IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 82-89
www.iosrjournals.org

Sentiment Analysis of English and Tamil Tweets using Path Length Similarity based Word Sense Disambiguation

Kausikaa.N[1], V.Uma[2]
[1](PG student,Department of Computer Science,Pondicherry University,India)
[2](Assistant Professor,Department of Computer Science,Pondicherry University,India)

Abstract: *In social media, users have the privilege of connecting with people and extensively communicate, share information, discuss topics of recent trends. Friendster, LinkedIn, Instagram, Twitter are some media through which users can perform the activities as mentioned earlier. Twitter is a well-known microblog which allows user to express their opinion or sentiment in the form of tweets within maximum length of 140 characters. The sentiment of the user can be analyzed and interpreted using the concept called Sentiment Analysis (SA).Twitter is widely used* in *almost all parts of the world thus ensures the presence of multilingual tweets in expressing their sentiments. In this paper, Sentiment Analysis is employed to determine the polarity of English and Tamil tweets, which is therefore bilingual and it is further subjected to word sense disambiguation to figure out the contextual usage and further the sentiment of the words are classified using the Support Vector Machine which derives the polarity value of the words as positive, negative and neutral.*

Keywords-*Bilingual, Path Length Similarity, Sentiment Analysis, Support Vector Machine, Word Sense Disambiguation.*

I Introduction

Microblogging is a mode through which the users of social media connect with each other in order to communicate the essential information among them in the form of shorter sentences namely instant messages [1].Twitter is one such well known example of microblogging for sharing information with other people through instant messages called 'tweets'. In Twitter, one user has the privilege of following the other and gets instantly updated about his/her expression of thoughts or opinions on a day-to-day basis. The tweets posted by one user can earn an extensive response in terms of forwarding (Re-tweet) the content, by using '@' for identifying the user and '#' for representing a topic, ranging from daily life to current events[2]. An opinion shared in twitter can be positive or negative, irrespective of the topic or trend, where the nature of the opinion can be further derived using sentiment analysis.

Sentiment Analysis (SA) can be defined as the process of learning users' attitude, opinions, and emotions subjected towards an item, which can thereby imply to individuals, events or topics of latest trend. Sentiment Analysis can be classified under three levels namely, document level, sentence level and aspect level [3]. As the Twitter platform uses tweets to express opinions in the form of sentences, sentence level sentiment analysis is employed for analyzing the sentiments. The Sentence level sentiment analysis, identifies whether the sentences are subjective or objective. If subjective, sentence level SA determines whether the sentence expresses positive or negative opinion.

When sentiment analysis is employed in Twitter, it can further be subjected to three different categories such as polarity, emoticon, and strength. Polarity based sentiment analysis extracts the information from the sentences and returns the possible values for polarity as positive, negative or neutral. Emoticon based sentiment analysis extracts information from the emoticons expressed in the tweets, which thereby have predefined values for each and every facial expressions. Strength based sentiment analysis extracts information from the text and returns numerical values for different opinion with respect to the intensity of opinion in the text [4].

Twitter uses different languages for expressing opinions which changes with respect to different geographical regions of the world. So there arises the need for handling multiple languages in the platform. In this context, Tamil is one such language which is restricted to a limited number of users, hence the focus is laid on the sentiment analysis of bilingual (both English and Tamil) tweets. Since Tweets are in informal in nature, the contextual meaning of the sentence varies often. In order to overcome this limitation, the Word Sense Disambiguation is employed.

Word Sense Disambiguation (WSD) is the ability to disambiguate a word that can have many senses based on itsusage context. Each word may be represented in one (monosemous) or more sense (polysemous)[5]. WSD can be achieved through engaging knowledge based method, using wordnet, which is a lexical database. In Wordnet, the words are grouped into synsets representing the meaning of the words and hence semantic similarity between two synset is computed by measuring the path-length based similarity.

SVM is proven best for sentiment based text classification, which has been trained with data to

determine the linear seperators in search space [6]. The different classes can be separated by seeking hyper-plane represented by vectors with maximum margin. The performance of SVM is evaluated using measures of precision and recall and as a result the final output of SVM include binary values for classes. The contribution of this work is manifold,

1.Translation of Tamil Tweets into English Tweets.
2.Finding out the semantic similarity using path-length similarity.
3.Classification of Sentiments using Support Vector Machine.

The rest of the paper is organized as follows, Section 2 discusses the related work; section 3 describes the proposed framework; section 4 presents the experimental evaluation and finally section 5 provides conclusion.

II Related Work

Jantima Polpinij (2014) contributed towards the problem of analyzing the feedback in multiple languages and proposed the classification of multilingual sentiments. The processing of framework includes two steps where first step deals with classification of reviews into two language classes and second step focuses on classifying textual dataset into positive and negative sentiments. The first step is processed with the help of lingual separation of review by employing character analysis. The second step uses Latent Semantic Indexing to group the similar words from group of documents based on word concept.

Zheng Lin, Xiaolong Jin, Xueke Xu et.al (2014) focuses on two problems which are, excessive dependence of contextual tools or resources which is not suitable in the case of minority languages and sentiment polarity of some parts of text differs with respect to the overall sentiment polarity.The solution to this problem is obtained by estimating the sentiment polarity of unlabeled data.The extraction of opinion lexicon is based on the presence of seed words.The extracted opinion lexicon is further used to extract the key sentence which is employed to capture the absolute sentiment of the review.The semi supervised learning method is opted because it combines both supervised and unsupervised approach in order to classify the sentiment.The semi supervised learning method fails to result in effective performance.

The multilingual sentiment analysis system is developed by using two different languages such as English and Spanish. The hybrid feature and data translated by the machine can extensively help in recognizing better befitting features. The obtained feature is used in the process of classification of sentiment which increases the precision of sentiment analysis. The sentiment dictionary is employed and is used to transform sentiment presence words which is contained in tweets into mismatched labels for describing their polarity. The ability of sentiment classification is tested by employing four separate pairs of linear classifiers namely objective vs subjective, positive vs negative vs neutral, positive vs very positive, negative vs very negative. The observed performance is worse to the core because of the use of linguistic processing (Alexandra Balahur, José M. Perea-Orteg 2014).

The Topic Adaptive Sentiment Classification(TASC) is built to overcome the issue of adapting to unpredictable topics and labelled data which forms extremely sparse text.The first step is to build a classifier which is trained on common features and then the data from various topics is mixed.The feature of the text includes temporal, emoticon, and punctuation present in text are used to build the classifier.The experiment is carried through by employing six topics from a list of published tweets.The multiclass SVM is modelled to adapt the adaptive nature of the system.TASC is compared with other supervised classifier to evaluate the performance of the system.As a result TASC performs well in adapting to unpredicted data than other classifiers (Shenghua Liu, Xueqi Cheng, Fuxin Li 2015).

Samir E.Abdelrahman, Ebtsam Abdelhakam Sayed, Reem Bahga(2014) presented about the integration of two SentiWordNet based on sentence level polarity classification. First step is employed to find the score of foremost word sense using word sense disambiguation algorithm. The second step is to collect list of non-zero values for prior sentiment words of subjectivity bound lexicons. The experiment was conducted based on different lexical resource which is merged with feature selection to train the classifier. The system fails to use best feature selection algorithm which results in improper classification of polarity of lexicon.

The phrase level sentiment analysis for Spanish is presented with the help of knowledge based word sense disambiguation. Spanish WordNet and WordNet-Pr are used as knowledge source and fuzzy clustering algorithm is used to identify whether the word expresses subjectivity and objectivity. The annotated semantic resource is used to determine the polarity value. The SemCor corpus is used to estimate the attributes which is chosen for sentiment analysis and Rule based classifier is built to detect the sentiments .This actually turns out badly when it is used for coarse grain method (Marco Antonio Sobrevilla Cabezudo,Nora La Serna Palomino,Rolando Magui˜na Perez,2015).

Umar Farooq, Tej Prasad Dhamala, Antoine Nongaillard et.al (2015) contributed towards improvising

the method of disambiguation of words which indirectly boosts up the performance of sentiment analysis. The feature level sentiment analysis is used to yield the synopsis of opinion. The heuristic method is employed to detect the text which expresses sentiment. The boundaries are clearly determined to identify the feature which is used to extract the opinion of text. The content matching mechanism is used to obtain the polarity of similar context from lexicon. The lexicon dictionary is built which supports both context matching mechanism and word sense disambiguation techniques with respect to specific domain.

TheresaWilson, JanyceWiebe, Paul Hoffmann (2009) present new point of view towards sentence level sentiment analysis which initially predicts whether the statement is neutral or polar. The neutral expression generally includes facts whereas the polar includes opinion. The subjective indicators are words and phrases that may be used to have certain subjective usages. The lexicon which are subjective in most context is noted as strongly subjective(strongsubj) and lexicons with limited subjective usage is noted as weakly subjective(weaksubj).The classification of polarity includes positive, negative, both and neutral. The features are allotted with respect to words, document, sentence and modification involving dependency parse tree for sentence. The sentiments of the polar phrase is identified with the help of above mentioned features.

Ana C.E.S Lima,Leandro N.de Castro (2012) contributes towards the automatic sentiment classification for tweets and is developed comprising of three modules namely Support counting module, Database selection module and classification module. The Support counting module deals with checking of documents which contain atleast one emoticon. The Database selection module separates the dataset into testing and training data. The classification module uses Naïve Bayes classifier to classify the tweets. The sentiment classification is achieved automatically by practicing three approaches namely emoticon based approach which classify based on the presence of emoticons, word based approach which classify based on presence of words and hybrid approach includes the combination of both word and emoticons. This system fails to add neutral words while sentiments of the tweet are classified.

Duyu Tang,Bing Qin,Furu Wei (2015) proposed the joint framework which addresses the sentence level sentiment classification and determines phrase level opinions based on the results of segmentation. The candidate generation model is established for segmentation of sentence and ranking model is accounted for score of the segment candidate and classification model is used to predict the polarity of sentiment. The segmentation ranking model is used to involve the features namely phrase embedding feature and segmentation specific feature. The phrase embedding is capable of embedding varied length of words and it is used by both sentence segmentation and sentiment classification. The efficiency of this approach is cross checked by using it on sentiment classification of two datasets.

In this paper, the sentiment analysis is imposed on tweets which are thereby bilingual in nature composing of Tamil as well as English tweets. The tweets are subjected to word sense disambiguation wherein the semantic distance between the synsets are computed based on edge based similarity method. As a result the polarity of sentiments are derived as positive, negative and neutral by using Support Vector Machine irrespective of the topic or trend to which the tweet belongs.

III Proposed Work

In this work, an architectural framework is proposed to identify the sentiments of both English and Tamil tweet with the help of word sense disambiguation. The proposed framework comprises of two different phases for analyzing the sentiment of tweets. In the first phase, the correct usage of the word sense is determined by using word sense disambiguation technique and hence the path length similarity between two synset is computed with the help of WordNet. In the second phase, the Support Vector Machine is used as a linear classifier to classify the sentiments of tweets by utilizing linear kernel which is the best suited for text classification.

Tweets are collected and gathered with the help of Twitter API. Further the entire set of tweets which is composed of both English and Tamil tweets are subjected to separation by identifying the words in Tamil corpus. If the word is available in the corpus then by default it is considered as a Tamil tweet, otherwise it is concluded as an English tweet.

1.Translation of Tamil Tweet

Tamil is one of the ancient Dravidian languages which has the word order of Subject Object Verb (SOV).Words in Tamil are made up of lexical roots which are meaningful units succeeded by one or more affixes.

The lexical roots and affixes are generally referred as morphemes which is concatenated with each other. The first part of construction of Tamil sentence is a lexical root which may or may not be succeeded by other functional or grammatical morpheme. For instance, நகரங்கள் ‘nagarangkaL’ can be split into நகரம் ‘nagaram’ and கள் ‘kal’.நகரம் is the lexical root and கள் is the affix. There is a possibility of construction of the sentence in Tamil with only verb, or only subject and object.

The main challenge of this work aims at translating Tamil into English sentence because the word order of English is Subject Verb Object (SVO) which differs from the word order of Tamil. With the help of Google Translator the Tamil words are managed to translate into English.

2.Preprocessing of Tweet

The next step is preprocessing of tweet which is essential and foremost process in finding out the sentiments. The preprocessing step includes

2.1Transformation of upper case to lower case

All the upper case letters are shifted into lower case.For instance "This is Awesome" is transformed into "this is awesome".

2.2Tokenisation

In this step, the lower case tweets are immediately split into tokens in order to remove the punctuation marks ("this", "is", "awesome").

2.3Stemming

In this step, tweet is stemmed in order to find out the root of the words.

2.4 POS Tagging

Finally, parts of speech tagging of word facilitates the further steps in indentifying the sentiment 'this' DT(determiner), 'is' VBZ(verb third person singular present),awesome JJ(Adjective).

3.Word Sense Disambiguation using WordNet

Word Sense Disambiguation (WSD) is the ability to disambiguate a word that can have many senses based on its usage context. Since the tweets are informal in nature the sense of the words in tweets are difficult to be determined. In order to overcome this limitation word sense disambiguation (WSD) is employed. There are different methods to approach WSD among which, knowledge based method is preferred to find out the sense of the words. WordNet, being one such knowledge source, figures out the correct contextual usage of the word. WordNet includes collection of words in the form of lexical database and groups these words based on the synonym set referred as synset. The members of synset includes semantic relationship between the members. WordNet includes noun, verb, adjective, adverb and bypass preposition and delimiters. Two synsets are connected with the help of semantic relations. The relation for noun include hypernymy, hyponymy, meronym, holonym and verb includes hypernymy, troponym and entailment.

The semantic relation of noun include

a) hypernymy: B is a hypernymy of A if every A is kind of B (bird is a hypernym of pigeon).
b) hyponymy: B is a hyponym of A if every B is kind of A (pigeon is a hyponym of bird).
c) meronymy:B is a meronym of A if B is a part of A (leaf is a meronym of tree).
d) holonymy:B is a holonym of A if A is part of B (tree is holonym of leaf).

The semantic relation of verb include

a) hypernymy: the verb B is a hypernym of the verb A if the activity A is a kind of B (dance is an hypernym of hip hop).
b) troponymy: the verb B is a troponym of the verb A if the activity B is doing A in some manner (hip hop is a troponym of dance)
c) entailment: the verb Y is entailed by X if by doing X you must be doing Y (sleep is entailed by snore).

The word senses are arranged in the form of taxonomy and to find the semantic similarity between two synset, the path length based similarity is considered. The path length based similarity includes three methods namely Node based method, Edge based method and hybrid method out of which the edge based method is employed by calculating the length of edges on finding the shortest path between the words.

3.1 Edge based similarity

This method exercises the shortest path between concepts C1 and C2 in WordNet to find out the semantic similarity between C1 and C2.The length of edges which has shortest path are brought together to compute the semantic similarity. There are different methods to compute the edge based similarity namely:

3.1.1 Leacock & Chodorow similarity(LCH)

The similarity measure of LCH considers the depth of the taxonomy:

$$Sim(C1,C2) = -\log(len /(2*D)) \quad (1)$$

where len is the length of the shortest path between the two synsets (using node-counting) and D is the maximum depth of the taxonomy[19].

3.1.2 Hirst & St-Onge similarity(hso)

The similarity measure takes the direction turns into account for computing the shortest path between the edges. Allowed direction turns are Upward, Downward and Horizontal and combination of these turns are used while computing the allowable shortest path between C1 and C2.

This can be computed using Eq(2):

$$Relhso(C1,C2)=C–len(C1,C2)-k*t(C1,C2) \quad (2)$$

where C and k are constant and C1,C2 are concepts and t is the number of changes in direction of path[18].

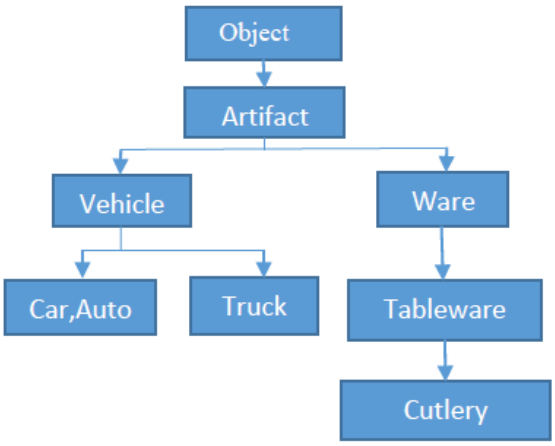

Fig 1:Representation of taxonomy in wordnet used to find path length similarity

The parent which is shared by two synset is known as sub-sumer. The least common subsumer of two synset is most specific subsumer of two synsets. In the above figure, the length between Vehicle and Truck is 1.The LCS of {cutlery} is tableware which is more specific than the sub-sumer ware. The value of path length similarity always ranges between 1 and -1.

4.Support Vector Machine

The sentiments are classified with the help of Support Vector Machine (SVM).SVM is one such classifier well suited for separating the text into different classes. SVM is based on the idea of decision plane which defines decision boundaries. Main usage of SVM is to predict linear separators in the search space which can best distinguish the different classes. Hyperplane affords the best separation between the classes, which represents the maximum margin for separation. SVM classification is the best method for text classification. Because of the sparse nature of text, it includes few features that are irrelevant, but it looks after the correlation and as a result gets separated into linear categories. SVM can predict a nonlinear decision surface in the actual feature space by plotting the data instances in non-linear fashion. A decision plane is having the responsibility of separating the set of entity having different class membership. The process of classification includes training and testing data. The training data is the initial input for the Support Vector Machine.

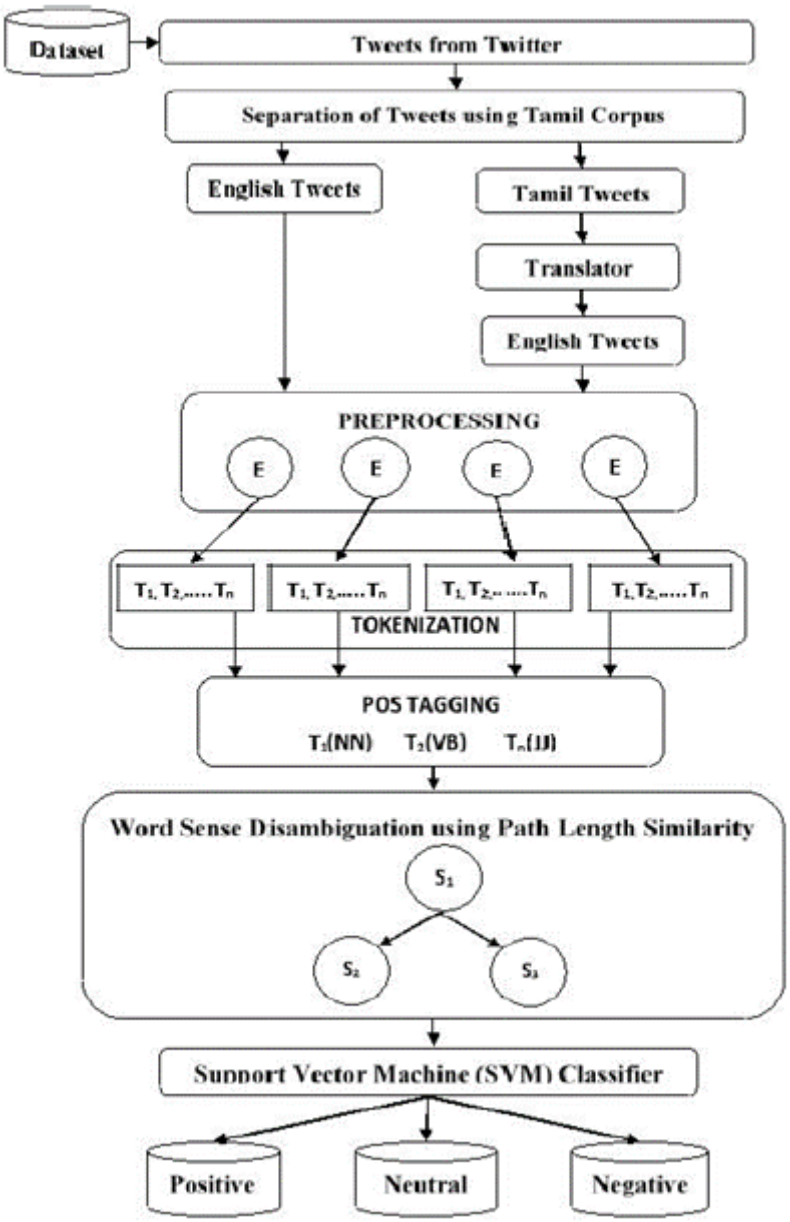

Fig 2:Architecture of the system

In the above diagram,E denotes English Tweets,T1,T2,Tn denotes words in tweets,T1(NN) denotes word is noun,T2(VB) denotes word is verb,T3(JJ) denotes word is adjective,S1,S2,S3 denotes nodes to find path similarity using the edges which connect these nodes.

IV Experiment Setup And Performance Evalution

The datasets are collected from Twitter with the help of TwitterAPI. Twitter is a microblogging service which is launched in 2006.Twitter is not only the best microblog but it also grows at faster rate. Users in the Twitter can communicate their ideas in the form of short informal text referred as tweets. The performance of Sentiment classification is evaluated using Accuracy, Precision, Recall and F_Measure. Evaluation is carried out using confusion matrix as shown in Table 1.

	Predicted Positive	Predicted Negative
Actual Positive Instance	Number of true positive Instance(TP)	Number of false negative Instance(FN)
Actual Negative Instance	Number of false positive Instance(FP)	Number of true negative Instance(TN)

Table 1:Confusion Matrix

Accuracy

Accuracy is proportion of total number of predictions that are correctly classified in class and is determined using Eq. (3):

$$Accuracy=(TN + TP)/(TN + TP + FN + FP) \quad (3)$$

Precision

Precision is percentage of all selected tweets that are correctly classified in class out of available tweets and is calculated using Eq. (4):

$$Precision=TP/(TP + FP) \quad (4)$$

Recall

Recall is percentage of correct tweets that are selected in class from all the available tweets and is calculated using Eq. (5):

$$Recall =TP/(TP + FN) \quad (5)$$

F-Measure

F-Measure is harmonic mean of precision and recall and is calculated using Eq. (6):

$$F\_Measure=2PR/(P + R) \quad (6)$$

4.1 Evaluation

Weka is a data mining tool which is used to perform classification of tweets using Support Vector Machine technique. Stringtoword Vector is used as filter to perform preprocessing step and 1000 tweets are used to model the training class and performance of system is measured using Accuracy, Precision, Recall and F-Measure. In this paper the classification model is evaluated using different combinations of Tweet size such as 200, 400, 600, 800, 1000 and different values of Precision, Recall and F-Measure are evaluated. Fig.3 shows the performance of the proposed system. It is found that the F-measure value for the proposed system using SVM is 0.741 wherein it is 0.68 in the sentiment classification systems using Naïve bayes which thereby proves that the proposed system is advantageous.

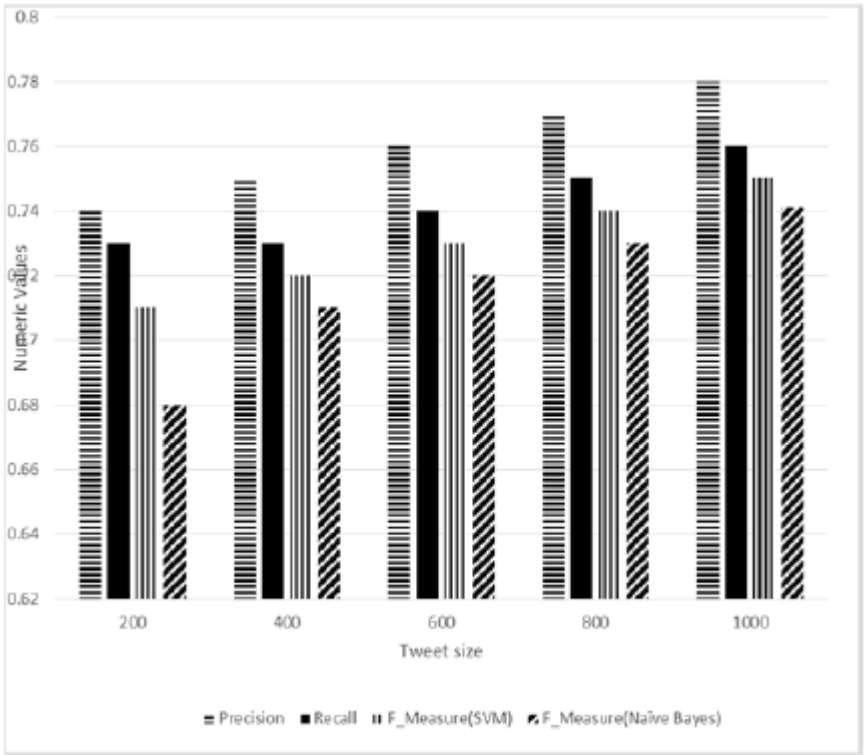

Fig 3:Result of SVM Classification.

V Conclusion

Twitter is a microblog which facilitates user to share their viewpoints on a variety of topics or trends. The opinion expressed by the user is subjected to Sentiment Analysis. The proposed system analyses the sentiment of English Tweets and Tamil Tweets. The Tamil Tweets which are gathered from Twitter API are translated into English Tweets using Google Translator. The tweets are composed of informal text which increases the probability of misunderstanding of word sense with respect to contextual usage. Word Sense Disambiguation is employed to overcome this issue and sematic similarity between the words is predicted using Path based similarity. The shortest distance between two synset is found using Edge based similarity. The Support Vector Machine classifier is used to predict the polarity values of the tweets. It is evident from the findings that the F-measure value for the existing system is 0.68 whereas the F-measure value is 0.741 for the proposed system thus proving the tweets collected are correct and are suitably classified based on polarity. The future work is aimed at involving the Tamil tweets without translation and are to be subjected for word sense disambiguation to find the polarity values.

References

[1] M.Albanese,A.Chianese,A.D'Acierno, V. Moscato, & A.Picariello, A multimedia recommender integrating object features and user behavior Multimedia Tools and Applications, 4(Springer,2010)563–585.

[2] Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng , Why We Twitter: Understanding Microblogging Usage and Communities.

[3] *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*,,2007,56-65.

[4] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, What is Twitter, a Social Network or a News Media?,*Proceedings of the 19th International Conference on World Wide Web*,2007,591-600.

[5] Manju Venugopalan, Deepa Gupta,Exploring sentiment analysis on twitter data,*Proceedings of Eighth International Conference on Contemporary Computing*,2015,241-243.

[5] WalaaMedhat,AhmedHassan,HodaKorashy,Sentiment analysis algorithms and applications: A survey,Ain Shanz Engineering Journal,5(4),2014,1093-1113.

[6] R. V. Vidhu Bhala, S. Abirami,Trends in word sense disambiguation,Artificial Intelligence Review,42(Spinger,2012)159-171.

[7] Anuj Sharma, Shubhamoy Dey,A Boosted SVM Based Sentiment Analysis Approach for Online Opinionated Text,Proceedings of the 2013 Research in Adaptive and Convergent Systems,28-34.

[8] Jantima Polpinij,Multilingual Sentiment Classification on Large Textual Data, IEEE Fourth International Conference on Big Data and Cloud Computing,2014,183-188.

[9] Zheng Lin, Xiaolong Jin, Xueke Xu et.al,Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies(IAT),02,2014,183-188.

[10] AlexandraBalahur,JoséM.Perea-Ortega,Sentiment analysis system adaptation for multilingual processing:The case of tweets,*Information Processing and Management*,51(4),2014,547-556.

[11] Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li,TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets,*IEEE Transaction on Knowledge and Data Engineering*,27(6),2015,1696-1709.

[12] SamirE.Abdelrahman,EbtsamAbdelhakamSayed,Reem Bahgat,The integration among disambiguation lexical resources for more effective phrase level contextual polarity recognition,*Proceedings of 9th International Conference on Computer Engineering Systems (ICCES)*,2014,86-91.

[13] Marco Antonio Sobrevilla Cabezudo, Nora La Serna Palomino, RolandoMagui˜naPerez,Improving subjectivity detection for Spanish texts using subjectivity word sense disambiguation based on knowledge, *Proceedings of Computing Conference (CLEI)*, 2015,1-7.

[14] Umar Farooq, Tej Prasad Dhamala, Antoine Nongaillard et.al , A Word Sense Disambiguation Method for Feature Level Sentiment Analysis,*Proceedings of 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*,2015,1-8.

[15] Theresa Wilson,Janyce Wiebe,Paul Hoffmann,Recognizing Contextual Polarity: An Exploration of Features for Phrase-level Sentiment Analysis,35(3),2009,399-433.

[16] Ana C.E.S Lima,Leandro N. deCastro ,Automatic sentiment analysis of Twitter messages,*Proceedings of Fourth International Conference on Computational Aspects of Social Networks (CASoN)*,2012,52-57.

[17] Duyu Tang, Bing Qin, Furu Wei et.al ,A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classificcation. *IEEE/ACMTransaction on Audio ,speech, and Language processing*,23(11),2015,1750-1761.

[18] G. Hirst,,D.St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms, *WordNet: An electronic Lexical Database* ,305,1998, 305– 332.

[19] C. Leacock, M.Chodorow, Combining local context and WordNet similarity for word sense identification,*WordNet: An electronic Lexical Database*, 49(2),1998, 265–283.