TWITTER DATA SENTIMENT ANALYSIS AND WORD CLOUD

¹Dr.R.Vijayabhanu, Assistant professor, Department of Computer Science, Avinashilingam Institute for Home science and Higher education for women, Coimbatore, India

²R.Shenbagam, PG Student, Department of Computer Science, Avinashilingam Institute for Home science and Higher education for women, Coimbatore, India

Abstract- Social networking plays a significant role in our day-to-day life by interacting around the world about any business and social products or services. The opinions, suggestions and the reviews on various disciplines on social mediums share a remarkable place in marketing and promoting. Sentiment analysis or opinion mining is the study of public opinions, sentiments, attitudes, and emotions that are expressed by the people and a dynamic area in natural language processing and text mining in current years. Being a huge social website, Twitter consists of large users posting and disputing their information. One such trending domain like Politics has been analyzed in this paper. Sentiment analysis is a chain of various processes such as data mining and machine learning. In this paper, sentiment analysis has been carried out on the politics dataset by sentiment scoring, classification and performance evaluation.

Keywords: - Sentiment Analysis, Opinion Mining, Natural Language Processing (NLP), Lexicons, Polarity Classification.

1. Introduction

Being an encouraging and active platform of interactions, Social Media becomes a library of volumes of various trending subjects revolving around the globe. One such application where people are free to post their opinions and suggestions about a commodity or have their comments on any discipline of public interest is Twitter. Around 200 million tweets per day are posted by the users as by communicating their thoughts on hot topics independently.

According to a couple of survey reports conducted among 2000 people had proved that about 81% of online users research at least once about a commodity and between 73% and 87% have reported that a major impact on the purchase of a product relies on the internet reviews [1]. Therefore, Sentiment analysis systems are being useful in almost every business and social domain because opinions are important to almost all communal activities which also a fundamental influencing factor of behaviours. Decision making is a complicated tasks, which rather be simplified with others' opinions. Sentiment Analysis is the computational handling of opinions, sentiments and subjectivity of text. But while managing Natural Language Processing (NLP), the problems such as Sentence boundary detection, Tokenization, Part-of-Speech assignment to individual words, Homographs, etc are include [2]. Thus, text mining a non-traditional information retrieval strategy to reduce efforts in obtaining information from large set of textual documents is required. The R, an open source programming language for statistical computing and graphics widely used and best known among statisticians and data miners for developing statistical software and data analysis can been used in such occasions [3].

In this paper, the sentiment analysis has been initiates with data mining process. The dataset has been pre-processed and subjected to sentiment scoring and sentiment classification. At last the performance evaluation has been achieved on polarity classification. To implement this scheme, Politics Dataset extracted from a Twitter account with 8000 tweets and 14 attributes has been used.

2. Literature Review

Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python [4] proposed by Bhumika Gupta *et al.* analysis the opinions of the Twitter users by combining various techniques and methodologies such as data mining and machine learning. An effective interrupted programming language like Python has been used for pre-processing the dataset. Scikit-learn, a Google Summer Code enhances the system with various machine learning classification algorithms and other tools of mining and analysis. The machine learning model broken up by the training data set provides a model of trained machine learning which has been further intermittent by the testing set from the original data set to provide a conclusive classified data. Thus, the paper enhance the sentiment analysis on the Twitter data with various sentiment classifiers and categories their result accurately.

Chetan Kaushik *et al.* has researched a paper on A Scalable, Lexicon Based Technique for Sentiment Analysis [5] by experimenting on sentiment analysis by using lexicon based technique rather than choosing a conventional text mining approaches. An expanded data set of around 6,74,412 tweets are studied using the Hadoop, an open source software. Being a flexible processing and analyzing technology, the Hadoop has been used in predictive analysis of both constructive and non-constructive data. The data set has been approached using sentiment lexicon or dictionary, thereby allocating the polarities of the words in their respective domains. To

optimize the time consumption, the feature detection of the subject towards the sentiments has been performed by using the Twitter hash tags. The negative and blind negations are focused during the polarity of the sentiments for the betterment of the output. Therefore sentiment calculations have been executed on each and every tweet thereby calculating the polarity score. Thus, the paper claims 73.5% of accuracy performed in 14.8 seconds on 6,74,412 tweets. Though the accuracy detected has not been comparative with the machine learning techniques but it has been better than other lexicon based approaches by consuming less time.

Sentiment Analysis on Twitter data [6] by Varsha Sahayak *et al.* describes a paper on sentiment analysis by using data miming and machine learning methodologies. The sentiments are categories by polarity of the words such as positive, negative or neutral. The lexicon-based approach of the proposal has been based on both the corpus-based and dictionary-based techniques. The initial step involves data mining of the tweets like extracting and pre-processing the data set. The extracted tweet data are classified by using machine leaning approaches such as Naive Bayes classifier or Support Vector Machines (SVMs). Thereby has been used as the training data set, the process of classification takes places followed by the sentiment scoring. Thus, the opinion mining has been executed automatically by classifying the sentiments of tweets obtained from Twitter data set.

The paper **Sentiment Analysis on Twitter** [7] directed by Akshi Kumar *et al.* demonstrates a novel approach of opinion mining by combining both the corpus-based and dictionary-based algorithms. The tweets from the Twitter dataset are preprocessed by using data mining techniques to extract the sentiments which are subjected to the scoring module. The semantic score of the sentiment carriers i.e. the parts of speech such as adverbs, verbs and adjectives are analyzed by the novel approach. The verbs and adverbs are scrutinized by the dictionary-based method while the adjectives are analyzed by the corpus-based method. A linear equation of letters, adjective groups, exclamations, emotions, etc has been used to calculate the tweet sentiment score respectively and to obtain the opinions on the Twitter data set.

Walaa Medhat *et al.* have studied a paper on **Sentiment Analysis Algorithms and Applications: A Survey** [8]. The paper illustrates a complete synopsis on Sentiment Analysis (SA) or Opinion Mining (OM). The classification in sentiment analysis can takes place either in following three ways such as document-level, sentence-level and aspect-level. The paper briefly discuss about both the machine learning and lexicon-based approaches. The feature selection during sentiment classification has been studied well. The illustration based on the number of articles that has been targeted for the sentiment analysis with percentages invokes the better understating of the subject. Various problems that might frequently occur during the opinion miming have been focused. Therefore, the survey paper has been played its importance in empowering the knowledge of Sentiment Analysis (SA).

3. Proposed System

The lexicon-based sentiment analysis consists of Dictionary-based and corpus-based approaches. The lack of finding the opinion words with respect to domain and context specific orientation is the major drawback of dictionary-based approach which luckily can be overwhelmed by the corpus-based approach. A collection of known and pre-defined sentiment terms are recognized as Sentiment Lexicons plays a vital role in lexicon sentiment analysis.

The proposed system consists of four phases:-

a. Data mining / Text mining: Pre-processing the raw data to extract the knowledge.

b. **Sentiment Score**: Loading of the predefined lexicons and calculating the positive / negative score by comparing the tweets terms with positive/negative term corpus and summing the occurrence count.

c. **Sentiment Classification**: The essential phase of organizing the text according to the opinion's; sentiment polarities (positive, negative, neutral) [9]. The pre-processed tweets are taken as input to sentiment classification which can be done by using two lexicons such as emotion lexicon and subjectivity lexicon. The emotion lexicon is used to classify the emotions, such as joy, fear, sadness, disgust, surprise, anger and unknown while the subjectivity lexicon is used to classify the polarity such as positive, negative and neutral.

d. **Performance evaluation:** Compares both the values of sentiment score and lexicon based algorithm and visualizes in the form of graph.

The system has been implemented in the R Studio which has been consider to be the best at text mining and preprocessing and it helps in the improvement of the system accuracy [10]. The systematic flow has been illustrated in the Fig1.



4. Module Description

4. 1. Load and standardizing the dataset

The raw data has been pre-processed to extract the usable dataset are loaded into the R studio as shown in the Fig 2.

And the last of the of the second sec	A form
Twitter Sentiment Analysis	
The set of	ter finning kente belan Analitikan Birawan Pe Sull Tere Teatment Ter Belandager Ann- Pe Teather Kentenset Person and Ann
A negative and the second s	The fail is achievished by perioding has dugrian Twitter of the Kowlion you
	RR Nominaationer Nominaationer
	Providence States of Colors
	All presents of location to how formed
	software for a second
	Light mode
	Page 4
	ty in And
	No. 8 Contraction of the second
Fig 2. To	p Trending Tweets Today

Text mining methodologies allows in highlighting the most frequently used keywords in a paragraph referred as the word cloud or the text cloud or the tag cloud, which is a visual representation of text data. In this work, the word cloud has been generated for Politics Dataset in which the most frequently used spot words are displayed in larger size and bolder format as shown in the Fig 3.

Name a survey of a loss of a stational of a station of	by housing boots have the second the part for the bir modern by the second of a
	Mole used weats avectable with the hundring
A DESCRIPTION OF THE OWNER OF THE OWNER	
	E description in the
	after the solid stated among
ments.	arriver and the second second
	AND A LOCAL DIVISION OF A LOCAL DIVISIONO OF A LOCALDI DIVISIONO OF A LOCAL D
	"Inditios"
	a pondoa Lat
	which the second party is not been as a second with the second party of the second par
	Value and a second
	second a property of the formation
	A work tool of a linear international straid part, provider rate to deput exercise the down
	vapone, and vauge the fore test is a construction of sectors because the experimental test or the sector of the se
Fig 1 Word Clay	id using 116 words based on Politics

A graphical representation of patterns in a set of data by plotting how often particular values of a measure occur is a frequency plot. The horizontal axis of a frequency plot graph shows grouping of a continuous measure (e.g. age, time, no. of occurrences) while vertical axis shows the number of times that a value in that group was seen. Here, the frequency plot has been generated for larger words in the word cloud generation.

4.2. Sentiment Score

The positive and negative lexicons are downloaded using the available links and loaded in R studio. The corpus contains around 6789 words. The positive / negative score are calculated by comparing the tweets terms with positive/negative term corpus and summing the occurrence count. The histograms of positive sentiments and negative sentiments are demonstrated in the Fig 4 and Fig 5 respectively.



The sentiment score is the more precise numerical representation of the sentiment polarity. It calculates the score for each tweet. The more number of words that matches the positive list from the tweet gives positive score and the number of words that matches the negative list from the tweet gives negative score which has been explained in the Fig 6.



4.3. Sentiment Classifications

R package sentiment by Timothy Jurka has a function that helps us to analyze some text and classify it in different types of emotions: anger, disgust, fear, joy, sadness, surprise and unknown. The emotion classification can be performed using the algorithm called lexicon based sentiment algorithm trained on Valitutti's emotion lexicon. Another function from sentiment package named classify polarity allows us to classify some text as positive or negative. In this case, the polarity classification can be done by using a lexicon based algorithm trained on Janyce Wiebe's subjectivity lexicon. The following Fig 7 indicates the pie chat of negative and positive sentiments.

Intrational and a second state	Information has produce integer from the Aphenese Aphenes
eyeka-	Pie Chort
	Nonfraent Longen
	Pentitus
	lingstrum
	A per char to a discuss matching pupple, which is already the science to increase the section of the having of a dash which and consider to its control of a dash with the section of the science of the
	Fig 7 Pie Chart

4.4. Performance Evaluation

While the performance evaluation on polarity classifications, both the values of sentiment score and lexicon based algorithm are compared and visualized in the form of excel sheets. By the performance evaluation and result shown, the lexicon based sentiment algorithm accuracy has been better than the sentiment score accuracy.



The above Fig 8 is a bar chart representation of top 20 tweets of hash tags and the following Fig 9 contains the Top 20 Tweeters.

	64	May Discovery	18			
	14	laget.				
	3.8		1.2			
	10	NUME .	1.5			
	1	0446,140				
	100	editivi				
	470	overti 4				
	82	#101.	10			
	10	c_pome	1			
	14.	Au .				
	12		<u>51</u> 0			
	12	-	12			
		And non	242			
	14	ri her	140			
		10.0				
		100				
	10	ut's	77415			
THE OF THE P.S. O	V on some low thank larks	en Taat feel - Serg			1000	121
n in Car Ver Per O	V etc. encomposition of Station Lines of 1 (11) 11) 111 (Over to Internal)	es haat feel - Sery		1	and and S factor	1 Part
Provide the Pro 0 To 10 art for the Pro 0 To 201 of advectors S 2	View acceptions france france a system () Over House D'Andata Sale parates 2013 V	ne bashiri - Jang 	here here Protided		S face	Bhart (ar
Here and the second of the second sec	V - a constant for States Tables A RELISE Constant for a lower Effort data bis president with V results	na Track find - Garge Then Tracking Then Tracking	There Said Tracilla	i nayar ta	Schese -	R har t
The last two two to 1 - 10 - 20 - Yes Two To 1 - 10 - 20 - Yes Two To 1 - 10 - 20 - 10 - 10 - 10 - 10 - 10 - 1	Viria anna hain 17000 Viria A 1903 R. Cherk Bann D'Ar dau Sa seanna 200 Y rgalas	ne familie - Song Son Fransy Tachiners Brier can the	Thema Tinge - Provident Tige Headbage of Low	i naya Pr	S Pasa . S Pasa .	Data (ar-
The second secon	Mental second long Matter Lance a DECESSE - Over Houry radia - Radra di hour Houriston da Radra di houristonico da Radra di houristonico da	The brown The brown The brown Deter case the Overset as	There is Theory - Ministration Togo Houstage of Lear Cha	i iti	ganta an Citata Ciel San	Prat 1 (u-) webbar-an' Webbar-an'
The Date Value Plane The Date Value Plane	27 otto demonstratives 17 Stands Lineau an International Control Alexandro Facilities Review of reservicions to are for anyon as the second standard and the second for anyon	No beauty See Young To beauty Date can be granted as	There a Table Tap Herburg of Low C	· ··	S Asia	in the second states
	V one or water bases that is a latter a triction of the second of Y and a second of the second of Y Reads of the second basis to us for anyon.	na haatiida Dang Tao harang Tao harang Data can tha Quatra tao Quatra tao	These Traces Tag History Styles Tag Sta	P	Ç Asın Ç Asın	in the land of the
	Constructional Florida Series STUELISE Overhour Parks Reaction described to all Reaction described	er haafed - Jay Jac Henry To heers Dese caalita Queto Erre Lant	There is Name - Provident Tap Histologi st (ver 19 00 Hestog keyendes in Arch	a mayar da]	Shann Shann Carl Non	E har 1 (ar- (ar- (ar- (ar-))) (ar- (ar-(ar-))) (ar- (ar-(ar-))) (ar- (ar- (ar-))) (ar- (ar- (ar- (ar-))) (ar- (ar- (ar-))) (ar- (ar- (ar-))) (ar- (ar-))) (ar- (ar-))) (ar- (ar-))) (ar- (ar-))) (ar- (ar-))) (ar- (ar-))) (ar-))) (ar- (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-))) (ar-)))) (ar-)))) (ar-)))) (ar-)))) (ar-)))) (ar-)))) (ar-)))) (ar-)))) (ar-))))) (ar-))))))))))))))))))))))))))))))))))))
	Month on example and an Alfred	er haafed-berg he henry he henry he henry he henry genoders journoters	Thema Theorem The State Tableting State 00 Tableting Negeritions in the Tor	a magan Pa	gana . Gan . Cel Sar	Elbert 1 Series Weiterzer Liefer Michael 2 ar Liefer Michael 2 ar Liefer
	Control of the second of the s	nd haalide - Sang San Tronsong Taxahanang Dari ya sang Dari ya sang Dari ya sang Dari ya sang Dari ya sang Dari ya sang Dari ya sang Da	There have Tractical Tractitional struct of Therefore provide a flat to the the provide a flat to the provide a flat to the the provide a flat to the prov	a surger Pr	To Passe	and a second sec
	V etc. completions fillend. John Controller V etc. Soor Status de session de V Vanisa. Rocar et noartitueis se us Krange.	n haddel dae he henry To heres Bre califs (arrest in terms heres	There is There: The section Tag iteration of A your Tag iteration of the section of the section of the section of the section	i saya Pr	Cool Same	The Part of the Pa
	V oto in memory of the distribution of the dis	ny faatiful - Jany Tuo Yoong Tuo Yoo	There a Tanger - Providence Tag Hostoage of Lyner Co Co Co Co Co Co Co Co Co Co Co Co Co	a manyan Pa	Solar and S	
	Construction of Construction o	ny haatika Jany Yaa Yaany Kathanin Kathanin Kathanin Kathani Kathani Kathani Kathani Kathani Kathani Kathani Kathani Kathani Kathani	There have investee in the second sec	a sayar Pr	See	en datar zelana en datar zelana en datar zelana eta lan esti bat eta lan esti bat
	X one completes filmed. John Controller (Control on X) Faints - Control of exercises to us Krange.	erinatio - Jaco San Transport Barteran Barteran Barteran Santar Santar Santar Santar	The share the second se) Saya Pr	State	A Constraint of the second sec
	V oto in memory of the distribution of the dis	erlaubi-ber Ter transp Terte caarita gurod m termini t	hers here in traction taintings Joint of	a saya Pa	Stream of the second se	Index 10 Parts 10 Par
	V ette ennen innen filmen innen grades i Recard ef naertbeels aus Kronges Recard ef naertbeels aus Kronges	in testing over the transport to be a second of generating to be a second of generating to a second of generating to a second of the second of the second testing test	There have needed to be a first the second s	a sugar to	Great See	
	X one of the second of the sec	instantial Carry Tax Young Young Y	There have the state of the sta	e surger pr	See See	
	V oto in memory of an all provided in the second se	In transition 2 any Tao transition Tao transition T	Here Sher Procides	a manyare fre	State.	
	Mente consequences Millande, Jackson Consecutives Consecutives and Mente Restarts of securitizes to use Restarts of securit	In transition Composition Contractions Destinations Desti	hensing Pocha	a sever for	State	Internet in the second
	March an experiment of March 2 Defends Control of the Second Sec	an Institut - Sary Tan Turang Tan Turang Tan Turang Tan Turang Tan Turang Tan Turang Tan Turang Tan Turang	There have the states		States	A read at the second se
	V ette entret index fillende Johnson Transfer in State - State	and matched - Samp	Persive Public Tarting June 0	a manyon fro and a toring	Sheet	

Hash tag frequency in the tweets of the tweeter has been visualized in the Fig 10.

5. Conclusion

People are using social media on daily basis to communicate their opinions on a wide variety of subjects, products and services, which has made it a rich resource for text mining and sentiment analysis. The sentiment classification done by using two lexicons such as emotion lexicon and subjectivity lexicon are visualized in the form of bar charts for better understanding. The performance evaluation of polarity classifications by comparing both the values of sentiment score and lexicon based sentiment algorithm are visualized in the form of graphs has enhances the system accuracy which has been consider better than the sentiment score accuracy.

References

[1] Opinion Mining and Sentiment Analysis, B. Pang, L. Lee, Foundation and Trends in Information retrieval, Vol 2, No 1-2, 2008, Pp: 1-135.

[2] Natural Language Processing: An Introducyion, Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, J Am Med Inform Assoc, 18, July 2011, Pp: 544-551.

[3] Txt Mining in R, Kasper Welbers, Wouter Van Atteveldt, Kenneth Benoit, Communication Methods and Measures, Vol 11, No 4, Nov 2017, Pp: 245-265.

[4]Study of Twitter Sentiment Analysisusing Machine Learning AlgorithmsonPython,BhumikaGupta,MonikaNegi, Kanika Vishwakarma, GoldiRawat,Priyanka Badhani, IJCA, Vol165, No 9, May 2017, Pp: 29-34.

[5] A Scalable, Lexicon basedTechnique for Sentiment Analysis, Chetan Kaushik, Atul Mishra,IJFCST, Vol4, No 5, Sept 2014, Pp:35-43.

[6] Sentiment Analysis on Twitter data, Varsha Sahayak, Vijaya Shete, Apashabi Pathan, IJIRAE, Issue 11, Vol 2, Jan 2015, Pp: 178-183.

[7] Sentiment Analysis on Twitter, Akshi Kumar, Teeja Mary Sebastian, IJCSI, Vol 9, Issue 4, No 3, July 2012, Pp: 372-378.

- [8] Sentiment Analysis Algorithms and Applications: A Survey, Walaa Medhat, Ahmed Hassan, Hoda Korashy, ASEJ, May 2014, Pp: 1093-1113.
- [9] Sentiment Classification and Polarity Shifting, Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu- ren Huang, Guodong Zhou, ICCL, Beijing, Aug 2010, Pp: 635-643.
- [10] Text Mining Scientific Articles using the R Language, Carlos A. S. J. Gulo, Thiago R. P. M. Rubio, DSIE 1st Edt, 2015, Pp: 60-69.

