A Comparative Study of Various Text Mining Approaches For Analysis of Drug Reaction Using Social Media Post

Smruti J. Dave[1], Prof. Hardik H. Maheta[2]
M.E. Final Year Student [1,] Assistant Professor[2], Department of Information Technology [1,2] , Shantilal Shah Engineering College, Bhavnagar,India.[1,2]
Email:smrutidave15@gmail.com[1] ,hardikmaheta@gmail.com[2]

Abstract- Adverse drug reactions can be referred as unwanted, uncomfortable or a harmful effect that any drug may cause. Adverse Drug Reactions are major problems as they may cause serious health related issues. So, identification of adverse effects of any drug is very important. Adverse effects of any drug are available from pharmaceutical companies when drug is launched in market. Still all adverse reactions cannot be identified during these clinical trials as it is applied on limited volunteered people. Some regulatory agencies monitor adverse reactions through surveys of medical practitioners. Currently large numbers of patients share their experiences on social media. These experiences of patients can be very helpful to detect the adverse effects of various drugs. We have done comparative study of various text mining approaches on social media posts to understand and obtain adverse drug reaction. Here, our main aim is to come up with a comparison of various text mining methodologies that can enable us to detect unreported adverse drug reactions using social media posts.

Index Terms - Adverse Drug Reaction, pharmacovigilance, Sentiment Analysis, LDA, Post tagger

1. INTRODUCTION

In the United States, at least 100,000 deaths are estimated due to adverse drug reactions (ADRs) every year in US hospitals [2]. So, the information of serious adverse drug effects is important factor in public health concerns. An Adverse Drug Reaction can be referred as any unexpected outcome to a medicinal product [1, 2]. An adverse drug event (ADE or AE) can also be called as any unfavorable and unintended sign, symptom, or disease associated with the use of a medicinal product [2]. Sentiment analysis is the study of users' opinions, reviews, attitude to topics, products, services, organizations, individuals and events or their attributes [28]. In recent years sentiment analysis has attracted the attention of the researchers in the field of data mining and machine learning [28]. So to understand the adverse effects of various drugs faced by the patients is taken in to consideration for this research.

There are large numbers of data sources available for finding and monitoring of the Adverse Drug Reactions [2]. These data sources contain spontaneous reported data, e-health record, pharmaceutical databases, bio medical literature etc [39], [40]. But all these data sources are limited by the issue of high cost, privacy and under reporting ratio. On the other side drug reaction data reported by users on the social media are free to access and thus can be used for pharmacovigilance. Large volume of patients' reviews on social media is increasing all over the world in recent years. So, researchers are now much interested in utilizing social media data for detection of Adverse Drug Reactions.

For example in one survey, they had found that twitter is growing by 1, 35,000 of users every day and generates around 9100 of tweets per second [3]. More over there are also some health related social networks like Daily strength and Medhelp are available on which the patient share their experiences [5]. In Health related data sets, users post the prescriptions of drugs, side effects and treatments. They also share their views, problems and results which makes social media an important source of data. There may be a situation where different patients suffering from a common problem. In such cases if they share their reviews and experiences on social media that may be helpful for pharmacovigilance. It can provide a clear view to both the health researchers and the patients.

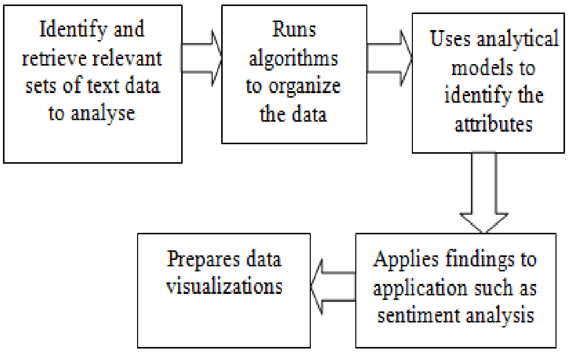
Sometimes the regular data sources can miss the new or rare events due to any drug that social media can provide and also provide an early access to the detected Adverse Drug Reaction which is a major benefit for the health related and pharmaceutical industries [2]. According to the regulatory authorities, social media is a way to obtain additional information from the general public [2][3][7].

The paper is organized as follows. Section II describes the importance of text mining process. Section III describes the description of different text mining approaches used in Adverse Drug Reaction detection. Section IV describes the comparison study of text mining approaches. Finally we conclude this paper with Section V.

2. TEXT MINING

Text mining is a major part of a field known as information mining that discovers intriguing examples from huge databases [11]. The main objective of text mining is to find and provide some information that is not yet known or recorded. Text mining is also known as text analytics [11]. It is the process of deriving high quality information from the text. It analyses and explores the large amount of unstructured text data aided by the software which can identify the patterns, concept or keywords from the unstructured data. Text mining has become more practical for data science researchers because of the development of deep learning and machine learning algorithms to analyze massive datasets [7]. Text mining is similar to the data mining in a way that instead of focusing on structured data, it focuses on the informal text. First step for the process of text mining is organize and structure the data in order to achieve quantitative and qualitative analysis. For doing the above step Natural Language Programming (NLP) technology can be used. Text mining approach is described in figure 1.

Fig. 1. Text mining process flow


Identify and retrieve relevant sets of text data to analyse
Runs algorithms to organize the data
Uses analytical models to identify the attributes
Applies findings to application such as sentiment analysis
Prepares data visualizations

Sentiment analysis is also known as opinion mining [17]. It is one of the widely used applications of text mining process to track the user sentiment from the online reviews or from social networks [8]. Pharmacovigilance from social media posts has been recent research topic .It has gone through remarkable progress over the recent years [7]. The past studies focused on combining lexicon analysis for the extraction of Adverse Drug Reaction to provide drug safety. The lexicon-based approaches has a number of limitations while applying to social media data. Because the users use phrases, descriptive symptom explanations, and idiomatic expressions, which are not available in existing lexicons [7]. Instead of using lexicon-based approaches, recent studies focus on the text mining. For extracting Adverse Drug Reactions it is required to extract both drug names and the adverse effect caused by the drug. Topic modeling is an approach which discovers hidden semantic structures in a corpus [10]. It assumes that the text document is the mixture of the words and the words are considered as topics. This technique finds hidden or latent topics from the text documents. Various approaches of text mining to detect Adverse Drug Reactions, and their usefulness are discussed in the next section.

3. VARIOUS APPROACHES USED IN LITERATURE

3.1 Information Retrieval

Information retrieval is the task of retrieving the related and useful information from the databases [11]. It is mainly concerned with providing information access to the user to large amount of unstructured data [11]. There are bunch of documents from which one want to retrieve the information. These documents need to be indexed in order to come up with a result very quickly. An index is a compressed version of the same information which these documents contain. When a user query comes in, index is queried and the documents are obtained which match the particular query. Further there is the ranking module, which tries to rank these reviews that are retrieved. For example, some blogs state the inconvenience faced by the users. These blogs must be at the top of a particular website. So they must be ranked accordingly. It focuses on the task of facilitating information instead of focusing on the analysis and pattern recognition in the text document; which is the main task of text mining [11], [12], [13]. Information retrieval gives less priority to the processing and transformation of the text. While in text mining, one should include information access for further enable users to understand the information and help in decision making [11].

3.2 Natural Language Processing

Natural language processing can be referred as the ability of a computer program to understand the human language as we speak it [16], [17]. Natural language processing includes two tasks. First is to understand the human language and second is to generate the sentence in machine understandable code [16]. So the two components of natural Language Processing are:

- Natural Language Understanding
- Natural Language Generation

Natural Language Understanding includes three types of ambiguity (1) Lexical ambiguity that is word level (2) Syntactical ambiguity (3) Referential ambiguity. Natural Language Generation includes Text planning, sentence planning and Text realization. Text planning is performed using the knowledge base. Sentence planning refers to make a sentence by arranging the words in a meaningful way.

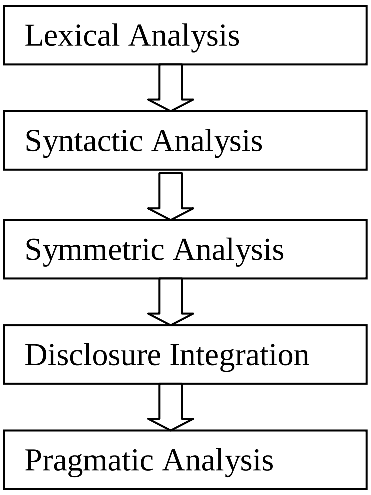


Fig 2: Natural language processing steps

It is a challenging task to develop an NLP application because computers need the input in form of a programming language which is structured and unambiguous. Human speech is not always precise. It is sometimes ambiguous and depends on many complex variables such as slang, regional dialects. There is a large amount of information stored in a free text files, for example patients' medical reports [17]. Natural language processing allows this kind of information to be accessible to computers and to find relevant information in these files. Sentiment Analysis is also one of the applications of Natural Language processing [18], [19]. Using sentiment analysis the polarity of the sentence on social media can be identified [18], [19].

3.3 Opinion Mining and Sentiment Analysis

Sentiment analysis is also known as opinion mining. It refers to the task of natural language processing that can identify the emotional tone of the text.

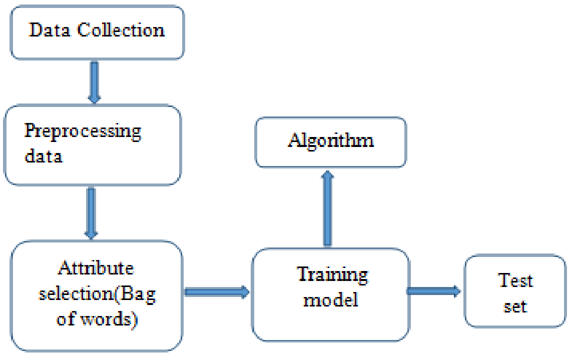


Fig 3: Sentiment Analysis steps

In sentiment analysis, the unstructured text is gathered from social media posts [22], [24]. For identifying the sentiment, rule-based automatic or hybrid methods can be used [25]. Rule-based methods perform sentiment analysis based on the pre-defined, lexicon-based rules. Automatic systems use machine learning techniques while Hybrid system combines both these approaches [25]. Sentiment analysis can extract the polarity or sentiment of the sentence in text.

3.4. Probabilistic Methods for Text Mining

There are various probabilistic methods for text mining such as probabilistic Latent semantic analysis and Latent Dirichlet Allocation (LDA) are unsupervised topic models .Where as conditional random fields is a supervised topic model [11].

Probabilistic models are the models which are based on the assumption that documents are mixture of topics and a topic is probability distribution of words. A topic model specifies a probabilistic procedure based on probabilistic rules, which describe how the words might be generated in the document based on random variables.

3.4.1 Latent Semantic Indexing

Latent Semantic Indexing also referred to as Latent Semantic Analysis [26], [27], [28]. It is used to improve the accuracy of information retrieval [35], [36]. It uses a method called singular value decomposition to scan the unstructured text documents and to identify the relation between the concepts of the text documents [26], [35]. The limitation of this model was that it is useful for probabilistic modeling of text but it is incomplete at the level of document [28]. In Latent Semantic Indexing model numbers of parameters are increased linearly along with the size of corpus which causes the problem of over fitting [28]. Latent Semantic Indexing is not able to assign probability to a document outside the training set [28]. No assumptions are made for how the weights can be calculated. So it is difficult to generalize this model for new documents [28].

3.4.2 Latent Dirichlet Allocation

It is easy for a human to understand a language as compared to the machine. For a machine, it is difficult task to understand human language. One solution of this is to group certain words in predetermined category and then merging useful words from stop words and find relation between two words in a sentence [37]. Latent Dirichlet Allocation is one technique to assist in modeling the data consisting of a large corpus of words [37]. Latent Dirichlet Allocation is an extension of Latent Semantic Indexing model [28], [29]. It overcomes the issue of Latent Semantic Indexing by treating the weights as hidden random variables [28], [29]. Latent Dirichlet Allocation is one of the most popular models of probabilistic topic models where its procedure connects the parameters of documents via hierarchical model [31]. It relates words and documents through the latent topics [31].

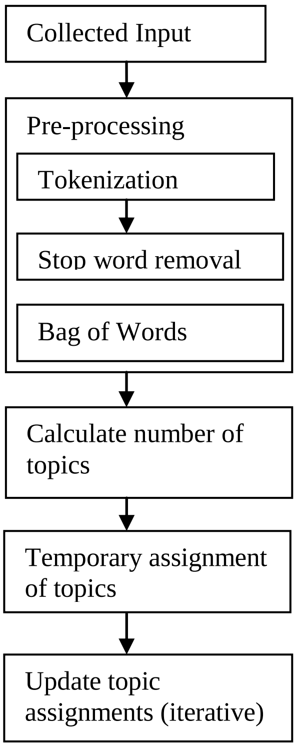

Collected Input
Pre-processing
Tokenization
Stop word removal
Bag of Words
Calculate number of topics
Temporary assignment of topics
Update topic assignments (iterative)

Fig 4: Latent Dirichlet Allocation

4. COMPARISION OF DIFFERENT TEXT MINING APPROACHES

Method name	**Usefulness**	**Merit**	**Limitation**
Information Retrieval	Information retrieval can crawl the user reviews related to drug and then adverse drug reaction can be extracted using information extraction.	Users are able to extract the information quickly because of the use of indexer. The ranking module in this system ranks the retrieved reviews.	Focus on the task of providing information instead of analysis and pattern recognition in text
Natural Language processing	It can be used in data cleaning process for punctuation removal from unstructured text and further for finding drug symptom dependencies.	Provide end-to-end training	This method works on human language, which is not always precise.
Sentiment Analysis	Sentiment Analysis extracts the polarity of the text from social media posts. Thus it shows that either particular drug has a positive effect or negative.	It is important as it considers the opinions , views and emotions of patients	The data collected from social media can be noisy sometimes
Latent Semantic Indexing	It identifies the relation between concepts of the text by scanning unstructured text documents. In case of finding adverse drug reaction this method identifies the relation between the entities such as drug-drug or drug-symptom.	For probabilistic modeling of text, there are some practical and scalable implementations already available.	This method does not provide accuracy because sometimes the order of words is ignored.
Latent Dirichlet Allocation	It can be used to discover the topics. These topics group drugs with similar safety concerns together.	It overcomes the issue of Latent Semantic Indexing.	It is an unsupervised model, but in some cases such as sentiment analysis; weak supervision is needed.

5. CONCLUSION

In this paper, we provide a comparative study of various text mining approaches to find Adverse Drug Reaction from social media posts. We have observed that unnoticed adverse reaction of drug can be found through patients’ reviews on social media. We have compared four different techniques. Among them, probabilistic topic model such as Latent Dirichlet Allocation is a technique which finds drug name and its related symptoms. Further Latent Dirichlet Allocation generates the drug-symptom pair. Other techniques like information retrieval or Natural Language Processing can extract the sentiment from the post but Latent Dirichlet Allocation can give the relation of the sentiment with its adverse effect.

REFERENCES

[1] https://www.who.int/medicines/areas/quality_safety/safety_efficacy/trainingcourses/definitions.pdf

[2] Chen, Xiaoyi, et al. "Mining patients' narratives in social media for pharmacovigilance: adverse effects and misuse of methylphenidate." *Frontiers in pharmacology* 9 (2018): 541.

[3] O’Connor, Karen, et al. "Pharmacovigilance on twitter? Mining tweets for adverse drug reactions." *AMIA annual symposium proceedings*. Vol. 2014. American Medical Informatics Association, 2014.

[4] Korkontzelos, Ioannis, et al. "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts." *Journal of biomedical informatics* 62 (2016): 148-158.

[5] Sarker, Abeed, and Graciela Gonzalez. "Portable automatic text classification for adverse drug reaction detection via multi-corpus training." *Journal of biomedical informatics* 53 (2015): 196-207.

[6] Eguale, Tewodros, et al. "Association of off-label drug use and adverse drug events in an adult

population." *JAMA internal medicine* 176.1 (2016): 55-63.

[7] Rajapaksha, Parinda, and Ruvan Weerasinghe. "Identifying adverse drug reactions by analyzing Twitter messages." *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2015.

[8] Korkontzelos, Ioannis, et al. "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts." *Journal of biomedical informatics* 62 (2016): 148-158.

[9] Abacha, Asma Ben, et al. "Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification." *Journal of biomedical informatics* 58 (2015): 122-132.

[10] Karimi, Sarvnaz, et al. "Cadec: A corpus of adverse drug event annotations." *Journal of biomedical informatics* 55 (2015): 73-81.

[11] Allahyari, Mehdi, et al. "A brief survey of text mining: Classification, clustering and extraction techniques." *arXiv preprint arXiv:1707.02919* (2017)

[12] Faloutsos, Christos, and Douglas W. Oard. *A survey of information retrieval and filtering methods*. 1998.

[13] Mogotsi, I. C. "Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval." (2010): 192-195.

[14] Rajman, Martin, and Romaric Besançon. "Text mining: natural language techniques and text mining applications." *Data mining and reverse engineering*. Springer, Boston, MA, 1998. 50-64.

[15] Manning, Christopher, et al. "The Stanford CoreNLP natural language processing toolkit." *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014.

[16] Manning, Christopher D., Christopher D. Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[17] https://searchbusinessanalytics.techtarget.com/definition/natural-language-processing-NLP

[18] Jingfang, L. I. U., et al. "Adverse Drug Reaction Related Post Detecting Using Sentiment Feature." *Iranian journal of public health* 47.6 (2018): 861.

[19] Korkontzelos, Ioannis, et al. "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts." *Journal of biomedical informatics* 62 (2016): 148-158.

[20] Steyvers, Mark, and Tom Griffiths. "Probabilistic topic models." *Handbook of latent semantic analysis* 427.7 (2007): 424-440.

[21] Xiao, Cao, et al. "Adverse drug reaction prediction with symbolic latent Dirichlet allocation." *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[22] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and Trends® in Information Retrieval*2.1–2 (2008): 1-135.

[23] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5.1 (2012): 1-167.

[24] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." *LREc*. Vol. 10. No. 2010. 2010.

[25] Liu, Bing, and Lei Zhang. "A survey of opinion mining and sentiment analysis." *Mining text data*. Springer, Boston, MA, 2012. 415-463.

[26] Hofmann, Thomas. "Probabilistic latent semantic indexing." *ACM SIGIR Forum*. Vol. 51. No. 2. ACM, 2017.

[27] Deerwester, Scott, et al. "Indexing by latent semantic analysis." *Journal of the American society for information science* 41.6 (1990): 391-407.

[28] Bagheri, Ayoub, Mohamad Saraee, and Franciska De Jong. "ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences." *Journal of Information Science* 40.5 (2014): 621-636.

[29] Steyvers, Mark, et al. "Probabilistic author-topic models for information discovery." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004.

[30] https://www.tidytextmining.com/topicmodeling.html

[31] AlSumait, Loulwah, Daniel Barbará, and Carlotta Domeniconi. "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking." *2008 eighth IEEE international conference on data mining*. IEEE, 2008.

[32] Cambria, Erik, et al. "Statistical approaches to concept-level sentiment analysis." *IEEE Intelligent Systems* 28.3 (2013): 6-9.

[33] Cambria, Erik. "Affective computing and sentiment analysis." *IEEE Intelligent Systems* 31.2 (2016): 102-107.

[34] Anick, Peter G., and Shivakumar Vaithyanathan. "Exploiting clustering and phrases for context-based information retrieval." *ACM SIGIR Forum*. Vol. 31. No. SI. ACM, 1997.

[35] https://www.searchenginejournal.com/latent-semantic-indexing-wont-help-seo/240705/

[36] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." *Discourse processes*25.2-3 (1998): 259-284.

[37] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research*3.Jan (2003): 993-1022.

[38] Porteous, Ian, et al. "Fast collapsed gibbs sampling for latent dirichlet allocation." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.

[39] https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018/home

[40] https://dailymed.nlm.nih.gov/dailymed/

[41] Shang, Ning, et al. "Identifying plausible adverse drug reactions using knowledge extracted from the literature." *Journal of biomedical informatics* 52 (2014): 293-310.

[42] Bisgin, Halil, et al. "Mining FDA drug labels using an unsupervised learning technique-topic modeling." *BMC bioinformatics*. Vol. 12. No. 10. BioMed Central, 2011.

[43] Biarez, Odile, et al. "Comparison and evaluation of nine bibliographic databases concerning adverse drug reactions." (1991): 1062-1065.

[44] Yang, Ming, Melody Kiang, and Wei Shang. "Filtering big data from social media–Building an early warning system for adverse drug reactions." *Journal of biomedical informatics* 54 (2015): 230-240.