# International Journal of Science, Engineering and Management (IJSEM) Vol 3, Issue 4, April 2018 Opinion Mining On Twitter Data for Sentiment Classification

<sup>[1]</sup> G. Anuradha, <sup>[2]</sup> CH. Raga Madhuri, <sup>[3]</sup> M.Srikanth <sup>[1]</sup> Associate Professor, VRSEC, VIJAYAWADA, <sup>[2] [3]</sup> Assistant professor, VRSEC, VIJAYAWADA

*Abstract-* Opinion mining and Sentiment analysis gives sentiments, opinions and subjectivity of text. Now a days many people express their opinions and ideas through social networking sites like Face book, Google+ and Twitter. These are platforms to allow people to share and express their views about topics, have discussion with different communities. Twitter data is short and continues data suitable for sentiment analysis. This paper focus on sentiment classification (positive, negative and neutral) which is multistep process involves preprocessing phase, parts of tagging (POT), and calculating polarity and apply three classification algorithms that are Decision tree, Naïve bayes and Support vector machine on twitter data in Jupiter. This paper also presents empirical comparison of classification algorithms in which decision tree algorithm is highest accuracy in comparison of all three algorithms considered in this study.

Keywords: Naïve-Bayes, Decision Tree, Support Vector Machine, POT, Polarity.

## I. INTRODUCTION

Micro blogging websites have become a source of various kinds of information. In micro blogs, people post real time messages about their opinions on various topics such as discuss current issues, complain and express positive sentiment for products they use in daily life. Twitter is such a platform where people expressing and sharing their thoughts and opinions on all kinds of information or opinions about products, politicians, events etc. Sentiment detection of tweet is one of the basic research topics over twitter data and it is very useful because it allows feedback to aggregate without manual intervention. From the sentiment analysis many people like consumer, businessman or may be typical person get benefit and many companies study user reactions and reply to users on micro blogs. One challenge is to build technology to detect and summarize an overall sentiment. In this paper, we look at Twitter and build models for classifying "tweets" into positive and negative sentiment. Twitter is a micro blogging platform where anyone can read or write short form of message which is called tweets. The amount of data accumulated on twitter is very huge. This data is unstructured and written in natural language. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negative or neutral with the help of different machine learning algorithm. We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. It includes rich structured information about the individuals involved in the communications. For example it maintains information of who follows whom and re-tweets and tags inside of tweet provide discourse information. Another reason is that the

amount of relevant data is much larger for twitter, as compared to traditional blogging sites. The response of twitter is more prompt and more general. Sentiment Analysis is process of collecting and analysing data based upon the person feelings, reviews and thoughts. Sentimental analysis often called as opinion mining as it mines the important feature from people opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data. Sentiment Analysis can be done at document, phrase and sentence level. In document level, summary of the entire document is taken first and then it analyse whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment. Sentimental Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelings or thoughts which they provide in form of text. Sentiment Analysis is domain centered, i.e. results of one domain cannot be applied to other domain. Sentimental Analysis is used in many real life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or market. The layout of the paper is as follows: Section II discusses about the Design Framework.. Section III presents Experiments and Results. Section IV concludes with the Conclusion.

#### **II. DESIGN FRAMEWORK**

In this section, we represent an overview of the architecture which resembles process in fig1. The following steps will expand the process of the proposed system,



# International Journal of Science, Engineering and Management (IJSEM) Vol 3, Issue 4, April 2018

- 1 Collect Tweets
- 2 Preprocessing of Tweets
- 3 Feature Extraction
- 4 Data Modeling



The Fig:1 shows the diagram in schematic form the general arrangement of the parts or the components of a complex system or process, such as an industrial apparatus or an electronics circuit. The block diagrams are heavily used in engineering in various hardware designs.

## 1. Collect Tweets

Twitter data is short and continues data suitable for sentiment analysis. To improve visibility of the content user generally used Emoticons for pictorially facial expressions, "@" symbol used to refer other users and Hashtags. Hashtags are used to mark topics.

The Initial Data or the Raw Data is collected from Twitter for analysis purpose. Overall 1400 tweets are collected.

## 2. Preprocessing of Tweets

The reviews are pre-processed primarily to make them noise free at the classification part as the review sentences may describe about two or more features which will be difficult to classify if the sentence has positivity and negativity associated[1]. So the features if two or more are present in a sentence are splitted in order to eliminate the above case And we also took into consideration that the reviews scraped from the websites [2] will contain some spelling mistakes which will be serious issue. Example: 'Gud' may mean nothing to the computer which if corrected does make some sense like positive or negative. Thus the preprocessing includes these steps to make review sentences noise-free along with Stemming as it is the standard procedure to make the words cut short to make match with the features.

## **3. Feature Extraction**

Large number of features without considering un important features [3] are extracted. Thus the sentences are first tokenized to check for the words which are adverbs/adjectives and nouns. Then we apply a POS tagging Adverbs/Adjectives and to select Nouns. Here adverbs/adjectives and nouns are only considered because the word which describes the feature in a review sentence is an adverb or a noun. Example: The body of the mobile is fragile. This sentence when tokenized and POS is applied look will like: [(The,DT),(body,VBP),(of,IN),(the,DT),(mobile, NN), (is,VBZ),(fragile,JJ)] Here we take the sentence to see if a feature is present in it or not to check if the sentence is worth further processing. If it does not have any feature, we continue with other review sentences. Senti-wordnet is a large database consisting of many words associated with its pos and neg score [4]. Example: 'Good' has a positivity score of 0.75 and negativity score of over 0.25

#### 4. Data Modeling

For sentiment classification on twitter data, here we are using decision tree[6], naïve bayes[5] and support vector machine. In the proposed feature selection, a Decision tree induction chooses relevant features using following formula:

$$\inf_{i=1}^{m} p_i \log_2(p)$$

For naïve bayes following is used:



Above, • P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).



# International Journal of Science, Engineering and Management (IJSEM) Vol 3, Issue 4, April 2018

- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor

SVM are supervised machine learning methods used for classification, regression and detection models. SVM are more effective for high dimensional space. SVCs are capable for multi-class classification. SVC and NuSVC are similar whereas, LinearSVC are based on linear kernels.

#### **III. IPLEMENTATION AND TESTING**

This is implemented in Jupyter on Windows Download the Jupyter software from the official website. http://jupyter.org/ and click on Download the Jupyter. Fig 2 shows tweets data set. Fig 3shows important features in given data set.

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1					-						
		n Tellan	-		- Phileseeper	-					
30	1410	-	111	I Di Brain	and a	一限	副目	游家	<b>前</b>	Xiata	ÀT (
-\$90er+	4.1			T HA DAVE	Der 10.4.1	12:00	faces, fat	bart Total	ine :	見たい	Ark to
lanet 1	1	Feel	19	ineres .	i here	and the second second	Mar Spin-	12		P Laur	Re-Se
		61		100							
	-										T.
	in the second	d'au									1
3	1	is to be from the	ARLINES								
<u>t</u>	-	a minute for he	w Roon Analys	lir.						_	
5	-0	Alter and the second second	manwith.	for last at \$20 det \$10 per	all. insurgered by	(getaction) ad re	planet.				
	184	Ren Marchant	iftee expe	ų.	and the second sec						
<u> </u>	21	out workly list th	unter .							-	
\$	Selet	Agen wheth	Torestine -	18.Toright							
4	herb	d in republicat	tobs .linkes	pro alterrity							
100	1000	the strategy and	now always the always	198 gH					_		-
úi.	Patt	In all the heat	THE PARTY OF	wine 12.40							
ju -	file	extendital aut	adurhe .							-	
24	0.04	incomplexed bitrayers	tes ay nov								
		and the second s									
111		kt	This is the s	ay their ghtnew		222121022	at 122200	ALC: 1			
1 15 8 16		kt; avitura	This is the w	ay i fasi nght now pletaly utaliets it now	. Funny, all I can do	h heither. hitp	//m/ac.ma	ITHE			
17 15 8 16 17 17		ikt; svitto nar Feeling sta	This is the s Fre com ingely fine.	ny i fasi nghtrew pistoly utaliess rtroo Nev ('n gonna golist	i. Funny, all I can do an to some Sensio	o in twitter. Hittp mic to colledinate	://m/ioc.maj	ITVE			
17 15 18 16 19 17 10 18		ikt;	This is the w r Precom ingely fine. F thurder ju	ny i fael right new pletely utalies it now New (in gonta ge lid at now50 scary(11	i. Funny, all I can do an to some Saniso	s in twitter, hittp: mic to collebrate	(//m/loc.ma)	1748			
T 15 8 14 17 17 10 18 17 19		iki; avittoma Roologito HUSErolio Ljuttotm	This is the w r I've com ingely fine. I thurster ju y beant off.	ny i fasi nght new pletely utaliers rt now New fin gonta go list et nowSO scary(11 if's only been growing	i. Forey, all Ican do an to some Sensio gfor well over a ye	o in twitter, hitp nic to colednate ar. Pin germa sta	//wy/oc.may	1741 Mariamano	shap	oy in the r	ventrie
17 15 18 14 19 17 10 38 11 19 12 20		iki; avitto ma Reeling sto HUGE roll o Ljust out m Very sad al	This is the s r	ny i fasi nght new platoly utaliasi it now New fini gonta go litt st now30 scaryi ili if's only been-growing	i. Funny, all i can do an to soma Samiso gfor well over a ye	a in tueldtuer. Hetpo mis to collediratio ur. Pim generus sta	(//m/ioc.ma) intitover. (Pr	JTKI Nahimanu	shap	ay a ber	eartine
17 15 8 18 9 17 10 18 11 19 12 20		iki; avitho mar Fooling sto HUGE roll o Ljust out m Very sad of	This is the w r Fire corre angely fine. If thursder ju y bearst off, rout itan,	ny theirghtnes plately useless three New I'm parsage let atroes. 30 scary(11 filority beer growing Filor2+ T	i. Forey, all i cardo en to come Servico gfor viell over a ye weets. T	onteitar. http: nationekkute ar. Imgomesta Data see	(//mp/ac.ma) ntitover. (Pr	ITHE Nationality	is hap	ay a der	nantme
T 15 8 18 9 17 0 18 1 19 1 20		kit;	This is the w r Fire com angely fine. If thurster ju y beard off, roud itam,	ng theil right new plately united if new New fin generage list at now50 scary(III f1 only been growing Fig2: T here a cod	i funy, all cando anto constantio gfor vellover a yo weets T	on norther, bety na the collectory or . I'm germa cha Data see	(//m/oc.ma) ntitover.ge	ITHE Normanu	it hap	ay a ber	natue
13 83 77 13 13 12 12 12 12		kt; aviterna Reelegits HUGEnsic Ljuttotm Veryadis [ 1.5	This is the w r Fire comp angely fine. If thurster pu y beard off. roud item.	ny fael right rev plandy usaless it now New (in generage list at root, 50 sam/11 fits only beer, proving Fig2: T word	Forey, all can do an to come Service gfor well over a yet weets T I s	ola tarittar. http eacto calebrate ar. Progerna da Data se	(//m/oc.ma) ntitover. (s et		i ti haqi	a) a ber	native
T 13 8 16 7 17 10 18 17 19 12 20 <b>T r</b>	, E	kit	This is the w rull fine comp angely fine. If thanker ju where for fill work from thanker is a state of the southern is a state is a	ny helingtree. Jean welen too New Inspiracy lit from 50 sam/11 from bergrown Fig2: T woord ['sourd	Forgul Londs in to one Series gfor well over upon weeets I s	oanattar. http natio oakabata n. I'mgomasta Data se	(/my/ac.ma) ntitovet. (Pr et	iha Maramato R	it hap	a) a ber	estre.
T 13 8 16 7 17 10 18 17 19 12 20 <b>T r</b> OL	it.	kit	This is the w number of the net o	ni heingtnes. Jeak usles tros Nev in prouge lit stros. 30 say (II from Jean power Fig2: T word L:sa	fory, all cards and core Series gfor well average weets I is i and	a taitar. http: nicto celebrat p. Tingorna da Data se	(/m/oc.ma) ntitover. (P 2t	ina Noramano R	it hap	a) a ber	ertre
10 81 97 03 10 22 22 17 22 20 17 00	t	kt; avitona Foologitu HUGEnilic Ljuttotm Verysetet [ 15	This is the w number of the other pury present off, to ut item. 3 =	ini heingtree. peoly usies it on New in proceed it those. 30 sam/ill those beer poors Fig2: T woond [:sam fr those to not	fory, ill or to internet serve gravel nerve weets I is d • . i end aile	o a haritar, lettp ench collebrar n. I'n gorrada Data se	(/mplac.ma) ntitiover. (Pr 2t	itea Noramaro	iti hap	a) a ber	estre.
пв 811 717 03 13 22 22 17 Ос	N IL	kt; avitona Foologitu HUGEndic Ljuttotm Verysatit [ 15	This is the w 	in heingtness. people sales theory here in presign in theory bergrown Fig2: T woord [:sa from tr al	tion, all contained producers are weets I is iend aile read	e hatte itte echoktes Data se	ellimpiocome entitover. (Pr	Inea Mariamano	iti haq	a) a ber	eatre.
15 88 97 03 15 12 20 17 00	1 It	kt	This is the w I'm comp angely time. If thander Ju y beamt off. could item. 3 =	in heinphree. people union theory New Improved in three Josephili from the people Fig2: T word I 's a fr tr tr tr d e	terny allocate antozen Similo (ternel average weets I is d', iend aile read ntis	e hatte itte echoktras Data se	ellimpiocome entitover. (Pr	Ina Natamatu	i ti haqi	a) a ber	native
13 8 8 9 7 0 3 13 13 12 12 12 17 0 10	N IL	kt	This is the w number of the order purposed of the order of the order o	heingtree. projection for New Ingenerge life stoor. Scientifi from beingering Fig2: T word [:sea fr 'tr all 'sea 'sf	weets I is d'. iend aried aried ntis nce	a ingenetari inte na ingeneta Data se	oll/myloc.ma ntitovet.@s et	ITHE Martamanu	i ti haqi	a) a ber	notre.
IB 8 H 7 J 0 B 1 B 2 D I I I I I I I I I I I I I I I I	v st	kt	This is the w number of the order purposed of the out item.	Fig2: T tword for the star for	tenulised entreminent (tenuliserse weets I is d	anattar. http natioolekina Data se	ell/myloc.ma etitovet.@s et	ITHE Non-server	iti haqi	9) 3 Der	estre.
TB 88 97 97 97 97 97 97 97 97 97 97 97 97 97	¥ JĽ	kt avhtora Rochgitz HUSEnile Ljuncolm Veryad of [15	This is the w much finance purposed of thander purposed thander purposed t	in heingtness. people sales theory here in presign in the in presign in the ingress point Fig2: T word [:sa from tr all de sf gen cr	tenularia antonismo genelaria d'. iend aile read ntis nce t', own	anattar. http enchoolekeene ar. In gorrada Data see	e//majoc.mag ntitover. (Pr 21	ine Normanu R	iti haqi	ay in the r	eartine.
	¥ JL	kt	This is the w n	heingtree. pinoly using theory New Ingenegating theory berground Fig2: T WOrd [:sa fr tr tr all de s1 ge	weets I is d'. iend aile ntis own t',	anathe htp nathaolithna Data se	el/majoc.maj artitover.@s et	Ina Matanatu	i ti haqi	a) in the r	native
TB 88 90 10 10 12 12 17 Oc	¥ JE	Bt. avhtoria Foologisti HUGEnilic Ljuncotin Verysatal	This is the in 	Fig2: T Word I 'sa free star in a la free star i tr i tr i tr i tr i tr i tr i tr i t	weets I affective is d', iend affective read nce: t', own t', ink	anatte htp extraction Data se	(//w/oc.ma) ntitoyer. (9) et	INA Matamatu	i ti haqi	ay it ber	estine.
10 80 90 10 20 17 00	TE	Bt. avhtora Foolog m HUGEntile Ljuncolm Very set al	This is the w nTre comp angely free. If thanker Ju y beard off. tout iten. 3 =	Fig2: T word for the star for t	tenulised anticentiens (tenulareau weets I s d	anatha htp achaiteas Data se	(//m/ioc.ma) antitover. (Ba	Ine	i ti haqi	कु त जेव	sertine.
TD SH TD SB TD SD TT OL	Y IL	Bt. avitto na Fooling dti HUGE nil o I junt of m Very set of C 1.5	This is the w 	Fig2: T word for the star for the star for the star word for the star for the star for for the star for for for for for for for for for fo	weets I s d'. iend aile ntis nce t', ink t', ink	anatta: htp extraoletere a l'egeratta Data se	(//m/oc.ma) atitover. Ba		i ti haqi	a de c	eartine.
TB SB VU SB SB SD SD SC CL	Y IL	bt authorna Foologen HUEenolo juntotm (15 [15]	This is the w number of the set o	Fig2: T Word I 'sa fr 'tr 'tr 's1 'ge 'ch 'ch 'ch	weets I is d', iend aile read ntis t', own t', ink' eati ch',	anatta htp achaitean Data se	(Vrayfac.ma) at it over. (P et		i ti haqi	a de c	estine.
TB SB VD OB TD ID ID ID ID ID	V JL	Bt; auhtorna HSErolo Justodin [ 1.5 [ 1.5	This is the w much more than any system of the water is the water of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the system of the	Fig2: T Word I 'sa from the second from the se	weets I aile aile read ntis nce t', ink t', ch', nded	anatta htp actoristicas Data se	ntitore: @		i ti haqi	ay a dec	
TD SH U U U U U U U U U U U U U U	v J L	kt; auhto na Feebig st. HUEndic Ljancetti (1.5 (1.5)	This is the so 	Fig2: T word for the star for t	tenulised mbomismus (brelareau weets I s d'. iend aile read ntis nce' t', ink' eati rry' ch', nded ifor	anattar, htp arattar, htp arattar aktore Data se	ntitoer (hydocae) et		i ti haqi	ay a dec	nutin
	v at	kt authorna HSErold Ljustoutm Very sold [ 1.5	This is the w I no com- angely fine. if thur the pu- point of the final of the final is a set of the final of the final of the final is a set of the final of the final of the final of the final is a set of the final	Fig2: T word for both the second second second second for the second sec	weets I s d', iend ai ai ai ai ai ai ai ai ai ai ai ai ai	anatta: htp extraoletras a l'egoratta Data se	ntitoer (f	Ine Automatico	i ti haqi	a dec	eartine
TD SH VU OB ID ID ID ID ID ID ID	v at	kt; auhtoma Hotenile Junotin [15]	This is the so I no com angely fine. I france pu to contian. ] : ] :	Fig2: T word from the second from the second for the second for the second for the second for th	weets I iendierue d'. iendierue d'. iendierue read nce t'. own t'. ink eatii rcy'. nded ifor	anatta htp achaitean Data se	nitore. (h	Ine Auranan R	i ti haqi	gy i the c	ertre
TB SB VU OB 22 22 Tr OL	¥ JE	kt; auhto na Feebig at HUEnild Ljancahn (115	This is the s much more regulation of the state of the flavor of the state of the s	Fig2: T Word Fig2: T Word I sa from bergener tro al sf sf sf sf sf sf sf sf sf sf sf sf sf	weets I aile aile read ntis nce t', own t', ink t', ch', nded ifor day ss,	anatta htp acto oktras Data se	ilmiocm, nitoer. B	ine Norman R	ti hap	9) i be	natine.
n D Su NU D D D D D D D D D D D D D D D D D D	¥ st	kt auhto na Feebig st. HUEndo Ljancetni Wery set at E 1 5	This is the so much so of the source of the source of the source of the source of the source of the source of the source of the source of the source of the source of the source of the source of the source of the source of the source of the source of the source of the	Fig2: T word for the figure from the figure fr	weets I ailed aile read ntis own t', nded ifor ass', read	anatta htp actorialization Data se	ritore P		ti hap	up, a ben	

Fig3: Extracted features for classification of data set

After extracting important features from tweets, assign polarity to each feature in three levels. These three level sentiment polarities are positive, negative, neutral. Fig4 shows performance of Naïve bayes, SVM, decision

tree algorithms

	Accuracy	Precision	Recall	F-score
Naïve Bayes	0.56128	0.4129	0.25169	0.22907
Support vector machine	0.8006	0.8459	0.62999	0.68586
Decision Tree	0.8961	0.8713	0.8493	0.8596

Fig4: Comparison table

## IV CONCLUSION

We presented results for sentiment analysis on Twitter. Twitter is a demandable micro blogging service which has been built to discover what is happening at any moment of time and anywhere in the world. By using machine learning algorithms i.e naïve bayes, decision tree, SVM, we conclude that it is easier to classify the tweets and more we improve the training data set and more we can get accurate results. We can also implement features like emoticons, neutralization, negation handling and capitalization or internationalization as they have recently become a huge part of the internet. This paper also presents empirical comparison of classification algorithms in which decision tree algorithm is highest accuracy in comparison of all three algorithms considered in this study.

#### REFERENCES

- Haddi E, Liu X, Shi Y. The role of text pre-processing in sentiment analysis. Elsevier Procedia Computer Science. 2013; 17:26–32.Available from: www.sciencedirect.com
- [2] Ceci M, Loglisci C, Macchia L. Ranking sentences for key phrase extraction: A relational data mining approach. Elsevier Procedia Computer Science. 2014; 38:52–9. Available from: www.sciencedirect.com C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.
- [3] Thamizharasi P, Sathiyavathi R. An approach to product rating based on aspect ranking in opinion mining. Indian Journal of Science and Technology. 2016; 9(14):1-6. doi:10.17485/ijst/2016/v9i14/89604



## International Journal of Science, Engineering and Management (IJSEM) Vol 3, Issue 4, April 2018

- [4] Singh A, Ullah ME. Aspect based sentiment analysis. Available from: http://home.iitk.ac.in/~enayat/files/absa\_report.pdf
- [5] Kiruthika M, Sanjana Woonna, Priyanka Giri-'Sentiment Analysis of Twitter data'- International Journal of Innovations in Engineering and Technology– Vol.No.6,Issue No.4, April 2016.
- [6] Durgesh Patel, Sakshi Saxena, Toran Verma-'Sentiment Analysis using Maximum Entropy Algorithm in Big Data' – International Journal of Innovative Research in Science, Engineering and Technology–Vol No.5,Issue No.5,May 2016.
- [7] Bholane Savita D., Prof.Deipali Gore 'Sentiment Analysis on Twitter Data Using Support Vector Machine'– International Journal of Computer Science Trends and Technology (IJCST) Volume 4, Issue 3, June 2016.
- [8] V.Lakshmi, K.Harika, H.Bavishya, Ch.SriHarsha 'Sentiment Analysis on Twitter Data'– International Research Journal of Engineering and Technology (IRJET) Volume 4, Issue 2, February 2017.