Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview

Braja Gopal Patra[1], Dipankar Das[1], Amitava Das[2], and Rajendra Prasath[3(✉)]

[1] Department of Computer Science and Engineering,
Jadavpur University, Kolkata India
{brajagopal.cse,dipankar.dipnil2005}@gmail.com

[2] Department of Computer Science and Engineering,
IIIT, Sri City, Chittoor India
amitava.santu@gmail.com

[3] Department of Computer and Information Science,
Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
drrprasath@gmail.com

Abstract. Sentiment Analysis in Twitter has been considered as a vital task for a decade from various academic and commercial perspectives. Several works have been performed on Twitter sentiment analysis or opinion mining for English in contrast to the Indian languages. Here, we summarize the objectives and evaluation of the sentiment analysis task in tweets for three Indian languages namely Bengali, Hindi and Tamil. This is the first attempt to sentiment analysis task in the context of Indian language tweets. The main objective of this task was to classify the tweets into positive, negative, and neutral polarity. For training and testing purpose, the tweets from each language were provided. Each of the participating teams was asked to submit two systems, constrained and unconstrained systems for each of the languages. We ranked the systems based on the accuracy of the systems. Total of six teams submitted the results and the maximum accuracy achieved for Bengali, Hindi, and Tamil are 43.2 %, 55.67 %, and 39.28 % respectively.

Keywords: Sentiment analysis · Tweets · Bengali · Hindi · Tamil

1 Introduction

Sentiment Analysis or Opinion Mining from electronic texts is a hard semantic disambiguation problem [1]. Sentiment analysis refers to the process of identifying the subjective responses or opinions about a specific topic. It is observed that sentiment analysis has become a main stream research during the last two decades with an immense possibility from the perspectives of both industry and academia [5]. Sentiment analysis task has been performed on English [6–8] as well as on Indian languages [3, 4] for the plain texts.

A part of this work was done at University College Cork, National University of Ireland, Cork, Ireland.

© Springer International Publishing Switzerland 2015

R. Prasath et al. (Eds.): MIKE 2015, LNAI 9468, pp. 650–655, 2015.
DOI: 10.1007/978-3-319-26832-3_61

On the other hand, the evolution of social media texts such as blogs, micro-blogs (e.g., Twitter), and chats (e.g., Facebook messages) has created not only many new opportunities for information access and language technology, but also many new challenges, making it one of the prime present-day research areas[1]. In case of social texts, the presence of misspellings, poor grammatical structure, emoticons, acronyms, and slang are very common and thus, making the task of sentiment analysis from these texts more difficult. Sentiment analysis becomes more challenging in case of the Twitter text when people try to project their sentiment using only 140 characters. Indeed sentiment analysis on social media text is a hot research discipline in present days, but most of the efforts so far have been made on English. Tasks like sentiment analysis in tweets [8], classifying figurative tweets [6], strength of the sentiment in figurative tweets [7] have been performed in English.

The shared task: *Sentiment Analysis in Indian Language tweets (SAIL-2015)* patronizes the Indian researchers to work on automatic sentiment analysis for their own languages by providing them relevant data. Prime motivation of the *SAIL-2015* is to gather researchers, experts and practitioners together to discuss, collaborate and instigate the sentiment analysis research particularly for Indian languages, which involves resource creation, sharing and future collaboration. In this task, we, the organizers provide tweets for three Indian languages namely Bengali, Hindi, and Tamil annotated with positive, negative, and neural polarity. The main objective of this task is to classify the tweets into positive, negative, and neutral categories.

In the remainder of this paper, we described the sentiment analysis task and the process of creating training and test data for three languages in Sect. 2. In Sect. 3, we presented the results of the participating systems and also analyzed their contributing features and results. Finally, we concluded our study with future steps and work in Sect. 4.

2 Task Description and Data Preparation

2.1 Task Description

Here, we describe the shared task of *SAIL-2015*. Given a tweet, the participants are asked to determine whether it expresses a positive or a negative or a neutral sentiment. If any tweet expresses both positive and negative sentiment, then the stronger one should be chosen. We asked the participants to submit two systems for each of the languages namely constrained and unconstrained systems. In case of the constrained system, the participants are only allowed to use corpus supplied by the organizers and at most the *SentiWordNet* for Indian languages by Das and Bandyopadhyay [1]. No external resource is allowed to develop or be used for the constrained systems. In contrast, the unconstrained system, the participants were allowed to use any external resource (POS tagger, NER, Parser, and additional data) to train their system and they have to mention those resources explicitly in their task reports, accordingly.

[1] http://amitavadas.com/SAIL/index.html.

2.2 Dataset

We collected the training and test datasets from Twitter over a period of three months. It is difficult to search tweets, specifically in Bengali or Hindi or Tamil. Therefore, we followed an interesting approach to collect the training and test data. First, we collected the monolingual corpus for each of the languages manually on different topics. Then, we removed the stop words and prepared a word frequency list. We searched each of the words exists in the frequency list in Twitter and collected the maximum of 2000 tweets for each of the words. We used the TWITTER4J[2], a Java implementation of Twitter API to download the tweets. The duplicate tweets were removed and the statistics of the training and test dataset are given in Table 1. We also counted the number of smiles or emoticons in each of the classes after normalizing them. For example, we normalized the happy smiley having multiple brackets, i.e. we converted ':)))))))))' to :). The statistics of the smiley for each of the classes separately is given in Table 2. We can observe that the usage of smiley is more in case of Tamil compared to Bengali and Hindi.

The undergraduate students annotated these data voluntarily. The annotators are the native speakers of the above mentioned languages. We employed 12, 4 and 2 annotators for annotating the Bengali, Hindi, and Tamil language tweets, respectively. Examples from each of the languages are given in Fig. 1.

Table 1. Data statistics

		Positive	**Negative**	**Neutral**	**Total**
Bengali	Training	277	354	368	999
	Test	213	151	135	499
Hindi	Training	168	559	494	1221
	Test	166	251	50	467
Tamil	Training	387	316	400	1103
	Test	208	158	194	560

Table 2. Smiley count for each class

	Training			**Test**		
	Positive	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**
	+ve/− ve	**+ve/− ve**	**+ve/− ve**	**+ve/− ve**	**+ve/− ve**	**+ve/− ve**
Bengali	21/0	10/15	14/4	18/9	10/15	10/10
Hindi	16/2	12/4	18/2	15/0	5/15	0/0
Tamil	24/3	10/7	28/6	22/8	15/9	12/0

[2] http://twitter4j.org/en/index.html.

3 Results and Discussion

3.1 Results

Initially, 21 teams from different institutes all over India have registered for the task and finally, six teams succeeded to submit the results. We calculated the accuracy of the positive, negative, neutral tweets individually as well as total with respect to different teams for all the submitted systems. The team IDs, individual results obtained for each polarity and the total accuracy are shown in Table 3.

Table 3. Results of each team as per evaluation criteria (B: Bengali, H: Hindi, T: Tamil, C: *Constrained*, U: *Unconstrained*)

Team names	Positive	Negative	Neutral	Total accuracy
JUTeam_KS (B_C)	23.94	60.26	47.41	41.2
JUTeam_KS (B_U)	21.13	63.58	45.18	40.40
JUTeam_KS (H_C)	2.41	88.45	22.0	50.75
JUTeam_KS (H_U)	3.61	82.87	28.00	48.82
IIT-TUDA (B_C)	23.47	59.60	56.30	**43.2**
IIT-TUDA (B_U)	24.88	54.30	55.56	42.0
IITTUDA (H_C)	9.04	73.70	64.0	49.68
IITTUDA (H_U)	4.22	69.72	68.0	46.25
ISMD (H_C)	4.22	58.17	72.0	40.47
ISMD (H_U)	1.81	42.63	72.0	31.26
AmritaCENNLP (B_C)	27.23	65.56	0.0	31.4
AmritaCENNLP (H_C)	36.14	64.94	2.0	47.96
AmritaCENNLP (T_C)	57.77	40.51	0.52	32.32
AMRITA-CEN (B_C)	29.58	34.44	39.26	33.6
AMRITA-CEN (H_C)	45.79	57.37	80.0	**55.67**
AMRITA-CEN (T_C)	29.81	26.58	59.79	**39.28**
AMRITA (H_C)	17.47	54.59	68.0	42.83

Bengali: **দেশে এখনো আড়াই কোটি লোক নিরক্ষর: প্রাথমিক ও গণশিক্ষামন্ত্রী** (*Negative*)
Transliteration: Deshe ekhono adai koti lok nirakhor: prathamik o ganasikhamontri
Translation: **Till date, two and half crores** of people are illiterate in the Nation : Primary and Mass Education Minister

Hindi: भारत के पूर्व राष्ट्रपति एपीजे अब्दुल कलाम को चीन की पेकिंग यूनिवर्सिटी में पढ़ाने का न्यौता मिला है। (*Positive*)
Transliteration: Bharat ke purb rastrapati APJ Abdul Kalam ko China ki Peking University main padane ka niyota mila hai |
Translation: The Former President of India APJ Abdul Kalam got an invitation for teaching at Peking University, China.

Tamil: கடவுள் என்று தனியாக ஒருவர் இல்லை...!! உன் கைகாசில் ஒரு குழந்தைக்குசாக்லேட் வாங்கிக்குடுத்துவிட்டு அந்த குழந்தையின் மு கத்தைபார் !! (*Positive*)
Transliteration: kadavul endru thaniyaga oruvar illai ... !! un kaikasil oru kizhandaikku chaaklet vanghikkuduththuvittu andha kuzhandaiyin mugaththaippaar!!
Translation: There is no one like GOD ... !! Offer a chocolate to a kid by spending your own money and then look at the face of the kid !!

Fig. 1. Examples of Indian language tweets

3.2 Discussion

Total of four teams have submitted the results for Bengali. It is observed that the maximum accuracy achieved for the Bengali language is 43.2 % by the team *IIT-TUDA*. For Hindi, total six teams have submitted the results and among them *AMRITA-CEN* achieved the maximum accuracy of 55.67 %. Only two teams have submitted the results for Tamil language and *AMRITA-CEN* is achieved the maximum accuracy of 39.28 %. Most of the teams used the *SentiWordNet* that was developed in [1] for the constrained system. Teams have used the features like *hash tags*, *re-tweet*, *TF-IDF scores of n-grams*, *links*, *question marks*, *exclamatory marks*, *smiley lists* and *SentiWordNet* for the sentiment analysis task.

The teams have used several well-known supervised classification algorithms like Decision Tree, Naïve Bayes, Multinomial Naïve Bayes and Support Vector Machines (SVMs). We observed that, the accuracies of the unconstrained systems are less compared to the constrained systems. The main reason may be the unavailability of basic NLP tools like POS taggers and NER for Indian language tweets, specifically.

The accuracies of the systems for the Indian language tweets are less as compared to the systems for English tweets as mentioned in [8]. The reason may be the scarcity of the sentiment lexicons for Indian languages tweets. However, a good number of sentiment lexicons available for the Hindi, Bengali and Tamil, but these are collected from the plain texts and not specialized for tweets. Because, in case of tweets, there are many spelling variations, acronyms and emoticons that make the sentiment analysis task more difficult and challenging as compared to other tasks on traditional sentiment analysis. It is also difficult to collect the monolingual Indian language tweets as most of the cases the tweets are written using English alphabets and sometime tweets are code mixed. The annotation of such monolingual tweets based on sentiment expressions requires the involvement of manpower and time.

4 Conclusion and Future Work

We described the first shared task organized for the sentiment analysis of Indian languages for Twitter data. This time, 21 teams have registered for the task and six teams successfully submitted their results using different features and machine learning algorithms. In future, we will try to draw the attentions from more number of teams to participate in this task. We hope this shared task will facilitate more research on the sentiment analysis for Indian language tweets by focusing different research challenges associated with it.

We are planning for a new edition of sentiment analysis task in Indian Languages tweets in the coming year, where focus will be on the Code-Mixed tweets. We are also planning to prepare data for classifying the figurative tweets for Indian Languages.

References

1. Das, A., Bandyopadhyay, S.: SentiWordNet for Indian languages. In: Proceedings of the 8th Workshop on Asian Language Resources (COLING 2010), Beijing, China, pp. 56–63 (2010)
2. Das, A., Bandyopadhyay, S.: Dr. sentiment knows everything! In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011 Demo Session), Portland, Oregon, USA, pp. 50–55 (2011)
3. Das, D., Bandyopadhyay, S.: Word to Sentence Level Emotion Tagging for Bengali Blogs. In: Proceedings of the ACL IJCNLP-2009, Suntec, Singapore, pp. 149–152 (2009)
4. Das, D., Bandyopadhyay, S.: Labeling emotion in Bengali blog corpus – a fine grained tagging at sentence level. In: Proceedings of the 8th Workshop on Asian Language Resources (COLING 2010), Beijing, China, pp. 47–55 (2010)
5. Patra, B.G., Takamura, H., Das, D., Okumura, M., Bandyopadhyay, S.: Construction of emotional lexicon using potts model. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP 2013), Japan, pp. 674–679 (2013)
6. Reyes, A., Rosso, P.: On the difficulty of automatically detecting irony: beyond a simple case of negation. Knowl. Inf. Syst. **40**(3), 595–614 (2014)
7. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: Semeval-2015 task 11: sentiment analysis of figurative language in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Co-located with NAACL, Denver, Colorado, pp. 470–478 (2015)
8. Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V.: SemEval-2015 task 10: sentiment analysis in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Co-located with NAACL, Denver, Colorado, pp. 451–463 (2015)