Automatic Drug Reaction Detection Using Sentimental Analysis

Sahana D S
PG Student
Dept of CSE, CIT Gubbi Tumkur.

Prof. Shantala C.P
Vice Principal and HOD,
Dept of CSE, CIT Gubbi Tumkur.

Girish L
Assistant Professor
Dept of CSE, CIT Gubbi Tumkur

***Abstract*—Nowadays various Pharmaceuticals industries are using the Social media in which industries may include sick persons, Doctors, Pharmaceuticals companies etc. There exists a lot of medical sites and conference blogs which are the part of sentimental analysis. These are used in many applications such as text based information (i.e., social media such as twitter, Facebook), market intelligence, drug observations ,taking the opinions about heath related information and also to detect the impact of Adverse Drug Reaction(ADR) automatically by making use of Pharmacovigilance techniques. By using the already existing tools which are used to examine the health related information and sentimental analysis we are not able to get the sufficient classification accuracies. And also they are not able to extract the exploratory and transferable features; hence they are lacking the biased features. In our design we use the Advanced NLP approaches to generate the functional features from the input and using the advanced machine learning algorithms to achieve the classification accuracies. Here we rely on text classification approach that generates huge set of features which represents many semantic properties like sentiment, polarity, topics which are useful to manifest the experience of the user when they talk about ADR. We present two datasets that are prepared by user to perform the job of ADR detection from the data posted by the user on internet. We also verify whether the combining training data which have been taken from different collections can improve the automatic classification accuracies and address the issue of data imbalance. This is very useful where there exists a large datasets and also reduces time and cost of those data.**

***Keywords* –Pharmacovigilance, NLP, Sentiment analysis, ADR, polarity**

I. INTRODUCTION

Since the last few years many of the Pharmacovigilance industries, medical institutions and organizations are increasingly using the web technologies [1] in their day to day operations. Web technologies like forums, blogs, social networks (like twitter, face book, whats-app, wiki) in order to give and get the feedback of certain medicines [2][3]. Social media analysis is becoming a strong tool which helps to mine the content across various domains on these kinds of social networks. In the recent years, one domain which has grown by massive proportions is Social media. Users will be focusing on many issues while they can exchange their views such as side effects caused by that particular medicines ,similar medicines and their effects , about the treatments. All will make those particular sites as unique and Robust [4] [5] [6]. One of the example of such sites is online health Community. Since the user posted data are publicly available, it is also helpful for the pharmaceutical companies to address the issues faster .They can also make the comparisons with their competitors and also they can post their marketing drug observation. By using these kinds of sites we can predict the ADR. There are many researches have been considering on dividing positive and negative comments in online medical sites , in order to study the user inclination about various drugs. This is very beneficial for the customers who wish to know about the sentiment of the product before purchasing that product and also for the companies to know about the sentiment of the customers of their products.

Due to the availability of large amount of user posted data on social media [7], ADR detection accuracies is lacking. There exists various NLP challenges such as misspellings, irrelevant words, doubtful and sarcastic expressions [8] to express adverse reactions. These kinds of issues fails to identify and generalize different posts, which affects the performance. There also exists many other problems like sparse feature vectors [9], Shallow processing is used since posts are very short and limited features are needed, colloquial style of expression and ambiguous with respect to sentiment, polarities and topics in social networks. In this paper we are considering the datasets of clinical and twitter data that consists of annotated posts, which are publicly available data to remove the problem of Sparsity [10] and to represent the information rich features. We also address the issues of automatic

ISSN: 2278 – 1323

detection of ADR text data coming from different sources such as twitter [11] and clinical data on which we are focusing on user posted data.

Currently , only some research attempts have been undertaken to apply the advanced NLP techniques to generate the information rich features like NLP-based Sentiment analysis techniques[12][13]. They are not considering the data from multiple social networks to improve the classification accuracies. Another main thing is lacking the availability of publicly available data in order to perform the research and also to compare with other data. Hence, we got a strong motivation to use the NLP approaches to achieve the better classification accuracies. So here we consider the data which is publicly available .We combine the data from multiple sources such as twitter and clinical data sources.

II. RELATED WORK

There exists a representational richness framework (FRRF) [11] on health related data and social media data that is based on sentimental analysis. This framework accepts the semantics, sentiments, topics, polarities and many more, in which people are able to exchange their experiences with others. But here there exist a large number of features which results in noise and redundancy. For that we need to apply to feature sub assumption phase to remove redundancy. For example: A study have been done and analyzed on the online health forums which analyses the emotional impact [14]. Here we used the machine learning and also text analytics, to evaluate the blogs, posts and to identify the sentimental change impact when they receive support change from other forums. To detect the sentimental polarities [15], there are certain tools such as Standalone and Trained workbench tools [16]. There also exists many comparison tools such as Senti-Strength [17], Sentiment140 [19], Opinion Finder [20], FSH [21] [22] and a word n-gram baseline [23]. Some of the challenges exists in online medical analysis like users in the forums and blogs uses the casual language while commenting on the medical sites, which are short. This leads to some of the issues on representation of text based analytics and also feature sparsity. This will be illustrated in FRRF [24] [25]. In this approach we need to use more generalized lexicon-based approaches [26].

And earlier, drug safety observations were dependent on the Spontaneous Reporting Systems (SRS). SRS are the passive observation systems that are mainly used to address the adverse drug reactions and are collected from the stake holders in the medical industries. The process of SRS will be carried by making use of the following three steps: Data Acquisition, Data Assessment and Data Interpretation. However, research tells that Spontaneous Reporting Systems such as Adverse Event reporting System (AERS), suffers from a wide range of limitations .Hence to overcome the limitations of SRS we use Advanced NLP approaches and Machine learning techniques to achieve classification accuracies and to generate information rich features. But the existing systems has several limitations like results in Under Reporting , results in Over reporting, Data obtained will be incomplete, Duplicate copy of data may generate, Spontaneous Reporting systems may vary because of severity of the reaction for that particular user and also may vary in time from the market introduction

There are some of the researches have been undertook previously on the sentimental analysis also like pang et al [27], which analyses the performance of different intellectuals on the movie reviews. Practically, it is very difficult to collect large data to train sentiment distributions for tweets. But it has a limitation that it doesn’t include the neutral class. Neutral class like, PT@friz motor company is not up to the expectation on mileage and speed http://gs.cli/9adbsCd#motor.Our goal is to traverse NLP approaches, which are used to extract the exploratory and transferable features from the text coming from clinical and social media. We try to solve the problem of data imbalance by using automatic ADR detection from social networks and by using machine learning we can improve the performance. We provide the data which is available publicly.

III. PROPOSED SYSTEM

NLP techniques have been used to discover ADR’s from the text by considering data from Electronic Health Records (EHR) [29-30], twitter data and clinical data. EHR [31] contains the complete records of the patients, about their treatments, previous history of the patients and also experiences ADR’s. It also observed that combining data from multiple sources say clinical and twitter helps to improve the ADR detection accuracies

Usually Social networks will be having a large volume of data which are posted by user and are used as tools for the Real Time Knowledge Discovery [32]. Due to the presence of the informal terms, ambiguous and other factors, the task of data mining is becoming difficult.The previous work to

ISSN: 2278 – 1323

use the user comments was considered by Lea man et al[33], showed that user comments from online Health communities can think about known ADR's and generates early warnings about the ADR's which are unknown.

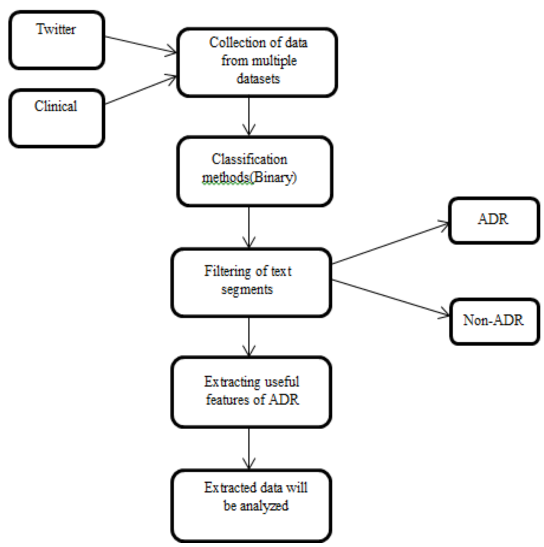

Fig1: ADR Pipeline

There exists lot of lexicon based approaches to detect the ADRs, but has a limitation on number of drugs and number of target ADR's. Here, we use the Lexicon –based approaches [34] for noticing the syntactic and semantic patterns from the user posted data and machine learning used to apply for ADR [35-37] relation extraction and to classify the drugs in to different groups [38].

Fig1 shows the architecture of the pipeline that we used to detect the ADRs from the Social media. The first step in our architecture is the collection of data from multiple data sets, we are considering the twitter and clinical data reports. Twitter incorporated data will be collected based on common and brand names of medicines along with spoken misspellings. Second dataset is the clinical data sets taken from online health community.It consists of the review pages of each of the medicines and allows each patients to exchange their personal knowledge about the treatments, diseases and other factors. The main advantage in Online Health community is it doesn't requires filtering , since each of the medicines will be having a specific review page in the daily strength along with that we also consider collection of ADE ,which is available publicly . It contains double-annotated sentences which indicates whether ADR's are present or not. But this datasets is not considered from social media, even though we consider because to perform the comparisons of the performances and to check how classification algorithms vary between datasets of different classes.

IV. IMPLEMENTATION

Our approach follow some steps to perform the analysis of the datasets, they are

(a) Splitting the reviews
(b) Feature extraction
(c) Identifying subjectivity using machine learning algorithm
(d) Sentiment identification
(e) Polarity estimation
(f) Analysis

Data sets: In our approach we use two datasets .One is the clinical dataset and the other one is the social media data [Twitter data]. Each of the datasets will contain the text segments that are the collection of sentences , group of sentences , blogs, websites and many more , which are publicly available to detect the presence or absence of ADR's.

- **Clinical dataset:** The clinical data that we are considering is the historical data, which are taking from publically available data that are taken from health reports. This data will be considered, in which it contains many attributes like Product, Substance, ADR, Age group, Gender, Clinical data and many others. But this dataset is not obtained from Social media. Even though it is not taken from social media, we are using this data to compare the performances with other datasets in order to verify how the classification algorithm differs from different datasets and how these data are used by training learners.Fig2 illustrates how the clinical data has been analyzed.

- **Twitter:** We are considering the data set from twitter, which is a popular site that is growing very rapidly in today's world. Initially we start up with a data collection process that includes identification of a set of drugs , followed by a collection of users comments along with drug name .To classify the tweets , we mainly focus on the drugs for the diseases and conditions of the patients . Tweets will be considered from the data using their brand names of each of the drugs because twitter may also contain variety of topics, which are irrelevant.Fig3 illustrates how sentiment has been identified

ISSN: 2278 – 1323

as positive, negative or neutral tweets and estimation of polarity have been done.

Here we are collecting our own data because there does not exists large publicly available datasets of messages in twitter along with sentiment. So to access the tweets we use an Interface called Application Programming Interface, which will be having a parameter which represents in which language we can retrieve the tweets. It also returns the positive and negative emotions. Training data will be pre-processed like emoticons will removed, the tweets with both positive and negative tweets will be removed, and repeated tweets will be removed.

(a) *Splitting the reviews:*

This will be performed for the twitter data. The reviews will be divided in to many sentences and make those particular sentences as a bunch of sentences. Then these bunch of sentences will be given to extract the required useful features because to remove the noisy data that may involve spelling mistakes, special characters and symbols.

(b) *Feature extraction :*

The main goal here is to clean the data. So we perform the pre-processing step that includes Tokenization, Normalization,POS tagging etc. [39].

- **Tokenization:** which includes emotions and abbreviations such as OMG, BRB, each of which are considered as tokens.
- **Normalization:** Abbreviations present in the tweets will be considered and then those abbreviations will be replaced with their full form.
- **POS tagging:** here each of the word in the sentences will be identifies and positions will be given.

By performing the above step we are also able to identify informal intensifier such as sentence may contain same words in capital letters and repetition of characters may have found, those things will be evaluated. In case of capital words in sentence, capital letters will be replaced with small letters and character repetition in words will be minimized to single character.

Example:
I LOVE singing!!! -------→I love singing,
I am hungryyyyyyyyyyyy!!!!----→I am hungry

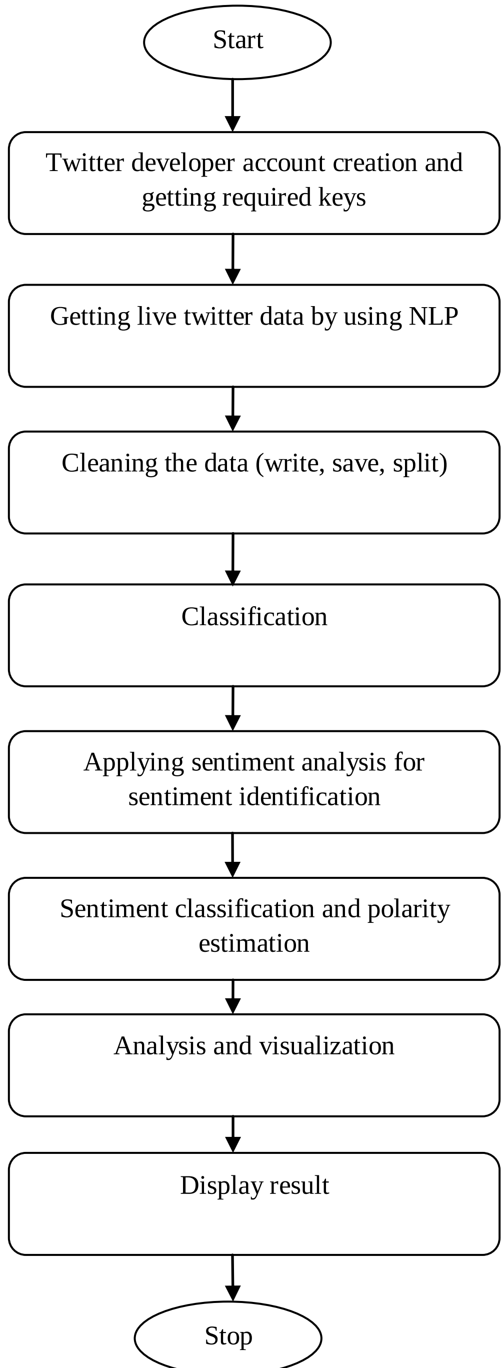

Fig3: Social Media

ISSN: 2278 – 1323

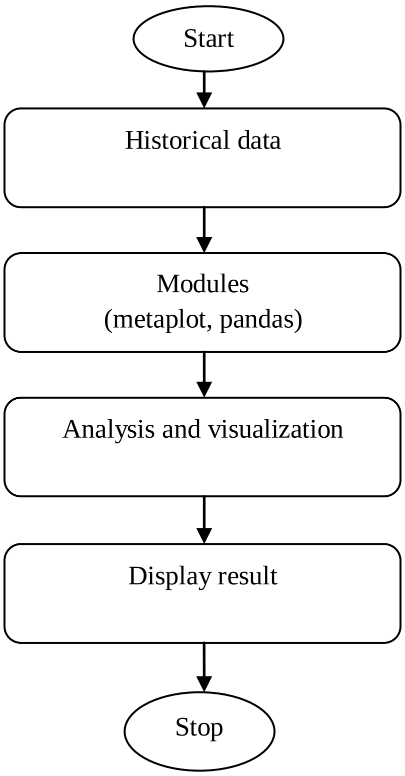

Fig2: Clinical data

(c)Identifying the subjectivity:

Subjective sentences are the sentences that are provided by author, by expressing their sentiment over an entity. There are various classification algorithms to identify the subjectivity of the given text. We majorly concentrating on Support Vector Machines (SVM) .It is a technique used in Supervised Machine learning approach in which we consider a given set of categories, which contain an arbitrary number of items the algorithm predicts which category a new items belongs to. Due to SVM's bi-classification nature, one to onevariant of SVM model was created for having multiclass classification. The main objective of SVM is that class getting the maximum votes will get selected. The main advantage why we go for SVM is , it produces accurate classifiers. We are dealing only with the subjective text segments and objective sentences will be treated as neutral.

(d)Sentiment Identification:

For each of the sentences sentiments will be identified like positive, negative or neutral. Here discrete emotions will not be considered, which does not detect sentimental strength of the text.

(e)Polarity estimation:

Here the object to which the opinion need to be provided will be extracted and for that particular object the estimation will be done

(f)Analysis:

Finally the result analysis will be done. For the clinical data, we are performing analysis by plotting thegraph and for twitter data we are ending with calculating the polarity estimation.

V. RESULT ANALYSIS

Results are evaluated for the clinical datasets by plotting the graph by considering the products of the particular datasets over the usage of that particular product. The graph shown in fig4 is plotted by taking the products in x axis, product usage in terms of percentage in y axis for our dataset, we get the graph as shown below:

(a) Result analysis of clinical data:

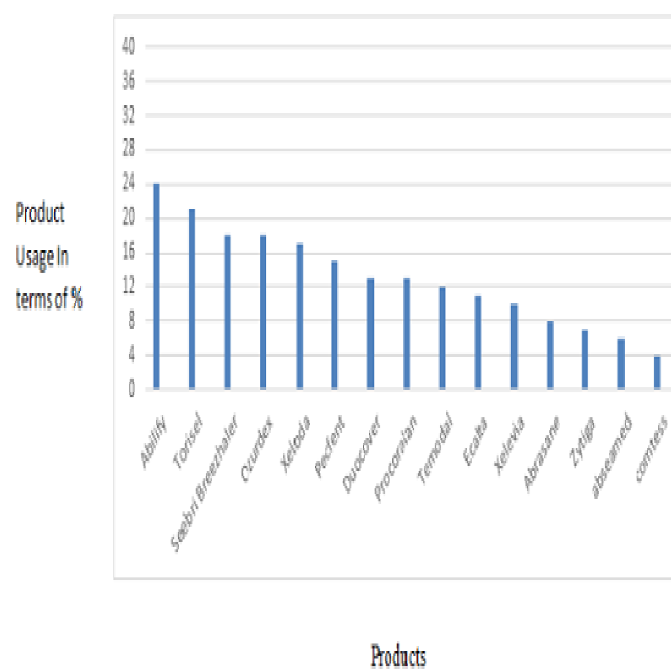

Fig4:Result analysis of twitter data

ISSN: 2278 – 1323

(b) Result analysis of twitter data

Fig5: Result analysis of twitter data

VI. CONCLUSION

Here we have concentrated on the problem of automatic text classification to detect the ADR's. Here we have used the NLP approaches and machine learning algorithms to improve the performance and classification accuracies.We are focusing both on social media data i.e., twitter data and clinical data which is publicly available from outside the social media to compare it scientifically. We have used the NLP techniques to extract the useful features and selection of those features in an efficient manner can improve the classification accuracies. Finally our experiments also shows that integration of data from several corpora can also improve the classification performance and time and costs may also be reduced.

FUTURE WORK AND ENHANCEMENTS

By using NLP and machine learning techniques we can classify the sentiments in the tweets. We feel that the classification accuracies can still be improved, in the other important aspects like

(a) Internationalization:

Generally we will be focusing only on English Sentences, but the social Medias like face book and twitter has many international users. Our approach need to be compatible to classify the sentiments in other languages also.

(b) Semantics:

Our machine learning algorithms used to classify the overall sentiment of the tweets. The polarity can be estimated on the way the users interpret the tweets.

(c) Emoticon Data need to be considered:

Emoticons were stripped out from the datasets, which means they were not influenced over the class. This need to be considered because emoticons are valuable features.

ACKNOWLEDGMENTS

We are grateful to express our sincere regards to all the faculties who supported and special thanks to our department forproviding the facilities that we needed to do this paper.

REFERENCES

[1]http://www.sssi.org.au/userfiles/event_doc1320138204.df

[2] B. Hesse, D. Hansen, T. Finholt, S. Munson, W. Kellogg and J. Thomas, \Social Participation in Health 2.0," IEEE Computer, vol. 43, no. 11, 2010.

[3] G. Eysenbach. \Medicine 2.0: Social Networking, Collab-oration, Participation, Apomediation, and Openness," Journal of Medical Internet Research, vol. 10, no. 3, p. e23, 2008

[4]Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009;
16:328–37.

[5] Friedman C. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In: Proceedings of the 12th conference on artificial intelligence in medicine (AIME); 2009.p. 1–5.

[6] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al.Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010;160(Pt 1):739–43.

[7] Ginn R, Pimpalkhute P, Nikfarjam A, Patki A, O'Connor K, Sarker A, et al. Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. In: Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing; 2014.

[8] A. Hassan, A. Abbasi, and D. Zeng. \Twitter Sentiment Analysis: A Bootstrap Ensemble Framework." Social Com-puting (SocialCom), 2013 International Conference on. IEEE,2013.

[9] A. Abbasi, H. Chen, and A. Salim. \Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classi-_cation in Web Forums," ACM Transactions on Information Systems, 26.3 (2008): no. 12.

[10]"Sentiment inTwitter Events" MikeThelwall, Kevan Buckley, and Georgios Paltoglouhttp://www.researchgate.net/profile/Mike_Thelwall/publication/220435012_Sentiment_in_Twitter_events/links/0912f51333fad59559000000.pdf

[11] Sharif H, Abbasi A, Zafar F, Zimbra D. Detecting adverse drug reactions using a sentiment classification framework. In: Proceedings of the sixth ASE international conference on social computing (SocialCom); 2014.

[12] "A Survey on Sentiment Analysis Technique in Web Opinion Mining", **S.Veeramani1, S. Karuppusamy2**http://ijsr.net/archive/v3i8/MDIwMTU4MTU%3D.pdf

[13]B. Qiu, K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G.E. Greer, and K. Portier. \Get online support, feel better{ sentiment analysis and dynamics in an online cancer survivor community." In Privacy, security, risk and trust, 2011 IEEE third international conference on social computing

[14] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas \Sentiment strength detection in short informal text." Journal of the American Society for Information Science and Technology (2010).

[15] A. Hassan, A. Abbasi, and D. Zeng. \Twitter Sentiment Analysis: A Bootstrap Ensemble Framework." Social Com-puting (SocialCom), 2013 International Conference on. IEEE,2013.

[16] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas \Sentiment strength detection in short informal text."Journal of the American Society for Information Science and Technology (2010).

[17] A. Go, R. Bhayani, and L. Huang. \Twitter sentiment clas-si_cation using distant supervision." CS224 Project Report,Stanford (2009): 1-12.

[18] T. Wilson, P. Ho_mann, S. Somasundaran, J. Kessler,J. Wiebe, Y. Choi and S. Patwardhan. \OpinionFinder: A system for subjectivity analysis." Proceedings of HLT/EMNLP on Interactive Demonstrations, 2005.

[19] A. Abbasi, S. France, Z Zhang, H. Chen. \Selecting at-tributes for sentiment classi_cation using feature relation net-works." Knowledge and Data Engineering, IEEE Transactions on 23.3 (2011): 447-462

[20]B. Pang, L. Lee and S. Vaithyanathan. \Thumbs up?: sen-timent classi_cation using machine learning techniques." Pro-ceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Com-putational Linguistics, 2002.

[21] A. Abbasi, F. M. Zahedi and S. Kaza. \Detecting Fake MedicalWeb Sites using Recursive Trust Labeling," ACM Trans-actions on Information Systems, 30.4 (2012): no. 22.

[22] A. Abbasi, T. Fu, D. Zeng, and D. Adjeroh. \Crawling Credible Online Medical Sentiments for Social Intelligence."Proceedings of the ASE/IEEE International Conference on So-cial Computing (2013)

[23] T. Fu, A. Abbasi, D. Zeng and H. Chen. \Sentimental Spi-dering:Leveraging Opinion Information in Focused Crawlers,"ACM Transactions on Information Systems, 30.4, 2012.

[24]Harpaz R, Vilar S, DuMouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2012;20(3):413–9.

[25] Tuarob S, Tucker CS, Salathe M, Ram N. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. J Biomed Inform 2014;49:255–68.

[26]Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: Proceedings of the 2010 workshop on biomedical natural language processing; 2010. p. 117–25.

[27] Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. J Biomed Inform 2011;44:989–96.

[28] Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety signal detection. In: Proceedings of the 2012 international workshop on smart health and wellbeing; 2012. p. 33–44.

[29] Yates A, Goharian N. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In: Proceedings of the 35th European conference on advances in information retrieval; 2013. p. 816–9.

[30] Jiang K, Zheng Y. Mining Twitter data for potential drug effects. Adv Data Min Appl 2013;8346:434–43.

[31] Liu X, Chen H. AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In: Proceedings of the 2013 international conference on smart health; 2013.p. 134–50.

[32]Nikfarjam A, Gonzalez G. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In: Proceedings of the American medical informatics association (AMIA) annual symposium; 2011. p. 1019–26.

[33]Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug event from medical case reports. J Biomed Semantics 2012;3(15).

[34]Liu X, Chen H. AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In: Proceedings of the 2013 international conference on smart health; 2013.p. 134–50.

ISSN: 2278 – 1323

[35]Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. In: Proceedings of the American medical informatics association (AMIA) annual symposium; 2011. p. 217–26.

[36]Pimpalkhute P, Patki A, Gonzalez G. Phonetic spelling filter for keyword selection in drug mention mining from social media. In: Proceedings of the American medical informatics association (AMIA) annual symposium; 2013.

[37]Fextor: a Feature Extraction Framework for Natural Language Processing. A case study in Word Sense Disambiguation, Relation Recognition and AnaphoraResolutin.

First Author: Sahana D S was born in Karnataka, India. She received the B.E Degree in Computer Science and Engineering from Visvesvaraya Technological University Belgaum Dist., India in 2013 and Mtech Degree in Computer Science Engineering from same University. Her research interests are in the area of Data Analysis, Sentimental Analysis and NLP.

Second Author: Shantala C P M-tech, was born in Karnataka, India. She is working as Vice Principal and HOD in Computer Science Engineering Dept., CIT, Gubbi, Tumkur (district), Karnataka, INDIA.

Third Author: Girish L M-tech, was born in Karnataka, India. He is working as Assistant Prof. in Computer Science Engineering Dept., CIT, Gubbi, Tumkur (district), Karnataka, INDIA.

ISSN: 2278 – 1323