

A Literature Review on Text Mining Of User Generated Content (UGC):Techniques, Tools and Applications**Mehak**

School of Computer Science & Engineering
Lovely Professional University, Phagwara, Punjab, India
mehakkatnoria93@gmail.com

Tarun Manchanda

School of Electrical Engineering & Computer Science
University of Ottawa, Ottawa, Canada
tarunmanchanda1903@gmail.com

Abstract—The area of text mining is vast. In simpler terms, it is a process of analyzing the text as a natural language and convert it from totally unstructured format to structured or semi structured format. It will be further utilized to understand the relations between different words and predict some hidden correlation between different components of natural language. Machine learning concepts are used over the field of text analytics and helps to distinguish among different sections of text using statistical modelling. This literature review will be highly focusing on various types of user generated data like data generated over social media, blogs, online user communities, speech to text or big data. This paper will talk about various techniques like supervised and unsupervised machine learning, various tools that can be used to perform text analytics, applications and gaps or limitation within existing research.

Keywords: *Text Analytics, Text Mining, Machine learning*

1. Introduction

The world has been changing at the fast pace and so is data in this world. The data is not only changing but also the speed of generating it has reached to extreme. The systems that are being used as source system for the data, social media or machine operated system are generating tons of data that has a lot of information that needs to be analyzed and understand how it is affecting the society and other systems. Major contribution is by user, the data that is being generated specially over social media has a lot of valuable information such as users are writing about different products and companies need to track down all their customer views and opinions [1]. Looking at the research available in the market currently most of the papers are focusing on single type of research method or techniques like syntactic or semantic analysis that might be good but is not the correct way to handle the user generated data

efficiently [2]. We need to customize the techniques, compare different approaches to each other to understand which type of data works better with type of techniques. Also, there are many tools available in market that help to perform the text mining needs to be analyzed to understand the feature set. Many research papers have been identified where the manual work over the large data set took so much time because of the limited resources. The limitations with regards to different tools and applications needs to be formulated together to compare and see which works better with respect to different techniques and scenario.

This research paper will talk about different ways to understand the techniques that can be formulated and effectively used while doing text analytics using machine learning. It also discusses about different ideas of the different authors about how we can use different algorithms and tune them. After that the different applications and tools are discussed which will helps to create an understanding of different automations that a person can perform while doing specific steps of text analytics.

For understanding the study of text mining and getting insights, the following research questions are proposed for this literature review:

Research Question 1: What are the different techniques of machine learning in the field of text analytics.

Research Question 2: What are the challenges of the different techniques performed by different authors.

Research Question 3: What are the different tools and applications for user generated data in machine learning.

2. Literature Review

Webopedia explains the user generated content as “UGC is the term used to describe any form of content such as video, blogs, discussion form posts, digital images, audio files, and other forms of media that was created by consumers or end-users of an online system or service and is publicly available to others consumers and end-users”. User-generated content is also called consumer generated media (CGM) [3]. So, it is understood that user generated content is mostly referred to the customers that are the main backbone of any business. The basic idea behind gathering the data is to understand people review on a certain product or

getting knowledge of the desires they have. The study of all these comes under the category of human-computer interaction. With the growing customers, the era of Big data has reached, and companies are looking for more advance systems to collect the data from these target systems that are generated in structured, semi-structured and unstructured formats. The data being collected is all about users' preferences, behavior and opinions which is the source for the researchers to collect using affordable manner [4,19]. The content being generated needs to be accessed by machine learning techniques. The mining techniques like clustering and classification can help to build machine learning models [5]. According to Lexalytics.com, Machine learning helps to do predictive and statistical modelling using various techniques[1]. There are many scenarios where the machine learning is applied to the user generated content to eventually helped the society. According to [3], the author has described the role of machine learning and user generated content in convincing their customers for the brand. Many companies need to be very quick in understanding the collection of data and filtering it based on the different criteria. They need to provide their customers with the purchasing decisions, but this could only possible if the technology is effectively used.

Olapic is the visual commerce platform that helps other brand in effectively using the UGC and providing the control over activating many featured with the tool Photosafe which first filters and then rank based on many factors [18]. As per the Adobe blog, they published an article about the learning drive of UGC from machine learning author made a statement "To better understand how machine learning is accelerating content marketing, it's helpful to look at the past: how artificial intelligence "AI" has evolved since the field first emerged in the 1950s, advancing from the automation of narrowly defined tasks to the automation of human-centered tasks, which include content curation, management, and optimization [6]". One of the research papers is about the twitter data where author categorized the tweets generated into positive and negative tweets through semantic analysis which is the one of the methods of text analytics. Semantic analysis is the task that is mostly referred to the natural language processing [7]. Another paper talks about doing research over the online cessation community that helps users to quit smoking. The UGC is used as the input to machine learning model and generated tailored content that can help more specific users [8].

The area covered in this literature includes health sector, social media post, media reporting, use case scenario, data engineering, decision support system, education, email support system. To gather the different research articles, we used the methodology of searching from

different database using keywords. The different digital libraries were references to pick the research article from online databases such as ProQuest, Web of Science, Scopus, IEEE Explorer, Springer and ResearchGate. All the seminal articles including journals and conference paper were obtained from 2004 till 2019. Most of them have been cited in many top ranked articles used for the study of user generated data and machine learning. The appearance of citation of these paper in other scholarly articles has been more than 30-40 times. The primary focus is to look for academic literature published from the above-mentioned database. For the external search or to learn more about the concepts, a study over the published literature in form of book was done that is Data Science (Second Edition) Concepts and Practice 2019 by Vijay Kotu&Bala Deshpande. This book deals with the concepts of text mining and explains the basic understanding of the need of text mining with the help of machine learning. Google search engine was also used as the secondary online search tool to find relevant search regarding the topic using relevant queries with '*'. Following Table I formulates the description of query used as the part of research in obtaining seminal literature for this literature review:

Table I: Query used in obtaining seminal literature

Database	Search Queries
ProQuest	• “Text mining” OR “text analytic” techniques in user generated content • “Text mining” AND “text analytic” techniques • “Text mining techniques user generated content”
Web of Science	• “text mining in user generated data” • “Text mining” OR “text analytic” techniques in user generated content • “Text mining” AND “text analytic” techniques • “Text mining techniques user generated content”
Springer	• Machine learning techniques in text analytics. • User generated content analytics. • “Text mining techniques user generated content”
IEEE Explorer	• Machine learning AND “text analytics” OR “user generated data” • Text mining AND “techniques” OR “tools”

Scopus	• Tools and techniques in text mining. • Text analytics and machine learning.
---------------	---

3. Techniques, Tools and Applications used for Text Mining of User Generated Content(UGC)

This section will talk about the all the results formulated from different papers in three different sections namely techniques, tools and applications.

3.1Techniquesimplemented for Text Mining of UGC:

Under machine learning we have many algorithms that can be used in combination of the data mining techniques. Classification techniques talks about the supervised learning where the data used is already labelled into different classes. Machine learning models can be trained with many algorithms like Naïve Bayes, Logistic Regression, Decision Tree, SVM and many more. The Research conducted for the data of the online smoking cessation society used classification as its basis for the research where they prepared their model based on different algorithms using F1 score metric and AUC metric to identify the accuracy of the labelled data. They were able to get higher accuracy with the Model 5 that was trained with AdaBoost using J48. The domain specific and author specific feature played a great role [9]. The paper about email filtering to identify the spam and useful emails using the semantic modelling along with feature dimensionality reduction technique which provided the better results when tested on different algorithms. They found that using featured dimensionality reduction over the such type data can help to get very specific results. Logistic Regression got the higher accuracy among other algorithms under the classification technique [2]. When it comes to text analytics the content generated by user can be even big beyond expectations and to handle this amount of data manual systems will fail. Looking at big social platforms like twitter generates tons of data which need sophisticated platform to run the different stages of the machine learning. The author of this paper has proposed a unique framework which includes a combination of supervised machine learning and lexicon approach for sentiment classification [10]. The paper talks about how we can handle huge the large amount of data with the help of big data instead of using the old techniques. The idea was developed to use Big data and Hadoop system and they have proposed the tool Apache spark for twitter data analysis. They have tried algorithms under supervised machine learning namely Naïve

Bayes and Decision tree and for lexicon approach namely pattern analyzer. The Decision tree was used under MLib spark library. Based on the result in form of precision, recall, F1-score and accuracy decision tree performed best with 100% everything whereas the Naïve Bayes gave the metrics to 86%. They used sentiment analysis to categorize the tweets into positive, neutral and negative tweets [11]. Now health is becoming an important topic and every platform over the web either its social platform or it is blog of search engine has discussion about the health. The model built by authors of this journal has been focused on monitoring the health of social media streams. They have used text mining processes along with the software as a service model by “MeaningCloud” that helps to extract relevant information from the overall text. They found that using natural language processing dictionary they built, and the rule-based technology has shown a good performance in terms of extraction of significant data. They performed natural language processing with the help of tool cloud service and extracted the keywords and built the dictionary [12].

Also, a quiet impressive approach using semantic analysis over the twitter data using the tweet in form of the big data and did data analysis. The user generated data they took was having text, audio and videos. The existing approach shows that previously they have used the NB, SVM, topic classification and achieved the max accuracy of 83% with SVM whereas the proposed approach included sentiment analysis and they built the predictive model with linear regression. The model was able to achieve 85% of accuracy when 10-fold cross validation was applied. They concluded that linear regressions work better than SVM or NB during predictive modelling [7].

3.2 Tools used for Text Mining of UGC:

Talking about the machine learning in UGC has never been a simple process and it required a lot of different steps involved in during the data preparation stage or data processing stage. Also, sometimes tuning is required to get to specific results and manual computation over huge amount of data are expensive in term of time. In one of the studies about ontology, authors of the paper found that the ontology process is itself a time-consuming task and the manual work requires huge man power. Therefore, the paper talks about different tools and benefits while using specific mining techniques. they have done a small discussion over the tools that are used to learn relationship, learn new concepts and those that helped to build a taxonomy. They were able to conclude that there is no tool exist that can process the learning

fully automatically. Some tools help in linguistic operations and some provide access to areas of lexical semantic knowledge. They have listed many tools in 3 categories namely Clustering, Linguistic patterns & Statistical approach but talking about the top 3 in each category in table 2 [17]. Another research was done which includes the operation of text mining performed by a tool R-Project. This tool used 35 scholarly articles from academic search engine related to business intelligence. They wanted to understand the trend of field BI over the years and over the different regions of the world. The research includes classification techniques and they were able to predict the trend on publication of BI papers in different regions and how the things has evolved over the years 2006-2013 [13].

Talking about the overall process involved in working with enterprise data sets is not a straight forward job when the data you receive is not captured in its best form and you need to do so many things to bring the data, prepare it, transform it and build warehouses. One of the tools that impressed me while writing this literature review is RTextTools which is the complete package of R that involves all the process from data preparation, pre-processing of data to building the classifier and getting the accuracy. This paper talks about the classification technique which uses many algorithms like SVM, Glmnet, SLDA, Random Forest, Maxnet, Boosting, Bagging, Tree & NNet. This tool helps to remove noise data, refine train models and lead to improve accuracy. This is the open source tool that can be used to advance level. This tool was able to calculate the assembled summary of the all the algorithms and presented the precision, recall and F-score. Based on F-score we can calculate that the top accuracy of the algorithms can be used again with respect to the different parameters. The data set that was used here by social scientist was from social media reporting which helps them to build policies over media.[14]

3.3 Applications of User Generated Content:

Our study has derived many practical applications that can provide the public attention towards the benefit it provides to society. This literature review talks about many applications of user generated data in many different sectors like social media, health, media reporting, education, organizational and personal benefits and support systems. This literature review talks about health sector where the applications built using user generated data is about how can we predict the smoking status of a person generating the post to understand the online

community [9] and one of research was on content generated through health portals and connecting to the online drug database to predict the possible health solutions [12]. Also, social media has a major impact on public. Many users spend a lot of time on social media talking about many different products or expressing their views or opinion on currently going issues. Twitter is a platform that is the mostly used platform and analyzing the tweets can let us now insights about the data over the time. One of the papers talks about gathering the twitter data and using machine learning predicting the tweets as positive, negative or neutral and over the different regions of the world. They have included the apache spark to handle a lot of twitter data for fast processing [11]. Another paper also talks about the twitter data through gathering different data about the time of elections from twitter and did some analysis to understand how many tweets were in favor of Hilary and Trump. The analysis of twitter data can help to understand many political scenarios and situations to predict something for the parties over the time frame [7].

One of the literature reviews is a small review on what are the different type of problem we face with noisy data that is the first stage of data preparation. They talk about issue with speech recognition and solving them with OCR technology. Also, they have provided many different solutions to help the researchers if they are going through similar problems [15]. Also talking about the decision support systems there is one paper that is critical towards solving the confusion people have while purchasing the laptop. The paper using artificial intelligence with machine learning and embeds the image mining in between to help people make wise decisions during the purchase. This could be very useful to society in terms of multiple options and making hard choices [16]. Another sector that can help big organizations to filter out the spam emails from important emails. This type of research was done using user generated data in form of emails and then understand the text through semantic methods. This could also keep away the potential attacks over the personal data [2].

Following Table II shows comparison of various text mining techniques:

Table II: Comparison Of Various Text Mining Techniques

Research Paper	Concepts				
	Source of data	Techniques	Algorithm Used	Metric Used	Tool Used
Mining user-generated content in an online smoking cessation community	Web	Classification	SVM, Decision Tree, LR, Tree -J48,	F-score, AUC	X

to identify smoking status: A machine learning approach			Naïve Bayes, AdaBoost		
Text Mining Business Intelligence: a small sample of what words can say	Research Paper/ Web	Classification	X	X	YES
Efficient email classification approach based on semantic methods	Email Data	Classification, Semantic Discovery, Feature dimensionality reduction	Naïve Bayes, RBF, SVM, LR, Random Forest, J-48	Accurac y, precisio n, Recall, F-score, Kappa	X
Text Analytics Framework using Apache Spark and Combination of Lexical and Machine Learning Techniques	Web/Twitter	Supervised ML, Lexicon approach, Hybrid	Naïve Bayes, Decision Tree, Pattern Analysis	F-score	YES
Turning user generated health- related content into actionable knowledge through text analytics services	CIMA, MEDRA, SpanishDrugEffect DB database	Entities Recognition, NLP, Relation Extraction, Syntactical Parsing	X	F-score, Precisio n, Recall	YES
Sentiment analysis in twitter data using data analytic techniques for predictive modelling	Web/Twitter	Sentiment Analysis, Classification, Predictive Modelling	Logical regression	Accurac y	X
RTextTools: A Supervised Learning Package for Text Classification	Web	Classification	SVM, GLMNET, SLDA, BOOSTING	Precisio n, Recall, F-Score	YES

4. Research Gap

This literature review has identified the following gaps in the respective study of user generated data and could be taken as the input to the future research.

- There are only few articles that doing a comparison of different techniques of mining. Most of the studies are related to specific topic/ technique such as semantic analysis done on twitter data or classifying email content. The basis of methodology used in

literature review needs a concept to be validated over different platforms to be excepted globally.

- ii. Most of the articles are based on single type of technique like classification is mostly used in the all the literature along but it should not be the case. As per the search criteria it revealed only those literature and there is no literature based on Association rule, clustering and Outlier detection. More than one technique should be used over the corpus to identify the validity of the model. Also, if these techniques like association and clustering are very good techniques and need to be analyzed further.
- iii. Some of the literature review are not using the tools to perform the data preparation and machine learning model stage. They are doing every task manually. It is very important for the research to be time effective and accurate. Doing manually can cause errors and can affect the results. Tools should be the used to improve the result plus large datasets should be handled with tools to perform fast executions.
- iv. Most of the literature reviews are based on calculation of f-score as a metric. Sometimes relying on single parameter can affect your research results based on algorithms. The accuracy of the model depends upon the different metrics and true positive rate. The mining technique need to include comparative study of different metrics like precision, recall, accuracy and f score. These metric changes can be taken into consideration for future study.

5. Conclusion

In this literature survey, how text mining is helpful and by which different methods data can be categorized using machine learning techniques. Different text mining techniques are reviewed using User Generated Content. Also, different tools and various applications of text mining are analyzed. Also, there are key finding from this survey which are mentioned below:

- i. Text mining literature synthesized in this literature review contains both the data from web and from other standalone databases. Also, most of the user generated data is coming from social media as source.
- ii. The approach towards the using techniques is becoming hybrid somehow, there are some articles that include combination of the few different techniques to predict better result.
- iii. Most of the research papers talks about using the tools for the better results. The results obtained through tool are impressive and hold a stronger ground

for prediction as they are accurate. One of the examples include using Apache spark for larger set of data which couldn't be possible by normal machine.

- iv. It is very important to handle data pre-processing properly since cleaning of data for text analytics is the important phase and it would be better to use tools for such process.
- v. Text analytics involves computation and therefore using multiple metrics can help us to derive better results as in case of Email classification.

References

- [1] S. Redmore, "Machine Learning for Natural Language Processing", [Online], Available:<https://www.lexalytics.com/lexablog/machine-learning-vs-natural-language-processing-part-1>.
- [2] Bahgat, M.Eman, S.Rady, W.Gad, and I.F. Moawad, "Efficient email classification approach based on semantic methods", *Ain Shams Engineering Journal*, vol. 9, no. 4 pp. 3259-3269, Dec 2018.
- [3] V.Beal, "UGC— user generated content", [Online], Available: <http://www.webopedia.com/ TERM/U/ UGC.html>.
- [4] Khan, S.Ullah, S.A.K. Bangash, and K.U.Khan, "Learning analytics in the era of big data: A systematic literature review protocol", *International Symposium on Wireless Systems and Networks (ISWSN)*, pp. 1-7. IEEE, Jan2018.
- [5] Kotu, Vijay, and B. Deshpande, "Data Science: Concepts and Practice", *Morgan Kaufmann*, Nov 2018.
- [6] S. Pombo, "The Future of Content Marketing: How Machine Learning and UGC Drive Scale", June 2018.

- [7] Sulthana, A. Razia, A. K. Jaithunbi, and L. Sai Ramesh, "Sentiment analysis in twitter data using data analytic techniques for predictive modelling.", *Journal of Physics: Conference Series*, vol. 1000,no. 1, IOP Publishing, Apr 2018.
- [8] X.Wang,K.Zhao, S.Cha, M.S. Amato, A.M.Cohn, J.L. Pearson, G.D. Papandonatos, and A.L. Graham, "Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach.", *Decision Support Systems*, vol. 116 , pp. 26-34, Jan 2019.
- [9] F.Xia, W.Wang, T.M.Bekele andH.Liu, "Big scholarly data: A survey.", *IEEE Transactions on Big Data*, vol 3, no. 1,pp. 18-35, Jan 2017.
- [10]Ghazal, A.Mohamed, O.Ibrahim, and A.S.Mostafa, "Educational Process Mining: A Systematic Literature Review.", *European Conference on Electrical Engineering and Computer Science (EECS)*. IEEE, pp. 198-203, Nov 2017.
- [11] A.P.Jain, and P.Dandannavar, "Text Analytics Framework using Apache Spark and Combination of Lexical and Machine Learning Techniques.", *Journal of Applied Information Science*,vol. 4, no. 1, 2016.
- [12] Martínez, Paloma, M.L.José, I.S.Bedmar, J.M.Schneider, A.Luna, and R.Revert. "Turning user generated health-related content into actionable knowledge through text analytics services.", *Computers in Industry*, vol. 78 pp. 43-56, May 2016.
- [13]Ishikiriya, C.Satiko, D.Miro, and C.F.S.Gomes, "Text Mining Business Intelligence: a small sample of what words can say",*Procedia Computer Science*, vol. 55, pp. 261-267, Jan 2015.
- [14]Jurka, P.Timothy, L.Collingwood, A.E. Boydston, E.Grossmanand W.Atteveldt,, "RTextTools: Automatic text classification via supervised learning",*R package version*, vol. 1, no. 9, 2013.

- [15]Knoblock, Craig, D.Lopresti, S.Royand L.V.Subramaniam, “Special issue on noisy text analytics”, *International Journal on Document Analysis and Recognition*, vol.10, no. 3, pp. 127-128, Dec 2007.
- [16] W.Y.Tayand C.K.Pang, “A Decision Support System for Laptop Procurement Using Mixed Text and Image Analytics”, *Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems*, vol 2, pp. 279-292. Springer, Cham, 2015.
- [17] G.Pérez, Asunción, and D.M.Macho,“An overview of methods and tools for ontology learning from texts”,*The knowledge engineering review*, vol. 19, no. 3,pp. 187-212, 2004.
- [18] A.Zackiewicz, “Think Tank: The Role of Machine Learning and User-Generated Content”,[Online],Available: “<https://wwd.com/business-news/business-features/olapic-think-tank-1202760069/>”, July 2018.
- [19]G.W.Kim, Y.M.Kim, J.S.Han, Y.L.Rhie andM.H.Yun, “Understanding characteristics of user-generated content as a source of extracting user value”, *International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 1751-1755, Dec 2016.