. A recommender system is used for matching items with customers on the basis of historical behavior. Mostly there are 2 types of recommender systems: The first type is collaborative filtering which helps to find the closeness of user's with a preference of other users but ignores specific properties of a user[7][8]. Next type is Content-based filtering which is primarily based on properties, the content of the user. However, the recommender system is used for recommending relevant products or assets either by using marketing strategies or by using social media networking. The simplest method for identifying relevant users (i.e. users whom we want to target with advertisements) would be to manually construct rules detecting whether the given messages match particular patterns associated with the products or brands that will be advertised[9][10][11]. However, this is a time-consuming approach that may require significant experience with the message and user patterns on Twitter. Therefore, in addition to using a subset of manual rules called text extractors (for which the tradeoff is computational efficiency versus more detailed targeting rules), for which develop a method that can construct probabilistic relevance models automatically based on product descriptions and then tune the models based on the observed performance.

This presents two significant challenges. First, a single product catalog can contain numerous types of products. Second, the type of language used in product catalogs can be quite different from the language used in social media. The main KPI are Attract, Convert, Close, Delights. Social media marketing uses various techniques like Search Engine Optimization, Search Engine Marketing, Email Marketing, Content Marketing, SMS marketing, Video marketing/video infographics. so as to communicate with target customers which may help to understand customer limitations and needs. Search Engine Optimization (SEO) is the practice of increasing the number and quality of visitors to a website by improving rankings in the algorithmic search engine results[10][11][12]. SEO, therefore, involves making sure a website is accessible, technically sound, uses words that people type into the search engines, and provides the excellent user experience, with useful and high quality, expert content that helps answers the user's query. SEO includes keyword research, Link building, Content delivery, Analytics. Search Engine Marketing also includes paid online advertising models, such as pay-per-click (PPC). PPC advertising is those such as Google AdWords and Bing, which only require

payment when the ad is clicked through to the target website. SEM also requires keyword analysis as the words and phrases used in the ad and site. Email marketing is a very effective tool, despite claims that it isn't as important as social these days[13][14]. Modern email marketing is just beginning to evolve so that it can be linked to a database in order to personalize it so that individual groups of customers can be sent mail based on previous purchases and interests. Content marketing is a technique where content is produced and distributed with the intention of providing relevant, interesting content to attract and engage a particular audience that a business is targeting[15]. The goal is to win customer loyalty and retain them.

Beyond this, it is important to consistently monitor and analyze the results from your efforts. Using this data-driven marketing approach to your content marketing will ensure you achieve the best results possible.

II. LITERATURE SURVEY

In an era of a machine, it is been sighted that slightly manual work is beginning to replace by machine for doing daily activities. Here the machine learning plays a vital role in making understanding the data to a machine or computer. Machine learning is mostly divided into two types of learning supervised and unsupervised learning[16][17][18]. Supervised learning is a learning of data from its past behavior that map input with output variable and to predict for new input data from the trained model while forming clusters or finding an association amongst input data is termed as unsupervised learning[19][20]. Machine learning can be used in various domains like marketing, banking & insurance, education, gaming, bioinformatics last but not least malware detection [7][19][20].

It is very difficult to find a loyal customer in each domain like retail, travel, education. Nowadays big companies are trying to find loyal and potential customers through social media via customer tweets or reviews against products[21][22]. They also find likelihood users as per user to user to behavior or on the content-based recommendation in order to do campaigning across email, messages, APN or BPN where users will be keep posted for new arrivals of a similar product.

A. Existing System

There are many existing systems all of them are less effective than the proposed system. Some of the existing systems are as follows:

The social network is increasing day by day and data generated from social media is huge in amount. This paper entails about topic modeling carried out on twitter data. Starts with data acquisition of twitter data using rest API, carried by text pre-processing steps like tokenization, stop word removal and many more. This processed text data is used in text modeling where word2vec comes into the picture. After word2vec, topic modeling i.e. LDA plays a role in extracting topics and subtopics from the corpus. At last data, exploration is carried out to showcase the topic and subtopics distribution[15][23].

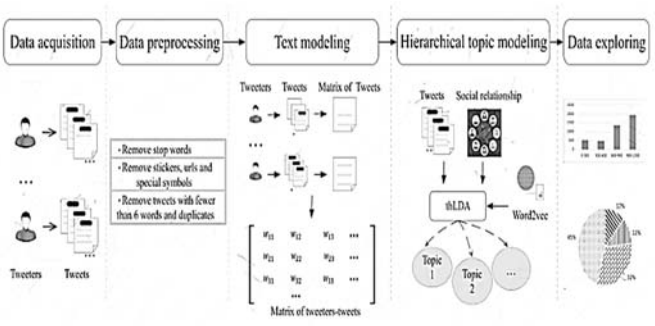

Fig. 1: The analysis carried out on twitter data

Above is a working architecture of topic modeling on twitter data which shows its whole working on the basis of its respective module.

$$P(c_m, z_{m,w}|\alpha, \beta, \gamma, Y, Y', H, W_m) = P(c_m|W_m, c_{-m}, z, \gamma, \beta, Y, H) x P(z_{m,w}|z_{-(m,w)}, W_m, Y', \alpha, \beta) \quad (1)$$

Equation 1 tells about topic modeling which is based on two parameters tweet document from topic tree and topic number of each collected token from tweet document collection.

$$P(c_m|c_{-m}, \gamma, Y_{k1}, Y_{kl+1}, H) = \frac{N_{(kl+l.j')}}{\gamma + m - 1} * Y_{kl,j}, K_{l+1,j'} * H \quad (2)$$

$$P(c_m|c_{-m}, \gamma) = \frac{\gamma}{\gamma + m - 1} \quad (3)$$

Equation 2 & 3 describes about probability distribution of tweet document while selecting from each node layer each node consist of two piece of data i.e. topic and tweet document in order to select node.

$$Y_{k1,k_{i+1}} = \frac{\sum_{i=1}^{n} f_{k_{l,i}} * \frac{\left(\sum_{i=1}^{n} f_{k_{l+i,j} * im(q_{k_{l,i}}, q_{k_{l+1,j}})}\right)}{\sum_{i=1}^{n} f_{k_{l+1,j}}}}{\sum_{i=1}^{n} f_{k_{l,i}}} \quad (4)$$

Equation 4 defines about calculating and extracting top n tokens in order to present their collection of their frequencies where each item give occurrence of words.

$$Sim(w_1, w_2) = \cos(V_1, V_2) = \frac{\sum_{i=1}^{x} (V_{1,i}, V_{2,i})}{\sqrt{\sum_{i=1}^{x} V_{1,i}^2} * \sqrt{\sum_{i=1}^{x} V_{2,i}^2}} \quad (5)$$

Equation 5 states about word2vec calculation consider two words obtain similarities for each word separate vectors are formed which is further used for finding word topic semantic that denotes frequency of occurrence of word with respect to topic[5][23][24].

The customer journey varies according to e-commerce sites. Multiple e-commerce sites follow a similar journey path in achieving their target. Customer journey is defined as customer interacting with the service provider in a given path and customer which follows a particular path multiple times till the purchase of the product is the loyal customer. The approach followed in this paper[6][25]: starts

with exploratory data analysis on dataset next baseline customer information is used for prediction and recommendation. After data determination data pre-processing is carried which includes data normalization, scaling, dealing with missing values, removing outlier & so on. Once processed data achieved further it used for prediction of the future on the basis of prediction.

To evaluate some metrics are used are as followed:

$$Recall = \frac{\#CorrectPredictions}{\#Items} \quad (6)$$

$$Precision = \frac{\#CorrectPredictions}{\#Predictions} \quad (7)$$

$$AP = \sum_{i=1}^{n} Precision(i)\, \Delta Recall(i) \quad (8)$$

$$MAP = \sum_{i=1}^{n} AP(i)/n \quad (9)$$

$$MAF1 = 2 * \frac{MAP*Recall}{MAP+Recall} \quad (10)$$

Equation 6, 7, 8, 9, 10 indicates Recall, Precision, Average Precision, Mean Average Precision, and Mean Average F1 score for evaluating the model & result[25][26][27].
These prediction are used for outlining very next event in customer journey for this purpose MAF1 is calculated to get top 5 prediction for each 3 predictor.

III. Implementation

This System contains complex scraper which is able to scrape all the necessary content from the webpage defined in the code for getting a very good amount of insight from the data. The System is not just a scrapper but also a data analyzer for insight extraction.

There's a section of review at every Amazon's page where each and every person is allowed to comment on the products review. As day-by-day this review information is keeping on increasing. This data basically show's costumer experience after buying that particular product. It helps the other customers for making decision regarding buying that product. This system have extracted all the products id's which is called "asin" number on Amazon and its corresponding reviews.

The system consist of module like Data gathering, Text preprocessing, Document matrix, Topic modeling, Data exploration; let's see one by one module in brief:

- Data Gathering

  It forms a tree structure. For accessing any node in the tree, it is supposed to iterate through the entire tree. If considering an example of accessing the span element from the tree of html, need to go through the following path:

  Html-body-div(left)-div(left)-span

  This is the way in which the html elements are accessed. Similarly accessed the elements of Amazon with the same algorithm of pattern matching. For pattern matching provided "xpath" for every required element, and then extracted those element with those "xpath".

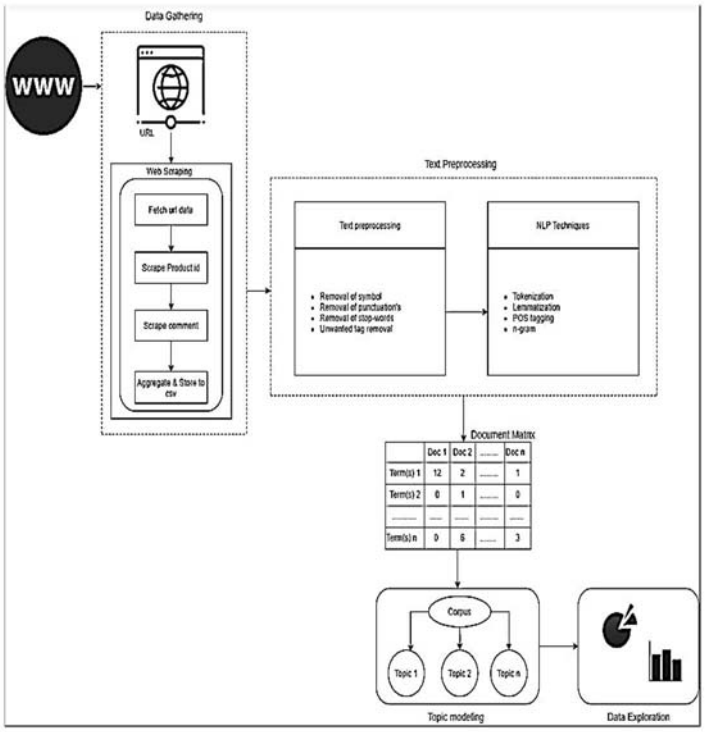

Fig. 2: System architecture

- Text preprocessing

  Text are termed as unstructured data to deal with huge amount of text processing need to be carried out in order to achieve processed data for analyzation. As machine don't understand the human language, to make data understandable to machine NLP is used. NLP stands for Natural Language Processing and it contains multiple techniques that help in breaking down of text data into a machine-understandable format to perform analysis. Below are the mentioned NLP techniques:

  i. Word/Sentence Tokenization: This is an early step of the NLP process that splits longer paragraph of text into smaller pieces or sentence these sentences are tokenized into tokens or words to perform further processing[25][29].
  ii. Stop Word Removal: Stop words are those words which are filtered out before further processing of text, since these words contribute little to overall meaning, given that they are generally the most common words in a language[25][29][30].
  iii. Normalization: Before further processing, the text needs to be normalized. Normalization generally refers to converting all text to the same case (upper or lower), removing punctuation, expanding contractions, converting numbers to their word equivalents, and so on.
  iv. Stemming & Lemmatization: Stemming is the process of eliminating affixes (suffixed, prefixes, infixes, and circumfixes) from a word to obtain a word stem and Lemmatization is it brings down particular word to root form.

v. Parts-of-speech (POS) Tagging: POS tagging consists of assigning a category tag to the tokenized parts of a sentence.

vi. For calculating frequency, but sometimes what happens is word signify more meaning when they are in the group of two or three known as unigrams the only one word whereas in BIGRAMS and TRIGRAMS two and three words respectively. This sometimes helps model to gain more information about the context provided.

- Document Matrix

A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the corpus and columns correspond to terms.

- Topic Modeling

Topic modeling is a way of extrapolating backward from a collection of documents to infer the discourses ("topics") that could have generated them.

From the entire document frequency of each word is calculated and the words with maximum frequency are taken as important keywords, but sometimes what happens is there are many words like "art", "artistic" and artist, here all the three words signify the same concept of "art". So, stemming and lemmatization is used for dealing with this problem. For each possible topic Z, we'll multiply the frequency of this word type W in Z by the number of other words in document D that already belong to Z. The result will represent the probability that this word came from Z. Here's the actual formula:

$$P(Z|W,D) = \frac{\#of\ word\ W\ in\ topic Z + \beta_w}{total\ tokens\ in\ Z + \beta} . (\#words\ in\ D\ that\ belong\ to\ Z + \alpha) \quad (11)$$

For topic modeling LDA algorithms is used to find topics along with keywords, wherein LDA stands for Latent Dirichlet Allocation.

Latent: This refers to everything that is not known a priori and are hidden in the data. Here, the themes or topics that document consists of are unknown, but they are believed to be present as the text is generated based on those topics.

Dirichlet: It is a 'distribution of distributions'. In the context of topic modeling, the Dirichlet is the distribution of topics in documents and distribution of words in the topic. It might not be very clear at this point of time, but it's fine as we will look at it in more detail in a while.

Allocation: This means that once we have Dirichlet, we will allocate topics to the documents and words of the document to topics.

LDA says is that each word in each document comes from a topic and the topic is selected from a per-document distribution over topics. So there are two matrices:

1. Θtd = P(t|d) which is the probability distribution of topics in documents

2. Φwt = P(w|t) which is the probability distribution of words in topics

The probability of a word given document i.e. P(w|d) is equal to:

$$\sum_{i \in T} p(w|t,d)p(t|d) \quad (12)$$

where T is the total number of topics. Also, let's assume that there is W number of words in our vocabulary for all the documents.

Gibbs Sampling: It is an algorithm for successively sampling conditional distributions of variables, whose distribution over states converges to the true distribution.

The mathematical manner, what it is doing is trying to find conditional probability distribution of a single word's topic assignment conditioned on the rest of the topic assignments. Ignoring all the mathematical calculations, what will get is a conditional probability equation that looks like this for a single word w in document d that belongs to topic k.

There are two parts in this sampling. First part tells us how much each topic is present in a document and the second part tells how much each topic likes a word.

IV. Experiment Results

In this system, Optimization of Social Media Comments to Improve Customer Journey Using Machine Learning, at first data is scraped from e-commerce site (Amazon) for any of the one domain, as in testing is done on television domain. After scraping whole extracted undergoes whole process in order to get top n topics in that domain. Below is the screenshot of user interface of system

The first step is to input the URL into text box which scrapes the asin no. i.e. product id from given webpage. Next it scrapes reviews with respect to each asin no, as there are multiple reviews for each item with multiple pages it bit takes time. Below is the screenshot of extracted asin no stored in list for scraping respective reviews. Once reviews are scraped for each asin number they are stored in a flat files and aggregated into one file which is further used for analysis.

These reviews goes under text preprocessing where tokenization, removal of stop words, lemmatization, normalization, POS tagging is performed in order to get processed data for further analysis. The term document matrix is calculated to find the frequency of occurrence of words in corpus. After Term document frequency matrix Topic modeling using LDA is performed on the generated corpus. It is a matrix factorization technique. In vector space, any corpus (collection of documents) can be represented as a document-term matrix.

In LDA models, each document is composed of multiple topics. But, typically only one of the topics is dominant. The dominant topic for each sentence and shows the weight of the topic and the keywords in a nicely formatted output. After

seeing the dominant topic it's time to check a word cloud with the size of the words proportional to the weight is a pleasant sight. The coloring of the topics taken here is followed in the subsequent plots as word cloud.

V. Evaluation Results

Now let's see how good the model is by using an evaluation metric i.e. Perplexity and Coherence Score. The perplexity metric as measuring how probable some new unseen data is given the model that was learned earlier. Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. Computed both metric as shown in table:

Perplexity:	-11.138445959692246
Coherence Score:	0.5694047777600276

Table 1: Model Evaluation

After all the processing it's time to explore the topic which will properly help to recognize the total number of topics and keywords present under those topics. As the domain is fixed it will help to understand customer issues, any warranty related issues or any power issues. As this analysis is done on television domain, this system will help to figure out customer limitations, satisfactions, if they have certain needs or they are facing some discrepancy from customer care or they not providing good service at the given time lots of issues can be solved using this system. Last but not least will check the exploration of topic or subtopics along with their keywords occurrence in the whole document as is with highlighting the first topic which will filter words and their occurrence.

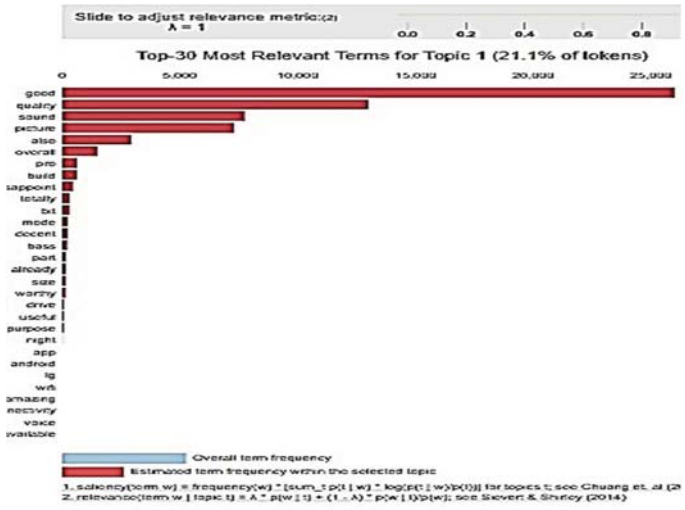

Fig. 3: Highlighting first topics-keywords

VI. Conclusion

The system Optimization of Social Media Comments to Improve Customer Journey Using Machine Learning, the aim is to detect a customer issue faced concerning the product in respective domains mentioned via reviews, which will also use in a brief understanding of warranty issues, power-related issue and much more positive semantic. To implement and measure the performance of the system, the real-time Amazon reviews data i.e. customer reviews are scraped from multiple pages and obtained a reasonably good result. This helps in understanding user's needs and problems faced by them. The use of Topic modeling implemented using LDA Algorithm is made to analyze the reviews and identify the topics on which customers are mostly expressing their opinions which saves valuable time of processing each review manually and increasing the speed of the system. The system highlight presentable output of topic & subtopics along with keywords present under those topics defining their frequency of occurrence. The system takes quite a time for scraping data from site but once it is scraped brings result faster showing topics keywords relation.

The future scope of system will be to detect an issue and also be able to recommend the user saving time to take precaution at earliest and also include a faster computation time. Topics will be purely generated by the system as per single product in the domain and this application will be made to run on mobile platforms too.

References

[1] Hanna M. Wallach, "Topic Modeling: Beyond Bag-of-Words," International Conference on Machine Learning, Cambridge, 2006.

[2] Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai, "Topic Modeling with Network Regularization," International World Wide Web Conference Committee, China, 2008.

[3] Xueliang Liu, Raphaël Troncy, Benoit Huet, "Using Social Media to Identify Events," International Conference on Social media, USA, 2011.

[4] Dharmesh Thakkar, Zhen Ming Jiang, Ahmed E. Hassan, "Retrieving Relevant Reports from a Customer Engagement Repository". ICSM, Canada, 2008.

[5] Liangjie Hong and Brian D. Davison, "Empirical Study of Topic Modeling in Twitter," Workshop on Social Media Analytics, USA, 2012.

[6] Deng Bin, Shao Peiji, Zhao Dan, "E-Commerce Reviews Management System Based on Online Customer Reviews Mining," International Conference on Innovative Computing and Communication, Asia-Pacific, 2010.

[7] Marco Pennacchiotti and Ana-Maria Popescu, "A Machine Learning Approach to Twitter User Classification," International AAAI Conference on Weblogs and Social Media, USA, 2011.

[8] Dr. O. Obulesu, M. Mahendra, M. ThrilokReddy, "Machine Learning Techniques and Tools: A Survey," International Conference on Inventive Research in Computing Applications (ICIRCA), 2018.

[9] Vedran Podobnik, "An Analysis of Facebook Social Media Marketing Key Performance Indicators: the Case of Premier League Brands," International Conference on Telecommunications, 2013.

[10] Suresh M, Rahul Mohan, "Application of Social Media as a Marketing Promotion Tool-A Review," IEEE International Conference on Computational Intelligence and Computing Research, 2016.

[11] Jorge Ale Chilet, Cuicui Chen, Yusan Lin, "Analyzing Social Media Marketing in the High-End Fashion Industry Using Named Entity Recognition," International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016.

[12] Jiawei Zhang, Senzhang Wang, Qianyi Zhan, Philip S. Yu, "Intertwined Viral Marketing in Social Networks,", 2016.

[13] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer, "Enterprise Data Analysis and Visualization: An Interview Study," IEEE Transactions On Visualization And Computer Graphics, VOL. 18, NO. 12, 2012.

[14] Mohammad Daoud, S. K.Naqvi, "Recommendation System Techniques in E-Commerce System," International Journal of Science and Research (IJSR), 2013.

[15] C.Emelda, "A Comparative Study on Sentiment Classification and Ranking on Product Reviews," International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1, 2014.

[16] Estee Van Der Walt and Jan Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," IEEE TRANSACTIONS and JOURNALS, 2017.

[17] Yilun Wei, Yingying Lao, Yudai Sato, Dongli Han, "Product-review Classification Combining Multiple Clustering Algorithms," 2019.

[18] Bushra Ramzan, Imran Sarwar Bajwa, Noreen Jamil, Riaz Ul Amin, "An Intelligent Data Analysis for Recommendation Systems Using Machine Learning," 2019.

[19] Jaspreet Singh, Gurvinder Singh and Rajinder Singh, "Optimization of sentiment analysis using machine learning classifiers," Springer, 2017.

[20] Saumya Chaturvedi, Vimal Mishra, Nitin Mishra, "Sentiment Analysis using Machine Learning for BI," IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017.

[21] S. Buckley, M. Ettl, P. Jain, R. Luss, "Social media and customer behavior analytics for personalized customer engagements," International Business Machines Corporation, Volume 58, 2014.

[22] Bogdan Batrinca, Philip C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," Springer, 2014.

[23] Dinjin Yu, Dengwei Xu, Dongjing Wang, and Zhiyong NI, "Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing," IEEE TRANSACTIONS and JOURNALS, 2017.

[24] Chong Wang, David M. Blei, "Collaborative Topic Modeling for Recommending Scientific Articles," Computer Science Department, USA, 2011.

[25] Joël Goossens, Tiblets Demewez, Marwan Hassani, "Effective Steering of Customer Journey via Order-Aware Recommendation," IEEE International Conference on Data Mining Workshops (ICDMW), 2018.

[26] Liaoliang Jiang, Yuting Cheng, Li Yang, Jing Li. Hongyang Yan,· Xiaoqin Wang, "A trust□based collaborative filtering algorithm for E□commerce recommendation system," Springer, 2018.

[27] Rahul Kumar Chaurasiya, Utkarsh Sahu, "Improving Performance of Product Recommendations Using User Reviews," International Conference and Workshops on Recent Advances and Innovations in Engineering, 2018.

[28] Yashvardhan Sharma, Jigar Bhatt, Rachit Magon, "A Multi Criteria Review-Based Hotel Recommendation System," IEEE International Conference on Computer and Information Technology, 2015.

[29] Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas By, "Sentiment Analysis on Social Media," International Conference on Advances in Social Networks Analysis and Mining, Italy, 2012.

[30] Xiaoying Wang, Chengliang Wang, "Recommendation system of e-commerce based on improved collaborative filtering algorithm," 2017.