IIT-TUDA: System for Sentiment Analysis in Indian Languages using Lexical Acquisition

Ayush Kumar[1], Sarah Kohail[2], Asif Ekbal[1], and Chris Biemann[2]

[1]Department of Computer Science and Engineering, IIT Patna, India
[2]Language Technology, Computer Science Department,
Technische Universität Darmstadt, Germany
{ayush.cs12,asif}@iitp.ac.in, {kohail,biem}@lt.informatik.tu-darmstadt.de

Abstract. Social networking platforms such as Facebook and Twitter have become a very popular communication tools among online users to share and express opinions and sentiment about the surrounding world. The availability of such opinionated text content has drawn much attention in the field of Natural Language Processing. Compared to other languages, such as English, little work has been done for Indian languages in this domain. In this paper, we present our contribution in classifying sentiment polarity for Indian tweets as a part of the shared task on Sentiment Analysis in Indian Languages (SAIL 2015). With the support of a distributional thesaurus (DTs) and sentence level co-occurrences, we expand existing Indian sentiment lexicons to reach a higher coverage on sentiment words. Our system achieves an accuracy of 43.20% and 49.68% for the constrained submission, and an accuracy of 42.0% and 46.25% for the unconstrained setup for Bengali and Hindi, respectively. This puts our system in the first position for Bengali and in the third position for Hindi, amongst six participating teams.

Keywords: Sentiment Analysis, Distributional Thesaurus, Co-occurrence, Indian Languages.

1 Introduction

Sentiment Analysis is a Natural Language Processing (NLP) task, which deals with finding the orientation of thoughts and opinions expressed in a piece of text [16]. Recently, a large body of work has been devoted to automating the process of analyzing and extracting sentiments from social media platforms and review forums [19, 20]. The rapid evolution in sentiment analysis has opened up the opportunities for governments and business organization to track the public opinion about their products and services.

Most of the existing work in sentiment analysis are dedicated to processing languages such as English, German and French. Sentiment analyzers developed for such languages are not directly applicable for Indian languages, which have their own challenges with respect to language constructs, morphological variation and grammatical differences.

Sentiment Analysis in Indian Language (SAIL) [17] tweets is the first attempt to bring together the researchers for resource creation and knowledge discovery in Hindi, Bengali and Tamil. Given a set of annotated tweets in these Indian languages, the task is to classify whether the tweet is of positive, negative, or neutral sentiment, which is also called polarity classification [2]. Teams are allowed to run their systems in two modes: constrained mode and unconstrained mode. In constrained mode, the participating team is only allowed to use the resources provided by the task organizers (i.e. tagger, parser, corpus). In contrast to this, participants were allowed to use any external resource in unconstrained mode.

Probably the most important resource for polarity classification is the sentiment lexicon. The sentiment lexicon is a list of words and phrases that convey sentiment polarities. It plays an essential role in most sentiment analysis applications [9]. Considering the lack and scarcity of available sentiment lexicons for Indian languages, we introduce an unsupervised approach for expanding a (small) Indian sentiment lexicon, leveraging distributional thesauri and sentence level co-occurrence statistics. Using the new expanded lexicon, we propose a sentiment classifier based on Support Vector Machines (SVM). We have participated in the SAIL task in two languages: Hindi and Bengali.

The remainder of this paper is organized as follows: Section 2 presents related works. Section 3 describes our method including dataset preprocessing, feature extraction and the lexicon expansion technique. Section 4 presents and discusses our experimentation results and evaluation, followed by conclusions and future work in the last section.

2 Related Work

Trends in the last few years show the inclination of research community towards social media like Twitter to sense public opinions, in commerce to anticipate stock market trends, to predict the outcome of elections [14],[5],[21]and even in disaster management [13] using a variety of approaches and experimental setups. However, most of the existing work in sentiment detection involve non-Indian languages except some prior work for Bengali [7]. The authors used SentiWordNet as well as a subjectivity lexicon to generate a lexical resource containing over 35,000 Bengali entries. Using the lexicon and features like positional aspect, the supervised sentiment classifier based on Conditional Random Field (CRF) achieved a precision of 74.6% and recall of 80.4% in the blog domain. A fall-back strategy for sentiment analysis in Hindi is reported in [11]. The results show that in-language sentiment analysis outperforms MT-based and resource-based sentiment analysis, where e.g. Hindi texts are translated automatically to English and are subsequently classified by an English sentiment analysis system.

3 Methodology

In this section, we discuss the process of building and training our sentiment classifier for the constrained and unconstrained runs. For the machine learning setup, we choose Support Victor Machine (SVM) as the classification model [6], as it can cope well with a large number of nominal features.

3.1 Preprocessing

We replace the URL links in all tweets with 'someurl', all @username with 'someuser' and multiple white spaces with single whitespace and tokenize the tweets in order to identify word tokens.

3.2 Features

We use the following features to train the SVM classifier:

- **Character and Word Features:** Writing style features like word and character n-grams features, often incorporated in stylometry research, have also shown to be effective in sentiment analysis [1]. They are also commonly applied to non-formal texts and user-generated content. For unstructured short texts like tweets, small values for n have shown to be most effective [10]. In our experiments, word unigrams and bigrams are extracted from the dataset. We also compute the n-gram overlap at the character level on the basis of character trigrams and quad-grams for word prefixes and suffixes.
- **SentiWordNet Features:** For this task, the organizer-provided Indian sentiment lexicons [8] include a list of positive, negative, neutral and ambiguous words with the corresponding part of speech (PoS) tags. We denote the words in SentiWordNet with a score of 1 if it is found in the positive list, -1 if it occurs in negative list and 0 if the word appears in the neutral list. Based on our annotation, we count the following features:
  1. Number of tokens in the tweet with $score(w) > 0$.
  2. Number of tokens in the tweet with $score(w) < 0$.
  3. Number of tokens in the tweet with $score(w) = 0$.

3.3 Lexical Acquisition

Lexical expansion [12] is an unsupervised technique that is based on the computation of distributional thesaurus [3]. While Miller et al. [12] used a DT for lexical expansions for knowledge-based word sense disambiguation, the expansion technique can also be used in other text processing applications. For rare words and unseen instances, lexical expansion can provide a useful back-off technique [4, 15].

For the unconstrained submission, we use an external dataset to generate separate lexicons for both Hindi and Bengali and run the same SVM model with

additional features derived from the lexicon. We now describe this expansion technique.

We exploit the concept of distributional thesaurus and sentence level co-occurrences from large background corpora[1] to build a lexicon, denoted later as DT_COOC, assigning each entry two scores between -1 to 1: one score computed over distributional similarity and the other obtained using the co-occurrences. We also assign a third score equal to -1 (absolute negative) and 1 (absolute positive) for each word in the lexicon. For background corpora, we use a Hindi corpus containing a total of 2,358,708 sentences (45,580,789 tokens) and a Bengali corpus of 109,855 sentences (1,511,208 tokens). Both corpora are constructed from online newspapers from 2011.

3.4 Distributional Thesaurus

A Distributional Thesaurus (DT) is an automatically computed resource that relates words according to their similarity. For every sufficiently frequent word in the corpus, we find out the most similar words as computed over the similarity of contexts these words appear in. We employ an open source implementation of the DT computation as described in [3], where complete details of the computation are described.

To illustrate this, a few examples of words and their distributionally most similar words are given in Figure 1. Our core assumption, which is backed up by data analysis, is that sentiment words are similar to other sentiment words. Moreover, while there are usually high similarities between words of positive and negative sentiment (such as 'good' and 'bad'), words tend to be similar to more words of the same sentiment.

Words	Similar Words from Distributional Thesaurus		
अतुलनीय (atulnIya)	अद्भुत (adabhuta)	महान (mahAna)	शानदार (shAnadAra)
तर्कसंगत (tarkasangata)	उचित (uchita)	सही (sahI)	गलत (galata)
धार्मिक (dhArmika)	सामाजिक (sAmAjika)	राजनीतिक (rAjanItika)	हिंदू (hindU)
ऊँची (UNchI)	ऊंची (UnchI)	लंबी (lambI)	छोटी (ChotI)
परम्परा (paramparA)	परंपरा (paramparA)	संस्कृति (sanskriti)	धर्म (dharma)

Fig. 1. Illustrative examples of words appearing in the DT expansion for Hindi.

3.5 Co-occurrences

We obtain a list of words that co-occur significantly with the other words in a sentence [18]. Some examples of words and their most significant co-occurrences are displayed in Figure 2. Our core assumption here is that sentence contexts

[1] from `http://corpora.informatik.uni-leipzig.de`

are mostly either positive, negative or neutral. While this does not hold in all cases, we have observed from data analysis that words of the same polarity tend to co-occur more than words of different polarity.

Words	Words from Co-Occurrence Lists		
अतुलनीय (atulnIya)	भारतीय (bhAratIya)	अन्य (anya)	वर्ष (warSha)
तर्कसंगत (tarkasangata)	कहना (kahanA)	ज्यादा (jyadA)	काफी (kAphI)
धार्मिक (dhArmika)	परंपराओं (paramparAon)	अपितु (apitu)	संतों (santon)
ऊँची (UnchI)	इमारत (imArata)	जाति (jAti)	जगहों (jagahon)
परम्परा (paramparA)	अनुसार (anusAra)	नई (nayI)	जीवन (jIvana)

Fig. 2. Illustrative examples of words appearing in the co-occurrence list for Hindi.

3.6 Construction of DT_COOC Lexicon

We use the given SentiWordNet for both the languages as the seed data for lexical expansion. We first filter out the candidate sentiment terms using DT expansion and then create a final lexicon using the agreement between the DT polarity list and COOC polarity list. In the subsequent sections we describe the steps in more details.

3.6.1 Finding the candidate sentiment terms: At first, after constructing the seed corpus, we obtain the top (i.e. most similar) 125 DT expansions for each word in the seed corpus. In context of further use, we define two terms: positive expansion list and negative expansion list. The DT expansion of positive and negative words in the seed corpus results in positive and negative expansion lists, respectively. To filter out the candidate terms from the noisy tokens, we rank each word in the complete expansion list with a score (*candidateScore*).

$$candidateScore = \frac{Number\ of\ expansion\ lists\ the\ word\ appears\ in}{Frequency\ of\ the\ word\ in\ the\ DT\ corpus} \quad (1)$$

Dividing through the frequency ensures that highly frequent words, which are similar to almost every word just because they occur in so many contexts, are down-ranked.

3.6.2 Calculating the DT score: Based upon *candidateScore*, we remove the 500 lowest-ranked terms for lexicon generation. Of the remaining words in the expansion, we compute another score (*score_DT*):

$$score\_DT = \frac{No.\ of\ positive\ expansions - No.\ of\ negative\ expansions}{No.\ of\ expansion\ lists\ the\ word\ appears\ in} \quad (2)$$

The DT score is a graded score between -1 and 1 that projects sentiment to new words, based on the known sentiment of distributionally similar words.

3.6.3 Calculating the COOC score: From the pruned list, we calculate a score *(score_COOC)* for each word using the sentence-based co-occurrences. We define the number of *pos co-occurences* as the total number of positive seed words with which word co-occurs. Accordingly, *neg co-occurrence* is defined analogously.

$$score\_COOC = \frac{No.\ of\ pos\ co\ occurrences - No.\ of\ neg\ co\ occurrences}{No.\ of\ seed\ words\ with\ which\ given\ word\ co-occurs} \quad (3)$$

3.6.4 Generating the final lexicon: To construct a final expanded lexicon, we consider the agreement between the two scored lists at the absolute polarity level: for the final lexicon, only those words are added where both methods agree on polarity. The statistics of the generated lexical corpus is given in Table 1:

Table 1. Statistics of induced lexicon for both languages.

Dataset	Positive	Negative	Neutral	Total
		First Expansion		
Hindi	3980	3331	357	7668
Bengali	1205	10005	600	11810
		Final Expansion		
Hindi	5521	3926	48	9495
Bengali	7213	1461	30	8704

In principle, the expansion procedure can be iterated to bootstrap sentiment lexicons: the output of one step can serve as the input of the next expansion step. Here, we explore two levels of expansion for Hindi, using described lexicon as the new seed. However, for Bengali DT_COOC Lexicon, we note that the expansion list is too skewed: he number of negative words in the lexical corpus is much higher than the positive ones. One possibility for the skewness might be the difference in the number of positive and negative words in the seed corpus. To overcome the skewness, we balance the Bengali seed by random sampling, removing negative instances randomly until we arrive at the same number of negative positive words. Finally, we perform all the steps sequentially to obtain the expanded lexical corpus. In preliminary experiments, however, we have not found this technique to be effective for Bengali, which might be related to the corpus size.

The statistics of the final expansion lexicons for both languages, as used in our experiments, are shown in Table 1.

4 Datasets and Experimental Results

To tune and to evaluate our approach, we perform five-fold cross validation on the training set. The datasets are annotated with three classes, namely positive, negative and neutral. The overall distribution of both train and test set per class label is given in Table 2. We used classification accuracy as a measure of sentiment polarity classification performance. Based on the cross validation results, the feature combination that we use for the various runs for both languages is given in Table 3. We make use of the LibLinear[2] SVM implementation.

Table 2. Distribution of training and test set for Hindi and Bengali language.

Dataset	Positive Tweets	Negative Tweets	Neutral Tweets	Total
		Hindi		
Training Set	168 (13.75%)	559 (45.74%)	494 (40.46%)	1221
Test Set	166 (35.54%)	251 (53.74%)	50 (10.70%)	467
		Bengali		
Training Set	277 (27.73%)	354 (35.43%)	368 (36.83%)	999
Test Set	213(42.60%)	151 (30.20%)	135 (27.00%)	499

Table 3. Feature combination for different modes of submission for both languages. Description of the features: 1. Word N-Gram, 2. SentiWordNet for respective language, 3. Character N-Gram of prefizes and suffixes of size 3 and 4, 4. DT_COOC Lexicon for respective language.

Mode	Hindi	Bengali
Constrained	1 + 2	1 + 2
Unconstrained	1 + 4	1 + 2 + 3 + 4

For Bengali, our system achieves an accuracy of 43.2% and 42.0% for the constrained and unconstrained runs, while we score an accuracy of 49.68% and 46.25% for Hindi in the constrained and unconstrained setups, respectively. The confusion matrix in Table 4 shows that the classifier performs very poorly on positive instances in comparison to other two classes in Hindi. The less percentage of positive tweets (13.75%) in the training set might be a cause for inaccurate classification. We also analyze the labeled data to determine the statistics of

[2] `http://liblinear.bwaldvogel.de/`

Table 4. Confusion Matrix for Hindi and Bengali

Class	Positive	Negative	Neutral
	Hindi		
Positive	7	65	94
Negative	2	175	74
Neutral	0	16	34
	Bengali		
Positive	53	52	109
Negative	17	82	52
Neutral	20	40	75

Table 5. Experimental results for feature ablation for Hindi and Bengali. The values in the parenthesis denotes the deviation from the score when all the features were taken into consideration.

Features	Accuracy: Hindi	Accuracy: Bengali
All	47.96	42.00
All-SentiWordNet	47.32 (-0.64)	41.20 (-0.80)
All-Word ngram	43.25 (-4.71)	38.40 (-2.80)
All-Character ngram	47.75 (-0.21)	42.20 (+0.20)
All-DT_COOC Lexicon	49.03 (+1.07)	42.20 (+0.20)

tokens that match in the training and test sets. The percentage of unique overlapping tokens between training and test set is 49.71% and 41.36% for Hindi and Bengali respectively. However, the values drop to 29.91% and 27.07% for the positive tweets in the two languages respectively. On further investigation, we find the token to have overlap between neutral tweets in the training set and positive tweets in test set, and this to be 45.21% for Hindi which is a possibility for a majority of positive instances classified as neutral. This shows that the training data is not rich enough to capture the new positive instances effectively. We also analyze the coverage of SentiWordNet and the induced DT_COOC Lexicon on the dataset. Assuming the adjectives to be the most dominating sentiment term in the tweet, we extract the sentiment terms in Hindi tweets using a POS Tagger[3]. Only 17.57% and 25.98% of the adjectives in the training and test set appear in the HindiSentiWordNet list. The coverage improves to 36.56% and 42.29% adjectives in the training and test set, respectively, while using DT_COOC Lexicon.

To get an insight to the contribution of each feature in the development of the system, we perform feature ablation experiment. Results of the detailed feature ablation study are shown in Table 5. We find that word ngram is the most

[3] `http://sivareddy.in/downloads\#hindi\_tools`

important feature in both languages which improves the accuracy by 2%-5%. The second most important feature is the SentiWordNet feature which helps in improving the results upto 0.8%. However, we observe a drop in performance with the induced lexicon, probably because the external dataset is from a different domain or the expansion is too aggressive. Since all participating systems achieved lower scores in their unconstrained runs, this could point either at overfitting on a small training set, or a selection of the data that was biased on the provided lexical resources.

5 Conclusions and Future Work

In this paper, we developed an SVM-based classifier for Indian tweet polarity classification. Our contribution is part of a recently held evaluation challenge (SAIL 2015), which aims to instigate researchers and experts to discuss and advance sentiment analysis research for Indian languages. Our system is based on supervised classification, SVM, which is enriched by using a lexicon expansion technique based on distributional thesauri and co-occurrences. Our system achieved the highest accuracy in Bengali in the competition, and we score third for Hindi amongst the participating teams.

However, there is a lot of scope of improvement and a large headroom – classification accuracies throughout do not even come close to a quality that would be useable in industrial applications. We first and foremost attribute this to the small amount of training and test data. We find that training set in both languages contain significant percentage (3.6% and 6.5% in Hindi and Bengali respectively) of quasi-duplicates (tweets that differ in just URL mention or spacing between punctuation marks or @mentions or simply identical duplicates) which would have resulted in overfitting. While it is very commendable that the effort for sentiment data creation for Indian language has started, it has to be expanded significantly in order to yield reliable results in the future.

In the future, we would like to create in-domain lexicon to test the effectiveness of our method since we still believe that expanding the lexicon with statistical methods is a simple yet effective method for increasing model coverage. We also plan to investigate and implement more features specific to the languages.

References

1. Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. 26(3), 12:1–12:34 (June 2008)
2. Almatrafi, O., Parack, S., Chavan, B.: Application of location-based sentiment analysis using twitter for identifying trends towards indian general elections 2014. In: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. pp. 41:1–41:5. IMCOM '15 (2015)
3. Biemann, C., Riedl, M.: Text: Now in 2d! a framework for lexical expansion with contextual similarity. Journal of Language Modelling 1(1), 55–95 (Apr 2013)

4. Biemann, C.: Unsupervised part-of-speech tagging in the large. Research on Language and Computation 7(2-4), 101–135 (2009)
5. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2011)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
7. Das, A., Bandyopadhyay, S.: Subjectivity detection in english and bengali: A crf-based approach. Proceeding of ICON, Hyderabad, India (2009)
8. Das, A., Bandyopadhyay, S.: Sentiwordnet for indian languages. Asian Federation for Natural Language Processing, China pp. 56–63 (2010)
9. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM 56(4), 82–89 (2013)
10. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural. Expert Systems with Applications 40(16), 6266 – 6282 (2013)
11. Joshi, A., Balamurali, A., Bhattacharyya, P.: A fall-back strategy for sentiment analysis in hindi: a case study. Proceedings of the 8th ICON, Kharagpur, India (2010)
12. Miller, T., Biemann, C., Zesch, T., Gurevych, I.: Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In: COLING. pp. 1781–1796 (2012)
13. Nagy, A., Stamberger, J.: Crowd sentiment detection during disasters and crises. In: Proceedings of the 9th International ISCRAM Conference, Vancouver, Canada. pp. 1–9 (2012)
14. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC 11(122-129), 1–2 (2010)
15. Panchenko, A., Beaufort, R., Naets, H., Fairon, C.: Towards detection of child sexual abuse media: Categorization of the associated filenames. In: Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 7814, pp. 776–779. Springer Berlin Heidelberg (2013)
16. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2(1-2), 1–135 (2008)
17. Patra, B.G., Das, D., Das, A., Prasath, R.: Shared task on sentiment analysis in indian languages (sail) tweets - an overview. In: Mining Intelligence and Knowledge Exploration - Third International Conference, MIKE-2015. Springer, Hyderabad, India (2015)
18. Quasthoff, U., Richter, M., Biemann, C.: Corpus portal for search in monolingual corpora. In: Proceedings of the fifth international conference on language resources and evaluation, Genoa, Italy (2006)
19. Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S.M., Ritter, A., Stoyanov, V.: Semeval-2015 task 10: Sentiment analysis in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval, Denver, Colorado (2015)
20. Rosenthal, S., Nakov, P., Ritter, A., Stoyanov, V.: Semeval-2014 task 9: Sentiment analysis in twitter. Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval, Dublin, Ireland (2014)
21. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM, Washington, DC 10, 178–185 (2010)