AMRITA-CEN@SAIL2015: Sentiment Analysis in Indian Languages

Shriya Se(✉), R. Vinayakumar, M. Anand Kumar, and K.P. Soman

Centre for Excellence in Computational Engineering and Networking,
Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore, India
{shriyaseshadrik.r,vinayakumarr77}@gmail.com,
m_anandkumar@cb.amrita.edu

Abstract. The contemporary work is done as slice of the shared task in Sentiment Analysis in Indian Languages (SAIL 2015), constrained variety. Social media allows people to create and share or exchange opinions based on many perspectives such as product reviews, movie reviews and also share their thoughts through personal blogs and many more platforms. The data available in the internet is huge and is also increasing exponentially. Due to social media, the momentousness of categorizing these data has also increased and it is very difficult to categorize such huge data manually. Hence, an improvised machine learning algorithm is necessary for wrenching out the information. This paper deals with finding the sentiment of the tweets for Indian languages. These sentiments are classified using various features which are extracted using words and binary features, etc. In this paper, a supervised algorithm is used for classifying the tweets into positive, negative and neutral labels using Naive Bayes classifier.

Keywords: Sentiment analysis · Features · Social media · Machine learning · Supervised algorithm · Naive Bayes classifier

1 Introduction

Opinion plays an important role in deciding about everything in the life as millions of people express their thoughts through personal blogs, social networking sites and many more. Opinions are private states, which are not directly observable by others but expressions of opinion can be reflected through actions including written and spoken languages [1]. Sentiment analysis which is also known as opinion mining is a task in Natural Language processing which deals with the discernment and categorization of opinions in narrative [2]. Predominantly, these opinions are classified into positive, negative and neutral classes and is thus helpful in many fields including marketing, sociology, psychology etc. [3] Sentiment Analysis is popular for English languages and it is found rarely for Indian Languages [4].

Twitter, a microblogging stage, permits its clients to post short messages about any subject what's more, tail others to get their posts. Many individuals

R. Prasath et al. (Eds.): MIKE 2015, LNAI 9468, pp. 703–710, 2015.
DOI: 10.1007/978-3-319-26832-3_67

use Twitter as a platform to communicate with each other. The objective of this examination is to study client opinion communicated on Twitter and to add to a system that permits observing it in the constant. Tweet planning among others include spelling correction, equivalent word substitution, hyperlink cancellation and stop words are performed [5]. Notion is physically ordered the slant into three classes: positive, neutral and negative, so as to make the preparation set for the classifier [6]. The classified tweets are utilized to make positive, neutral and negative feeling lists. The sensational increment in the utilization of the internet as a method for correspondence has been joined by a sensational change in the way individuals express their opinion and perspective [7]. They can express their surveys online about items and administrations and also the perspectives about anything by means of social network (i.e. web journals, examination discussions). Sentiwordnet is one of the widely used lexicon resources for sentiment analysis, emotional analysis, opinion mining [8]. Sentiwordnet is an automatically created lexicon with positive and negative scores [9,10].

The contemporary work is done as slice of shared task in Sentiment Analysis in Indian Language (SAIL) 2015, constrain category. The task which contains three classes (positive, negative, neutral) of twitter data in three languages - Hindi, Bengali and Tamil is to identify the sentiment of the given tweet in a given language. The main objective of the share task is to stimulate researchers to accomplish sentiment analysis in their endemic language. Section 2 provides a view about the methodology used in the system; Sect. 3 discusses about the short analysis of the dataset provided to the work; Sect. 4 discusses about various experiments and their results.

2 Methodology

In the proposed system, feature extraction is the most crucial process as the accuracy of the classifier is based on the extracted feature. The flow of the proposed system is depicted in the Fig. 1. Generally for the text classification problem, preprocessing is to be done and is mandatory especially for twitter dataset. The preprocessing steps include normalization and tokenization. In tokenization, the tweets are further chunked into small instances called tokens. These tokens are normalized using normalization process in which superficial variations are removed from the words and are thus converted to the similar form. The common type of normalization includes case folding and stemming [11]. Predominantly, stemming is avoided for Indian languages in case of text classification as this leads to stem the useful information into its root form. Case folding is used mainly for English language as it has upper case and lower case letters [12]. This is not needed in case of Indian languages as no such case differences exist. The terms which are normalized using the system are listed in Table 1. Along with the features, the machine also learns from the training dataset which is already labelled. A small part of the data from the training dataset say about 10 % is taken and given for the validation process. This is given as an input for the Naive Bayes classifier and

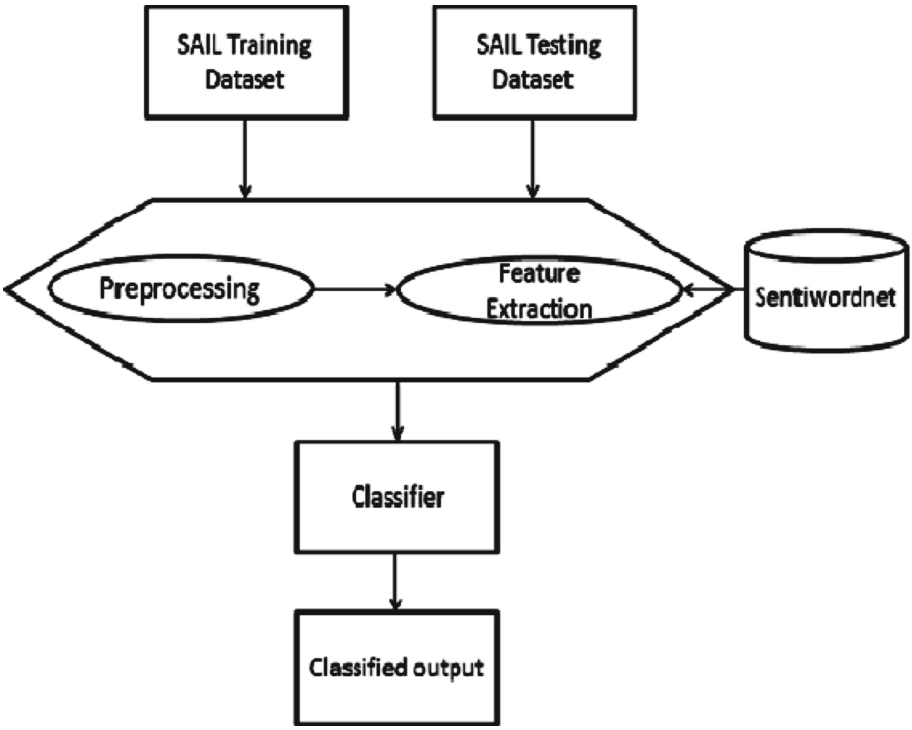

Fig. 1. Flow diagram of the proposed system

Table 1. Normalized symbol

Sl.No	Symbols	Features
1	@	User
2	#	Hash
3	1,2	Numbers
4	:-)	Emoticons
5	!,?	Punct
6	https://	URL

the classified outputs are taken into consideration. Naive Bayes classifier is chosen for the classification purpose as the size of the dataset is very small. In machine learning hunk SciPy library is used for classification.

2.1 Feature Extraction

The words of the sentiwordnet are taken as features because it contains the classified words of respective languages. The binary features are the features in

Table 2. Binary feature description

Sl.No	Symbol	Binary Features	Description
1	HA	#	If # is present in the tweet then 1, else 0
2	RT	RT	If RT is present in the tweet then 1, else 0
3	AR	@	If @ is present in the tweet then 1, else 0
4	LI	http://	If link is present in the tweet then 1, else 0
5	QU	?	If ? is present in the tweet then 1, else 0
6	EX	!	If ! is present in the tweet then 1, else 0
7	SEN_PO	Positive	If the word is from Sentiwordnet Positive file then 1, else 0
8	SEN_NE	Negative	If the word is from Sentiwordnet negative file then 1, else 0
9	SEN_NU	Neutral	If the word is from Sentiwordnet neutral file then 1, else 0

which if a symbol is present in the tweet, it is marked 1. These features are extracted from the twitter dataset as it contains various special characters such as @, RT, # and few more which are enlisted in the Table 2. The stop words are removed from the tweets. All special symbols are removed except for the question mark and the exclamation mark as these punctuations has the ability to change the meaning of a particular tweet.

2.2 Naive Bayes Algorithm

Naive Bayes has been used in information retrieval for many years and recently it has been used for many machine learning researches [13]. Multinomial Naive Bayes has been carried out for this work [14]. In recent years, the work has been focused on two basic instantiations of the classifier Bernoulli model and multinomial model [15]. Bernoulli model represents the document as a vector of binary features whereas the multinomial model uses vector of integer feature to represent documents [16]. The multinomial model works on the assumption that the probability of each word event in a document is independent of the words context and position in the document [17]. To normalize the error in the Naive Bayes, a small correction known as Laplacian Smoothing is included [18,19]. Generally the Naive Bayes is mathematically represented as,

$$P(c|d) = p(c) \prod_{1 \leq k \leq n_d \leq n} p(t_k|c) \tag{1}$$

As shown in the Fig. 2, the tweets are taken and if it includes any punctuations other than exclamations and question marks, it is removed and the tweet ids are also processed. If the tweet has any binary feature, they are marked 1.

S.No	Before Preprocessing	After Preprocessing
1	5,08771E+17 @ImJames_நான் உள்ள இருக்கலாமான்னு கேட்டாக்க,வெளில போய்யா மொதல்லங்கறா இந்த நர்ஸூ//	AR-1 நான் உள்ள இருக்கலாமான்னு கேட்டாக்க வெளில போய்யா மொதல்லங்கறா இந்த நர்ஸூ
2	508707986391719936 चंद नोटों के लिए,-ईमान को बेआबरू होने दिया जाये,इतना सस्ता इंसान का मोल नहीं, जिसे बाजार में जाके बेच दिया जाये .... रवि कवि	चंद नोटों के लिए ईमान को बेआबरू होने दिया जाये, इतना सस्ता इंसान का मोल नहीं जिसे बाजार में जाके बेच दिया जाये रवि कवि
3	508638661265485824 ডি মারিয়া ও আলোনসোকে ছেড়ে দূর্বল হয়নি রিয়াল	ডি মারিয়া ও আলোনসোকে ছেড়ে দূর্বল হয়নি রিয়াল

Fig. 2. Tweets-before and after preprocessing

3 An Analysis of SAIL Dataset

SAIL stands for Sentiment Analysis for Indian Languages. They have released twitter dataset for three languages namely Tamil, Hindi, Bengali. The size of the training and testing dataset is shown in Table 3 in which approximately 27 % of training data of Tamil and Hindi and 54 % of the Bengali data contain URLs. Most of the Tamil tweets are regarding movie reviews and comments about some actors and actresses whereas the Hindi tweets are based on politics. The dataset has issues such as single tweet are there in both of the positive and negative training data and also there are tweets which are repeated. Many of these tweets are misspelt which affect the accuracy of the classifier. The training dataset of Tamil tweets contains more colloquial words which is not present in Sentiwordnet and hence it is not clean whereas the Hindi and Bengali tweets are conventional. In the test data, many of the ambiguous tweets are present. As these data are already ambigiuous, the accuracy drops.

Table 3. Twitter training dataset for SAIL 2015 shared task

Language	Training data				Test data
	Positive	Negative	Neutral	Total	
Tamil	387	316	400	1103	560
Hindi	168	545	493	1222	467
Bengali	277	354	368	999	500

4 Experiments and Results

The experiment is conducted on Windows 64-bit machine with i7 core processor and 8 GB RAM. The tweets from the dataset are taken. The initial step is preprocessing in which steps such as normalization and tokenization are done and the output of this step is raw tokens. These tokens are then given as an input for feature extractor. The feature extractor will take the tokens as input and extract the features from these tokens. The words, Sentiwordnet are taken as features and binary features are also included so as to improve the feature extractor. Sentiwordnet, hashtags, retweet, links, question marks, exclamatory marks are taken as binary features which means if any of these features are present then the output will be 1 else 0. The description of the binary feature is well illustrated in the given Table 2 In this paper, engrossment is given for feature extraction. The features are extracted from the training dataset and stored in a text file. Using the features that are extracted, the classification step is proceeded to. There are different algorithms that are used for the classification in Machine learning. The algorithm which is used in this paper is Naive Bayes classification algorithm. This Naive Bayes algorithm works on the principle of Bayes theorem. The training dataset and testing dataset is given as input for the classifier, in which 10 % of the training data is taken as a validation data. The data are classified using Naive Bayes and the output of the classifier will be the labelled tweets with positive, negative and neutral. The accuracy of the classifier is verified using F-score. The F-score is calculated for all the three classes. It is given below

$$F - scorepos = \frac{numberofpositivedocument}{Totalnumberofdocument} \tag{2}$$

$$F - scoreneg = \frac{numberofnegativedocument}{Totalnumberofdocument} \tag{3}$$

$$F - scoreneu = \frac{numberofneutraldocument}{Totalnumberofdocument} \tag{4}$$

The F-score and the accuracy of the system for all the three languages are given in the Table 4. Table 5 shows accuracy obtained for the proposed system by SAIL. The number of tweets that are classified into their respective classes in three languages are shown in the Fig. 3.

Table 4. Accuracy and F-score of the proposed system(Cross Validated)

Language	F-score	Accuracy
Tamil	0.4832	0.5612
Hindi	0.5219	0.5322
Bengali	0.3942	0.4171

Table 5. Testing result of the proposed system by SAIL

Language	Accuracy (in %)	Positive (in %)	Negative (in %)	Neutral (in %)
Tamil	39.28 %	29.81 %	26.58 %	59.79 %
Hindi	55.67 %	45.79 %	57.37 %	80.0 %
Bengali	33.6 %	29.58 %	34.44 %	39.26 %

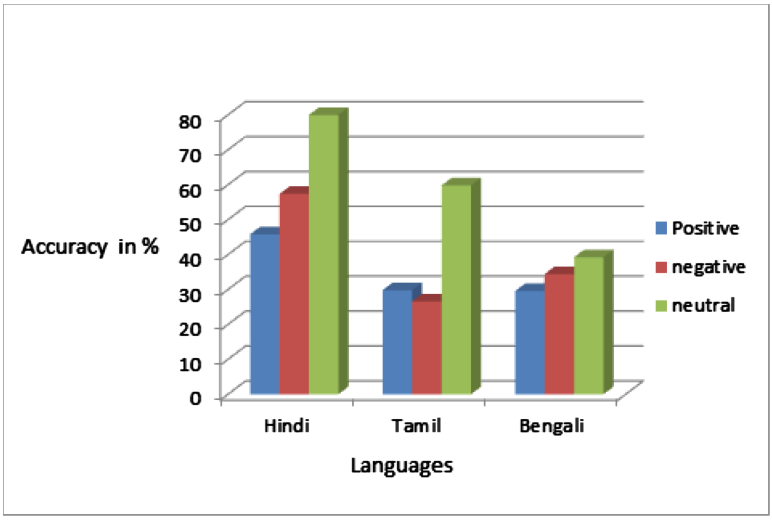

Fig. 3. Bar chart representation of the final result

5 Conclusion and Future Work

We have presented a method to classify twitter data based on the sentiment which is highly useful in the field of information retrieval (IR). In this work, we classify the tweets of the Indian languages into positive, negative and neutral classes. Generally before classifying the tweets, preprocessing is done. This is carried out in order to eliminate the unwanted symbols and also to retrieve words which are highly useful for analysing sentiment. The preprocessing step is taken extra care of and it gave better result after classification. Naive Bayes algorithm is used which gives a better classification result. This method can also be extended using SVM classifier and also the unsuupervised way of implementation can be done as future work.

References

1. Fink, C.R., et al.: Coarse- and fine-grained sentiment analysis of social media text. Johns Hopkins APL Tech. Dig. **30**(1), 22–30 (2011)
2. Balahur, A.: Sentiment analysis in social media texts. In: 4th Workshop on Computational Approaches (2013)

3. Hutto, C.J., Gilbertl, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
4. Arunselvan, S.J., Anand kumar, M., et al.: Sentiment analysis of tamil moovie reviews via feature frequency count. IJAER **10**, 17934–17939 (2015)
5. Jansen, B.J., et al.: Twitter power: tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol. **60**(11), 2169–2188 (2009)
6. Hiroshi, K., et al.: Deeper sentiment analysis using machine translation technology. In: 20th International Conference on Computational Linguistics (2004)
7. John, G.H., Langley, P.: Estimating continuous distributions in Bayesian classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence (1995)
8. Godbole, N., et al.: Large-scale sentiment analysis for news and blogs. ICWSM **7**, 21 (2007)
9. Kouloumpis, E.: Twitter sentiment analysis: the good the bad and the omg!. Icwsm **11**, 538–541 (2011)
10. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
11. Mikolov, T., et al.: Efficient estimation of word representations in vector space (2013). arXiv preprint arXiv:1301.3781
12. Turney, P.D., et al.: From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. **37**(1), 141–188 (2010)
13. Rennie, J.D., et al.: Tackling the poor assumptions of naive bayes text classifiers. In: ICML, vol. 3 (2003)
14. Jordan, A.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. Adv. Neural Inf. Process. Syst. **14**, 841 (2002)
15. Panda, M., Abraham, A., Patra, M.R.: Discriminative multinomial naive bayes for network intrusion detection, pp. 5–10 (2010)
16. Juan, A., Ney, H.: Reversing and smoothing the multinomial naive bayes text classifier. In: PRIS (2002)
17. McCallum, A., Nigam, K.A.: Comparison of Event Models for Naive Bayes Text Classification. In: AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41–48 (1998)
18. Lewis, D.D.: Naive bayes at forty: the independence assumption in information retrieval. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398. Springer, Heidelberg (1998)
19. Amor, N.B., et al.: Naive bayes vs decision trees in intrusion detection systems. In: 2004 ACM Symposium on Applied Computing (2004)