A Comparative Analysis of Lexical/NLP Method with WEKA’s Bayes Classifier

Karuna Gull
Dept. of Computer Science,
K.L.E.I.T., V.T.U.,
Hubli, India
e-mail: karunagull74@gmail.com

Sudip Padhye
Dept. of Computer Science,
K.L.E.I.T. V.T.U.,
Hubli, India
e-mail: sudip.padhye@gmail.com

Dr. Sandeep Sharma
Dept. of Computer Science,
C.S.J.M.University
Kanpur, UP
e-mail: karunagull123@gmail.com

Dr. Subodh Jain
Dept. of Computer Science,
SVN University,
Sagar, MP, India
e-mail: karuna7674@gmail.com

Abstract**—** Various websites are available as source of microblogs. This is due to nature of microblogs on which people post real time messages about their attitudes on a various topics, talk about present issues, criticize, and articulate positive or negative sentiment for products they use in daily life. That’s why, manufacturing companies of such products have started to take these microblogs to get a sense of general sentiment for their product. Reply can be given by the companies on microblogs for the reactions of the users. Thus challenge is to build a technique to detect and summarize an overall sentiment. The proposed methodology examines sentiments on Twitter data contextually. Sentiment Analysis is the major aspect of present day NLP. Also, Twitter has emerged as the most important data source for present day NLP. In the work carried out, tweets are extracted from Twitter using Twitter API after authentication, a fine pre-processing is dealt and provided for further processing. Later, tag each word with their respective parts of speech using Part-Of-Speech (POS) tagger. SentiWordNet, WordNet and NLP weight assignment policies are used to assign weights and provide results. The analysis of same data set is also done with Naïve Bayes classifier using WEKA - the data mining tool. Then results of both – the proposed method and Naïve Bayes are compared. (Then finally comparison between the results of proposed method with Naïve Bayes classier is done.) The investigation proved that our method i.e. NLP technique works better than that of Naïve Bayes Classifier. And this study also proves that the training set to the classier matters a lot in Machine Learning - “Expected output can be accurate if and only if the training of a classifier is better”.

Keywords- Natural Language Processing, Sentiment Analysis, Attitude Scrutiny, SentiWordNet, WordNet, Data Pre-processing, POS Tagging, Contextual Sentiment, WEKA Tool, Naïve Bayes Classifier, Twitter

*****

I. INTRODUCTION

A. *Introduction to Social Networking Sites (SNS) or Microblogs*

Micro-blog data like Twitter, on which users post real time reactions to and opinions about “everything”, poses newer and different challenges. For our project, we have chosen Twitter as source of user reviews because Twitter has the highest number of active users and members worldwide. Also, it has limit on message length and thus storing each tweet becomes easy.

B. *Introduction to Attitude Scrutiny*

Companies discovered that the fame of social media creates a realistic environment for the tone of the customers, Facebook and Twitter’s combined membership is over 1.5 billion people worldwide. Internet users also blog and add comments to film, article reviews on sites. This creates a foremost opportunity for companies to take an advantage of customer opinions spread on sites to better understand what people are saying on a topic or about a company or a person. Opinion Analysis is one of the most active research areas in Data mining, Web mining, and Text mining. In fact, this research has moved outside of computer science to the management/social sciences. The growing importance of sentiment analysis coincides with the growth of social media such as Twitter, social networks, reviews, blogs, and forum discussions etc. Thus now we have a huge volume of opinionated data recorded in digital form for study. Unstructured data refers to information that doesn’t have a pre-defined data archetype. Unstructured information is typically textual data, but may also contain numerical data, and factual details. This results in data that is obscure, irregular and ambiguous, thus making it difficult to analyze using conventional computing means. Much of the data in the web, in the form of blogs, news, social media platforms is unstructured. But they serve as a potential vast source of information, if processed efficiently. In this project, we tried to collect specified no. of reviews on specified topic, pre-processed them to clean the unwanted data and processed those using NLP techniques.

C. Introduction to Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve: natural language understanding, enabling computers to derive meaning from human or natural language input; and others involve natural language generation. Modern NLP algorithms are based on machine learning, especially statistical machine learning. The paradigm of machine learning is different from that of most prior attempts at language processing. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. The machine-learning paradigm calls instead for using general learning algorithms — often, although not always, grounded in statistical inference — to automatically learn such rules through the analysis of large corpora of typical real-world examples. A corpus (plural, "corpora") is a set of documents (or sometimes, individual sentences) that have been hand-annotated with the correct values to be learned. Many different classes of machine learning algorithms have been applied to NLP tasks. These algorithms take as input a large set of "features" that are generated from the input data. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

II. Liturature Review

Twitter has become an exclusive SNS that is chosen for every updates over the world. It is a place where people gather and confer their interests. And analyzing the comments, shares, favorites of the users help to realize the influential users and track their interests. As per the related work done by us, the summary of papers referred by us are as below:

Abraham Lincoln, 16th US President, quoted that “Public sentiment is everything. With public sentiment, nothing can fail. Without it, nothing can succeed.”

Apoorv Agarwal [1] examined sentiment analysis on Twitter data. The contributions of the paper are: (1) Introduced POS-specific prior polarity features. (2) Explored the use of a tree kernel to obviate the need for tedious feature engineering. The new features (in conjunction with previously proposed features) and the tree kernel perform approximately at the same level, both outperforming the state-of-the-art baseline.

Basant Agarwal [2] proposed a novel sentiment analysis model based on common-sense knowledge extracted from ConceptNet based ontology and context information. ConceptNet based ontology is used to determine the domain specific concepts which in turn produced the domain specific important features. Further, the polarities of the extracted concepts are determined using the contextual polarity lexicon which we developed by considering the context information of a word.

Nikhil R. [3] specifies unstructured data refers to information that doesn’t have a pre-defined data archetype. Unstructured information is typically textual data, but may also contain numerical data, and factual details. This results in data that is obscure, irregular and ambiguous, thus making it difficult to analyze using conventional computing means. Much of the data in the web, in the form of blogs, news, social media platforms is unstructured. But they serve as a potential vast source of information, if processed efficiently. In this paper, the basics of harnessing unstructured data from the web and the techniques to process it are discussed.

The survey made, clearly specified about the unstructured nature of data. Hence there was a need for fine tuning the data extracted. As, the input is from social networking site the vague data needs to be preprocessed properly. In the proposed methodology, the phase of preprocessing was neatly undertaken because this phase helps to train the system. And training data is directly proportional to the final result. Considering these observation the training was rigorously done and came up atlast with a result that worked better when compared with the Naïve Classifier using a WEKA suite. The comparison and graphical analysis were made on same data set for both the algorithms.

III. Methodology

A. Architecture Diagram

Proposed system reveals the process of Sentiment Analysis taking twitter as a case study. The proposed system emphasize on classification of tweets using NLP algorithm, which in turn helps in predicting the opinions of people. Sentiment Analysis of customers opinions assist in gaining the wider community attitudes behind certain subject. This benefits the organization to promote and advance their brand. The layout of the proposed model reveals the ordered steps to meet the objective of the problem definition. This project contributes to the field of sentiment analysis, which aims to extract attitudes, opinions from tweets.

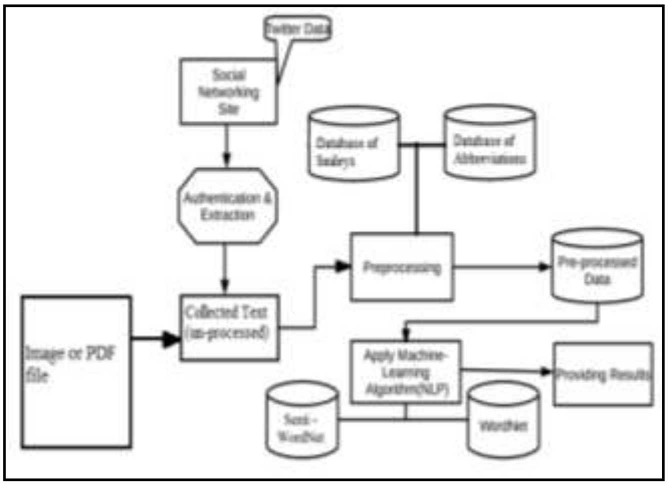

Fig 1: Architecture Diagram

The architecture shown in the Fig. 1 works as follows:

1. Authenticate the app to social networking sites – Twitter using available APIs.
2. Retrieve the data through the access tokens specified for each user.
3. Reduce the set of variables (features) by tokenizing and cleansing the tweets extracted.
4. Apply the Natural Language Processing technique to classify the pre-processed tweets.
5. Determine the polarity of sentiment expressed towards a company, product, individual or any given topic automatically.
6. Generate a summary of the sentiment analysis (positive, negative or neutral) of the collected tweets.
7. The result of our work is compared with the Weka Tool's Naïve Bayes analysis, which show that how our work is better than the traditional Naïve Bayes approach.

B. *Elaboration of Steps*

The steps of proposed system are:

1) Registration of application [6]

i. Authentication Process

User has to authenticate his/her application to Social Networking Site (Twitter) via his/her Twitter account. This process issues keys like Consumer key, token access key etc. to the end user.

ii. Extraction Process

Using the authentication keys, user is enabled to extract the stats (Tweets, Re-tweets, Timeline etc.) of his/her account. In Twitter, individual message implies the stats info of a user. Instead of tweets, text can also be extracted from image or pdf file using API.

2) Data cleaning and Pre-processing

Removal of unwanted characters (stop words, punctuations, special characters, links, @ etc.) from the statuses collected from Social Networking Sites. Take only text part of stats information and user names for further processing. The text part is tokenized into set of tokens.

i. Translation

Translation is carried out using Yandex Translator API which requires an API Key. This key is obtained by registering with the Yandex website. Translation facility helps in translating the non-English text input into English text.

ii. Regular expressions

Java supports regular expressions through Pattern and Matcher objects. Regular expressions are used for various purposes such removal of @username, URLs, hashes in hash tags, punctuations, multiple spaces and replacement of multiple alphabets with single alphabet.

iii. Resolving smileys and abbreviations

Smileys and abbreviations are nowadays used very widely. They are good source of sentiments. Thus it is not advisable to remove them. The smileys and abbreviations are resolved with their corresponding pair-word. The smileys are replaced with the corresponding expression of the smileys in textual format whereas the abbreviations are replaced with their equivalent expanded form.

iv. Spelling corrector

Spelling corrector is the technique which is popularly used in SEO (Search Engine Optimization). It takes care of the spelling mistakes caused while typing by users. Using the English dictionary, it corrects accordingly to the closely matching English spelling.

v. Lemmatization

Lemmatization is the technique used to find root words of the given word. This is done using dictionary which resolves the word in any form to its root word equivalent. Unlike Stemming, Lemmatization does not remove prefixes and suffixes of the word to obtain root word. Thus Lemmatization is more efficient than Stemming.

3) Assignment of weights using NLP

i. POS Tagging

POS Tagger tags Part-Of-Speech (POS) to each word in the sentence. This helps in for NLP (Natural Language Processing) analysis. Based on the POS of each token, weights are assigned and later total weight is calculated.

ii. Filtering

Data filtering is also performed on the collected data by removing unwanted words that don't contribute to the actual sentiment of the data. These unwanted words are usually single character words, articles, preposition, conjunction or pronouns. Thus it is advisable to use POS-tagging technique to eliminate such words instead of eliminating individually by creating database and using Regular expressions, because it is very cumbersome to match each and also time-consuming.

iii. Tokenization

Tokenization is done word-by-word so that the each word can be assigned with weights to compute the overall sentiment. Tokenization is done by taking a space as a delimiter.

iv. Assignment of weights using NLP

Based on the POS Tagging, weights are assigned to each token. For Nouns, Pronouns, Preposition, Conjunction and Interjection fixed weights are assigned. Fixed weights for noun, pronoun and interjection are 0.1 & that for preposition and conjunction is 0.2. For Verbs, Adverbs and Adjectives weights are assigned based on the polarity using SentiWordNet. If the word is not present in SentiWordNet, then its synonym is found out using WordNet and that synonym is checked into the SentiWordNet. If the synonym is present in the SentiWordNet then its corresponding weight is taken as the weight of the original word. On the other hand, if any of the synonyms are not present in the SentiWordNet then fixed weights are assigned. Fixed weight for verb is 0.8 and that for adjective and adverb is 0.5.

Clearing previously stored tweets and writing pre-processed & analyzed tweets to file based on sentiment.

4) Consolidating Discovered Knowledge

Tweets are segregated and stored so that they can be given as input to weka using TextDirectoryLoader for Naïve Bayes

Classification algorithm. Using this, we can get the accuracy of the algorithm and can be compared easily with our project accuracy. It also depicts certain limitations of Naïve Bayes algorithm. Displaying results in comparison with the WEKA tool, or using them inside the proposed system in the form of APIs.

a) Open Weka in Explorer mode.
b) Then in preprocessing menu, Open file Test data using TextDirectoryLoader.
c) Apply StringToWordVector filter present in Unsupervised-attribute folder (change the parameters if required) andthen Save this file.
d) Similarly open the Training data using TextDirectoryLoader.
e) Similarly, apply StringToWordVector filter present in Unsupervised-attribute folder (change the parameters if required).
f) Then in Classify menu, select Naïve Bayes Classifier. Provide the previously saved Test Data file as Supplied test set.
g) Then click on Start to get the results.
h) After analysis, we get the results as correctly classified, incorrectly classified etc.

IV. Dynamics of the Work Carried

The Overview of the work carried out is shown in Fig .2. and it is working as follows:

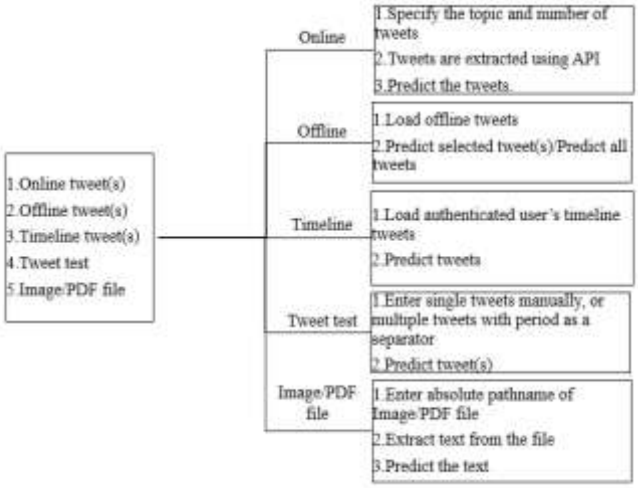
1.Online tweet(s)
2.Offline tweet(s)
3.Timeline tweet(s)
4.Tweet test
5.Image/PDF file
Online
1.Specify the topic and number of tweets
2.Tweets are extracted using API
3.Predict the tweets
Offline
1.Load offline tweets
2.Predict selected tweet(s)/Predict all tweets
Timeline
1.Load authenticated user's timeline tweets
2.Predict tweets
Tweet test
1.Enter single tweets manually, or multiple tweets with period as a separator
2.Predict tweet(s)
Image/PDF file
1.Enter absolute pathname of Image/PDF file
2.Extract text from the file
3.Predict the text

Fig. 2 : Overview of complete system

1. In the Online mode of analyzing the tweets, (i)the topic related to what exactly, the tweets are to be retrieved needs to be specified, (ii)specify the number of tweets to be extracted, (iii)tweets are extracted using twitter API's, (iv)extracted tweets are subjected to analysis.
2. In the Offline mode of analyzing the tweets, (i)the tweets stored in the local database are extracted, (ii)we can subject selected tweets out of the extracted tweets for analysis or all the tweets can be subjected for analysis.
3. In the Timeline mode of analyzing the tweets, the authenticated users timeline tweets are extracted from random topics and then the extracted tweets are subjected to analysis
4. In the Tweet test mode of analyzing the tweets, user can manually enter one or more tweets which are subjected to analysis.

V. Experimental Outcomes

Table 1. Sample Tweets extracted from Twitter

RT @zaynskordei: the seanianastans, would blame ricky for global warming if they could, always posting videos of ariana and seans perform…
RT @maxplanckpress: Global warming will lead to increase in #climate refugees from Middle East & North Africa, study suggests: https://t.co…
RT @larissawaters: You cant be serious about addressing global warming if you approve #coal mines & adopt Liberal cuts to clean energy htt…
#ZEROGOD #BENEATHTHEROCK Greatest threats R not global warming, weapons mass destruction, war, pestilence ISIS. Its remembering passwords!
RT @ZacEfron: Found this amazing beast on @UniStudios #backlot. I love you. So sorry ure going extinct because of global warming. https://…
RT @premierleague: Congratulations @PetrCech - winner of the 2015/16 Barclays Golden Glove! https://t.co/va6mwx5Rst
Yippie bought a laptop. Thank you Mom and Dad for the best gift ever.#veryhappy. #enjoying #party
: :

Table 1. shows the sample tweets extracted from Twitter account. After applying preprocessing on sample tweets , the weights evaluated by proposed method uing the NLP Technique and the final result is shown in Table 2.

Table 2. Pre-processed and Analyzed tweets

Pre-processed Sample tweets	Tweet weights	Result
re tweet the Seana stuns would blame Ricky for global warming if they could always posting videos of Ariana and Sean s perform	-2.703695	Neg
re tweet global warming will lead to increase in climate refugees from middle east Nore teeth Africa study suggests suggests	-4.368751	Neg
re tweet you cant be serious about addressing global warming if you approve coal mines adopt liberal cuts to clean energy	-0.168248	Neu
zero god zero god greatest threats r not global warming weapons mass destruction war pestilence Isis its remembering passwords	-4.205683	Neg
re tweet found this amazing beast on backlog i love you so sorry you re going extinct because of global warming warming	-4.659121	Neg
be tweet congratulation winner of the barclay 's golden glove	2.318728	Pos
yippie buy a laptop thank you mom and dad for the best gift ever very happy enjoy party	4.459526	Pos
:		

VI. OTHER OBSERVATIONS

Software testing is a process of verifying and validating that a software application or program. It meets the business and technical requirements that guided its design and development, and also works as expected.

In Table.3 different test cases are considered with few sample tweets as an input and based on it the observed output is written. Table.4 shows the sample inputs along with expected and observed output.

Table 3:Expected Output for different Test cases

Case	Test Case	Input	Expected output	Observed output
1.	Fetching tweets	Access token, Access secret, Consumer Key, Consumer Secret, Topic and no of tweets<=300 with Internet Connectivity	Specified no. of tweets on specified topic	Specified no. of tweets on specified topic
2.	Fetching tweets	Access token, Access secret, Consumer Key, Consumer Secret, Topic and no of tweets>300 with Internet Connectivity	Specified no. of tweets on specified topic	Maximum limit is 300
3.	Fetching tweets	Access token, Access secret, Consumer Key, Consumer Secret, Topic and no of tweets<=300 with no Internet Connectivity	Specified no. of tweets on specified topic	Connection Exception
4.	Translation	I am very happy. It was very good day!	No Translation required	No Translation required
5.	Translation	Estoy muy feliz. que era muy buen día !	I am very happy. It was very good day!	I am very happy. It was very good day!
6.	Translation	मैंबहुतख़ुशहूँ।यहबहुतअच्छादिनथा!	I am very happy. It was very good day!	Language not supported
7.	Translation	Estoy muy feliz. que era muy buen día ! (No Internet Connectivity)	I am very happy. It was very good day!	No Internet Connectivity
8.	Regular expressions	@timjones_17 – It was an average movie. #bored http://www.movies.com☹	it was an average movie bored	it was an average movie bored
9.	Spelling	There was	there was	there was

	corrector	smoeniose in the room.	some noise in the room	some noise in the room
10.	Lemmatization	I was playing football in rain	I am play football in rain	I am play football in rain
11.	POS Tagger	I am very happy. It was very good day!	i_LSbe_VBvery_RB happy_JJit_PRPbe_VBvery_RB good_JJday_NN	i_LSbe_VBvery_RBhappy_JJit_PRPbe_VBvery_RBgood_JJday_NN

Table 4: System - testing Test Case table

Case No.	Test Case	Input	Expected output	Observed output
1.	For English tweet	I am very happy. It was very good day! #happy	Positive	Positive
2.	For Non-English tweet	Estoy muy feliz. que era muy buen día !	Positive	Positive (Translated text: I am very happy. It was very good day!)
3.	For English tweet	He is a good man !	Neutral	Neutral
4.	For Non-English tweet	@Isis_2210 @Peitinhos2 @GatasNudes @NudesArtisticoagorana DM	Neutral	Neutral (Translated text: @Isis_2210 @Peitinhos2 @GatasNudes @NudesArtistico now in direct message)
5.	For English tweet	@MarkSchwab @Clevetroit the Geneva convention prohibits feeding Cincitucky chili to prisoners. ISIS even finds that practice cruel.	Negative	Negative

VII. Comparison Study Between Attitude Scrutiny Against Weka Tool's Naïve Bayes

Composed data for our work will be now given as input for the WEKA tool. Collection of the analysed results from WEKA tool's Naïve-Bayes implementation for the given data set is done which is shown in fig. 3.

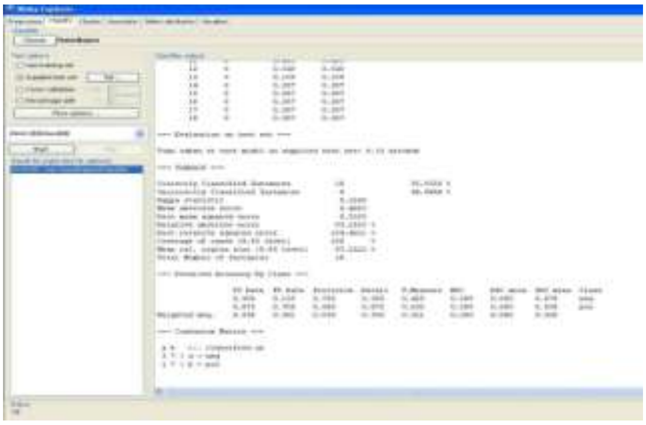

Fig. 3: Analysed results from Weka Tool's Naïve Bayes implementation for the same data set.

Table 5. shows True Positive, False Negative, False Positive and True Negative values for WEKA's Naïve Bayes implementation and our work Attitude Scrutiny Application.

Table 5. TP, FN,FP and TN values for NB and NLP algorithms

	Naïve Bayes	NLP
True Positives	3	8.25
False Negatives	7	0.75
False Positives	1	0.75
True Negatives	7	8.25

Fig. 4. shows graph of TP, FN, FP and TN values of both algorithm.

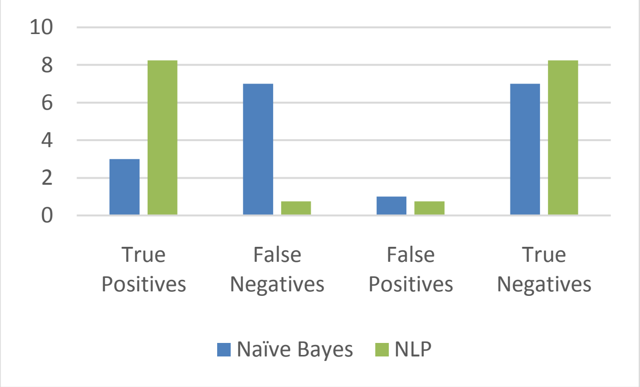

Fig. 4: Graph showing comparison of TP, FN, FP and TN values of Attitude Scrutiny (NLP) against Weka Tool's Naïve Bayes

VIII. Conclusion and Future Scope

In Social sites, beyond simple like's, thumbs up, we can check out qualitatively into people's views, motivation and opinions at quantitative level. Project tried to meet this by collecting the tweets from twitter. In this proposed work, we tried to analyze the data based on NLP by assigning weights to each token in the sentence. Later, the total weight of the sentence is calculated and tested for the range in which the values lie. Based on this, the sentences are classified into three classes- Positive, Negative and Neutral. Finally, the results are provided in the form of pie charts.

The result provided by the application is compared as shown in fig. 5 with the results given by the WEKA tool's Naïve-Bayes implementation. Our application gives better results than WEKA's Naïve-Bayes analysis because Naïve-Bayes algorithm requires lot of training set which consumes more memory, also our application is based on POS tagging whereas, Naïve-Bayes is based on conditional probability. If the word is not present in the Naïve-Bayes training set then it will ignore whereas, our application assigns polarized weights based on the POS tagging.

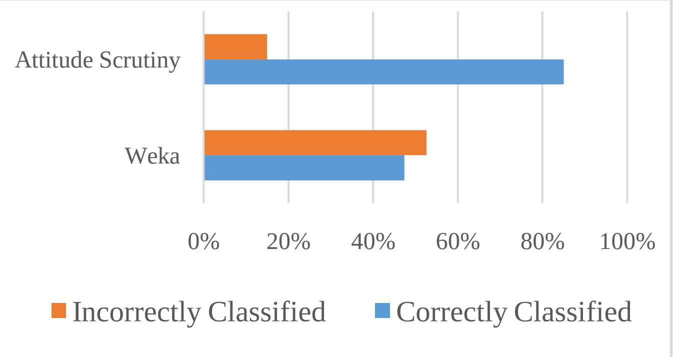

Fig. 5: Comparison of Attitude Scrutiny Application against Weka Tool's Naïve Bayes

The work focused on tweets of twitter, further this can be extended on developing a system that extracts the comments of customer from different sites (Data warehouse) and analyze it by using different algorithm and help to know the attitude of customers on the product or company or individual. This helps the organizations to know the flaws if present related to their product and will help to know the alterations that customer expects from them. As a future scope, the model can be extended further to provide different databases for different contexts and also web based GUI which can be used to provide services to several number of people over Internet. Also, the accuracy can be improved by removing stop-words from the collected tweets. Furthermore, the project can be extended to accept image and voice as an input & extracting the text in them for further analysis. Retrieving unique tweets and processing of Big-data using MapReduce algorithm can be added for offline processing.

REFERRENCES

[1] Sentiment Analysis of Twitter-Data by Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow & Rebecca Passonneau (Department Of Computer Science, Columbia University, NY, USA). July-2011.

[2] A Research Article on Sentiment Analysis using Common Sense & Context Information by Basant Agarwal, Namita Mittal, Pooja Bansal & Sonal Garg(Department Of Computer Science & Engineering - SKIT & MNIT, Jaipur) February-2015@HIDAWI Publishing Corporation.

[3] A Survey on Text-Mining & Sentiment Analysis for Unstructured Web Data by Nikhil R, Nikhil Tikoo, Sukrit Kurle, Hari Sravan Pisupati & Dr.Prasad G.R.(Department Of Computer Science & Engineering, BMS College Of Engineering, Bangalore)@Journal Of Emerging Technologies & Innovative Research(JETIR-Vol.2 Issue 4).April-2015.

[4] Baccianella S, Esuli A and Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC 2010 May 17 (Vol. 10, pp. 2200-2204).

[5] Esuli A, Sebastiani F. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC 2006 May 22 (Vol. 6, pp. 417-422).

[6] Narashima S. Purohit, Meghana Bhat, Akshata B. Angadi, Karuna C. Gull, (2015) "Crawling through Web to Extract the Data from Social Networking Site – Twitter", IEEE National Conference on Parallel Computing Technologies –PARCOMPUTECH,India,2015. doi:10.1109/PARCOMPTECH.2015.7084522, ISBN:978-1-4799-6916-6,pp.1-6. Available:http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7084522