

Konkani SentiWordNet - Resource For Sentiment Analysis Using Supervised Learning Approach

Ashweta Fondekar, Jyoti D Pawar, Ramdas N Karmali

Goa University

Department of Computer Science and Technology, Goa University Taleigao Plateau, Goa-403206

ashu.fondekar57@gmail.com, jdp@unigoa.ac.in, rnk@unigoa.ac.in

Abstract

Sentiment Analysis (SA) is the process of analyzing and predicting the hidden attitude/opinion in the given text expressed by an individual. Till now, ample amount of work has been carried out for the English language. But, no work is performed for the language Konkani in the field of Sentiment Analysis. Lexicon-based SA is a good beginning for any language, especially if the digital content is limited. Hence, the main motive of this paper is; to present the sentiment lexicon called SentiWordNet for Konkani language. The process of creating Konkani SentiWordNet is under progress using the Supervised Learning Approach. In this approach, the training set is generated using a Synset Projection Approach and Support Vector Machine (SVM) algorithm to classify the data. The reason behind using the Synset Projection Approach for building a training dataset is; English Sentiwordnet is developed using Semi-Supervised Approach where the training dataset is generated using WordNet lexical relations but; in Konkani WordNet, lexical relations are not yet developed. Hence, Synset Projection Approach is preferred. Conducted experimental results for the proposed algorithm are reported in this paper.

Keywords: Sentiment Analysis, sentiment lexicon, Konkani SentiWordNet (K-SWN), Hindi SentiWordNet (H-SWN), English SentiWordNet (E-SWN), IndoWordNet, Supervised Learning Approach, Synset Projection Approach

1. Introduction

Nowadays, as mentioned by (Pontiki et al., 2015) sentiments expressed by the people plays a crucial role in decision-making such as which product to buy, which movie to watch, which the political party to be supported etc. These Sentiment values of a document, text, article and the topic are computed using Sentiment Analysis algorithms. Most of the work in Sentiment Analysis has been carried out for the English language. For which, many of the resources are already developed and made available for the use, such as SentiWordNet 3.0 (Esuli and Sebastiani, 2006).

SentiWordNet is a lexical resource where each synset of the WordNet has an additional field of sentiment/polarity information associated with it. Polarity information includes polarity labels(positive, negative, neutral) with corresponding scores describing how positive, negative or neutral a given synsets is. These scores of the single synset range from 0.0 to 1.0 and its total sum should be equal to 1.

We know that web content is enriched with English data. But, in recent times, an observation have been made that non-English data are increasing at an exponential rate. Such content also contributes largely in decision-making. Hence, the need to perform text processing on such content to generate valuable information from it.

Konkani language belongs to non-English language category. It is the official language of the state Goa and also it is a part of Indo-Aryan Languages. It is very difficult task to perform Sentiment Analysis on the text, document or article present in the Konkani language due to lack of resource availability. Therefore, to perform Sentiment Analysis for the Konkani language, there is a need to develop resources required for it.

So far no work has been performed in the field of a Sentiment Analysis for the Konkani language. Therefore, the

attempt is made to build Konkani SentiWordNet, which is a very useful resource for the Lexicon-based Sentiment Analysis. Another reason of building Konkani SentiWordNet is; to extend existing Konkani WordNet¹ where lexical relations for the Konkani WordNet can be developed using polarity(positive and negative) information of each synset. The present work is about generating sentiment lexicon for Konkani language named Konkani SentiwordNet using the Supervised Learning Approach. In this approach, we use Support Vector Machine (SVM) as a Supervised Learning Algorithm for the data classification and prediction. To implement an SVM Algorithm, training and testing datasets are very much essential and hence, to generate this required training dataset we use a Synset Projection Approach and to generate testing dataset we use human annotator.

Once training dataset is obtained from the Synset Projection approach, it is manually verified by a human annotator. In Synset Projection Approach, IndoWordNet by (Bhattacharya, 2010) and Hindi SentiWordNet by (Joshi et al., 2010) are two main resources which play the key role in training set generation task.

IndoWordNet is a knowledge base where most of the Indian language WordNets are linked to each other using unique synset identification number called as synset id of each synset.

In this paper, our main contribution is generating a training set using Synset Projection Approach, manual verification of training data, training an SVM model using the obtained training dataset and passing the human annotated testing data to it, where the SVM model makes prediction of polarity class labels for each synset given in the testing file. By following this procedure we are building a Konkani SentiWordNet i.e. sentiment lexicon for Sentiment Analy-

¹<http://konkaniwordnet.unigoa.ac.in/>

sis. Evaluation of SVM Model prediction accuracy is being carried out using the testing dataset. In evaluation task, predicted synset polarity labels by the SVM Model are compared with a human annotated synset polarity class labels and the model efficiency is calculated using precision, recall, F-score measure and accuracy.

Synset Projection Approach is used in the creation of a Hindi SentiWordNet by (Joshi et al., 2010) where it is mentioned that the synset coverage of H-SWN is 10 percent of the English SentiWordNet as the IndoWordNet linking task is still in progress. This is the second reason; we are using Synset Projection Approach in the creation of a training dataset for the Konkani language rather than using it as an approach for building a Konkani SentiWordNet.

2. Related Work

As described in (Das and Bandyopadhyay, 2010), till date, a SentiWordNet is being developed for English, Hindi, Telugu and Bengali languages. In (Das and Bandyopadhyay, 2010) paper, a game called Dr. Sentiment has been introduced in order to create SentiWordNet for Hindi, Telugu and Bengali languages. At present using online game approach, Bengali SentiWordNet contains 20,546 entries, Hindi SentiWordNet contains 13,889 and Telugu SentiWordNet contains 10,204 unique entries. (Esuli and Sebastiani, 2006) created an English SentiWordNet using Semi-Supervised approach, where it contains overall $\sim 1,17,684$ synsets. Here, glosses of each synset are properly analyzed and processed in order to perform Semi-Supervised synset classification.

One of the examples is being taken from the English SentiWordNet², where *pretty#1* is an instant (synset) of the English SentiWordNet along with its concept and polarity scores are as given follow:

pretty#1 pleasing by delicacy or grace; not imposing; "pretty girl"; "pretty song"; "pretty room", Positive score (*pretty#1*) = 0.875, Negative score (*pretty#1*) = 0.125 and Neutral score (*pretty#1*) = 0.0 and total sum of the scores is (0.875+0.125+0.0) = 1.0.



Figure 1: Visualisation of synset *pretty#1* in English SentiWordNet.

Hindi SentiWordNet (H-SWN) developed at IIT-Bombay using two existing lexical resources, they are English-Hindi WordNet linking by (Karthikeyan and Arun, 2010) and SentiWordNet of the English language by (Esuli and Sebastiani, 2006). The overall synset coverage of the H-SWN is

~ 16000 , which is just 10 percent of the English SentiWordNet. This approach is highly dependent on Hindi-English WordNet linkage (IndoWordNet), where this linking task is still under progress as mentioned in (Joshi et al., 2010).

3. Need For a Konkani SentiWordNet

- As of now, no attempt being made to work for a Konkani language in the field of Sentiment Analysis. On the other hand, the English language is far ahead in this field. Therefore to begin with the new language Lexicon- based Sentiment Analysis is most preferable. But, so far no sentiment lexicon is created for Konkani language and hence, there is a need to develop a SentiWordNet (lexicon) for the Konkani language.
- Such resources are also useful in the task of a code mixed data (Barman et al., 2014) Sentiment Analysis.

4. Approach used

This paper mainly focuses on the creation of a Konkani SentiWordNet using the Supervised Learning Approach. As SVM is the Supervised Learning Algorithm and Konkani being the new language, there is a need to create the training and testing datasets from scratch. The training and testing datasets are used to train and test the SVM algorithm.

4.1. Generating a Training Dataset

Synset Projection Approach is used to generate the training set. This section describes the steps undertaken to generate training dataset as follows:

- Projecting synsets from the Hindi SentiWordNet to the Konkani synset file along with their polarity labels by using Synset Projection Approach is shown diagrammatically in figure 2.
 - In the first step, a synset is extracted from a Hindi SentiWordNet along with its corresponding polarity labels, synset id and polarity scores.
 - Since, Konkani WordNet and Hindi WordNet are linked to each other using common synset id.
 - Search is made with the help of the synset id in a Konkani WordNet to find whether entry of corresponding extracted synset is present in it or not.
 - If an entry of a synset is not found then, it is discarded.
 - If an entry of a synset is found in a Konkani WordNet then, the same synset from a Hindi SentiWordNet, along with its sentiment polarity labels are projected to the Konkani synset file.
- Discarded synsets which are absent in the Konkani WordNet but present in Hindi WordNet are stored in the file so that later on, it can be added to Konkani WordNet.
- Konkani synset file contains a list of synsets which have prior assigned three polarity labels such as positive, negative and neutral (also called as an objective

²<http://sentiwordnet.isti.cnr.it/search.php?q=pretty>

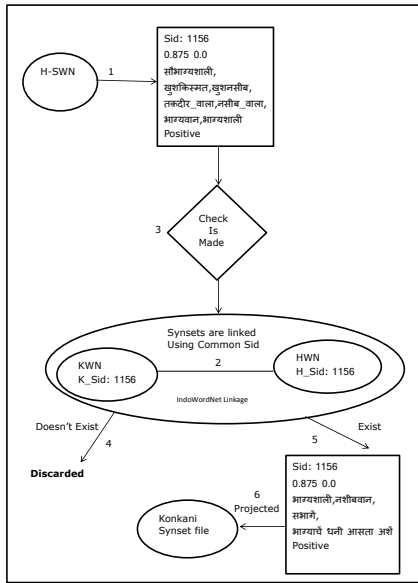


Figure 2: flow diagram of Synset Projection Approach.

label). There are total 2920 synset entries in Konkani synset file with four POS categories. Obtained results are depicted in Table 1.

POS Category	Number of Synsets
Adjectives	1293
Adverbs	65
Verbs	368
Nouns	1194
Total No. of Synsets	2920

Table 1: Statistics of Konkani synset file along with its POS categories.

- In the first step, we are concern about only binary classification i.e. a given synset has a positive or negative label. Hence, we extract only those synsets which have either positive or negative labels from the Konkani synset file. The count of positive, negative, and neutral synsets from the Konkani synset file is given in table 2.

Polarity labels	Number of Synsets
Positive	160
Negative	209
Neutral	2551
Total No. of Synsets	2920

Table 2: Count of positive, negative and neutral synset in a Konkani synset file

- Then, the obtained positive and negative synsets are given to the human annotator for verification and results are as follows:

- Out of 160 positive synsets, the annotator detected 18 negative, 1 redundant while remaining as positive synsets.
- Out of 209 negative synsets, the annotator detected 26 positive and 183 negative synsets.
- Now, 26 positive synsets are added to positive synset set containing 141 positive synset entries and 18 negative synsets are added to negative synset set containing 183 negative synset entries.
- Total estimation count of positive and negative synsets after manual verification and correction is given in table 3

Total no. of positive synsets	141+26 = 167
Total no. of negative synsets	183+18 = 201

Table 3: Estimation count of positive and negative synsets after manual verification and correction

- After manual verification and correction of positive and negative synsets, 167 positive and 167 negative synsets are kept for training an SVM model. The reason behind keeping 167 negative synsets for the training rather than 201 negative synsets is; in the training dataset, the proportion of both positive and negative synsets must be same to get fair results.
- Therefore, the training set contains 334 synset entries along with their polarity labels +1 or -1.
- Next, each synset from training set is replaced by its corresponding concept and examples using Konkani WordNet API³

4.2. Generating a Testing Dataset

Testing dataset is created manually by assigning sentiment polarity labels to 80 synsets. Among which 23 are positive and 57 are negative. This dataset is required, to check whether a trained SVM model gives a correct polarity label to each synset from the testing dataset or not. Before giving the test data to SVM model, all synsets are replaced by gloss and examples of the corresponding synset. Then the textual content of testing data is converted to numerical content. Further, same preprocessing steps are followed as training dataset.

4.3. Getting training and testing data into SVM data format

Initially, the content of the training and test dataset is present in the textual form. The training dataset contains 334 synset entries and test dataset contain 80 synset entries. The format of data(training/testing) once all synsets are replaced by its corresponding gloss/concept and examples looks like as follows:

```
< polarity label -1 or 1> <concept> <examples of synset 1>
< polarity label -1 or 1> <concept> <examples of synset 2>
```

³<http://indradhanush.unigoa.ac.in>

<polarity label -1 or 1> <concept> <examples of synset n>

We are using Libsvm tool⁴ for the classification and prediction of polarity class label for the given synset. Libsvm tool accepts the training or the testing data as an input if only if data is present in the particular format. This format is obtained using following steps.

- Creating a vocabulary
 - In this step, unique words from overall available data (training and testing) are fetched and stored in the vocabulary text file.
- Generating a document-term matrix for each sentence which is present in the obtained training and testing dataset.
 - In this matrix, data representation is done in the following way. Here, numerical data representation is shown for two textual sentences:
 +1 1:2 0:1 4:1 9:1
 -1 0:1 7:1 6:1 9:1
 +1 and -1 represents class labels i.e. positive or negative.
 <Index value of a word in the vocabulary from a sentence > : <number of times a word occurs in the sentence i.e. frequency count of a word in the sentence>
 In this manner both testing and training data are represented in a document-term matrix format.
- Sorting index values of each word from a sentence in the ascending order.
 - An example is given below for two sentences:
 +1 0:1 1:2 4:1 9:1
 -1 0:1 6:1 7:1 9:1

4.4. Training an SVM Model

Support Vector Machine (SVM) is one of the Supervised learning algorithms. Given a dataset, it does classification of data into two classes by drawing hyperplane between the data points in such a manner that it always try to maximize the margin. Here, we use positive and negative polarity class labels.

SVM training is performed using Libsvm packages(Chang et al., 2011). Libsvm uses Radial Basis Function (RBF) kernel by default for the classification. It is also named Gaussian kernel. The overall flow of the proposed approach is shown in figure 3.

4.5. Experimental Results

We give human annotated testing data to the trained SVM model, where it does the prediction for each synset present in the testing dataset. Based on the SVM model predicted class labels and human annotated class labels, SVM model

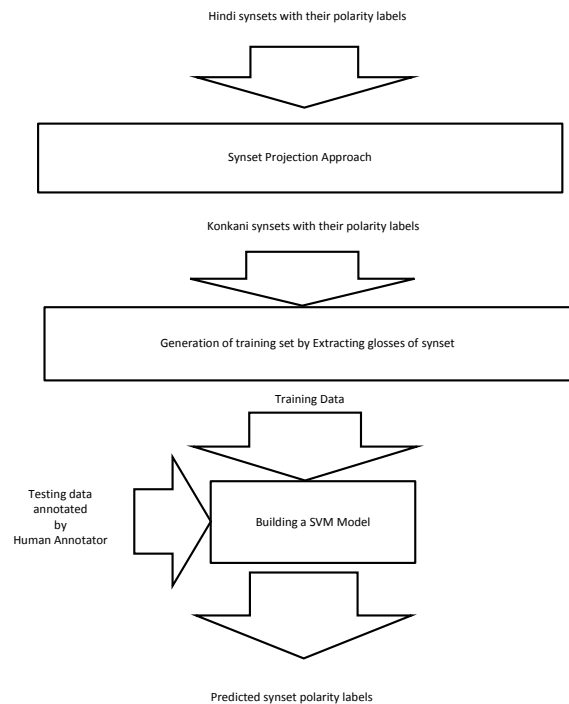


Figure 3: Flow diagram of Proposed System.

efficiency is calculated using following parameters such as precision, recall, f-score and accuracy. Results of the experiment are depicted in table 3.

Parameters used for the measure	Scores
True Positive	22
True Negative	16
False Positive	41
False Negative	1
Precision Rate	0.349
Recall Rate	0.9565
F-Score	0.5114
Accuracy	0.475

Table 4: Experimental results to check the SVM model accuracy

4.5.1. Key Observation

The SVM model evaluation is performed using two parameters namely "F-score measure" and "accuracy" where, it is being observed that to obtain a good F-score measure along with good accuracy, a more training data is needed to train the SVM model.

5. Conclusion and Future Work

In this paper, we present the Konkani SentiWordNet by using a Supervised Learning Algorithm where we use Synset Projection Approach for generating a training dataset.

⁴<http://www.csie.ntu.edu.tw/~cjlin/libsvm>

To generate testing dataset we use human annotator who does manual annotation. The two main reasons behind using the proposed approach are:

- The H-SWN creation approach depends on the English-Hindi WordNet Linking task, which is still in progress. Therefore, we use this approach to get training dataset ready for the Konkani language.
- In the E-SWN creation approach, a training dataset is created using synset lexical relations, which are present in the English WordNet but, not yet developed in the Konkani WordNet.

This proposed approach gives accuracy 0.475 and 0.5114 F-Score measure. Based on these outcomes we conclude that there is a need for a more training data for the further improvement of F-Score measure and accuracy.

6. Bibliographical References

- Barman, U., Das, A., Wagner, J., and Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *ACL14*.
- Bhattacharya, P. (2010). Indowordnet. In *LREC10*.
- Chang, Chih-Chung, Lin, and Chih-Jen. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.
- Das, A. and Bandyopadhyay. (2010). Sentiwordnet for indian languages. In *In the 8th Workshop on Asian Language Resources (ALR), COLING 2010.*, pages 56–63, August, Beijing, China.
- Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC06*, Rome, Italy.
- Joshi, A., Balamurali, and Bhattacharyya, P. (2010). Fall-back strategy for sentiment analysis in hindi: a case study. Dept. of Computer and Science Engineering, IITB-Monash Research Academy, IIT Bombay.
- Karthikeyan and Arun. (2010). *Hindi English WordNet linkage*. Dual degree thesis, Dept. of Computer and Science Engineering, IIT Bombay.
- Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). *Aspect based sentiment analysis*. Denver, Colorado.