Resource Creation Towards Automated Sentiment Analysis in Telugu (a low resource language) and Integrating Multiple Domain Sources to Enhance Sentiment Prediction

Gangula Rama Rohit Reddy, Radhika Mamidi
LTRC IIIT Hyderabad, LTRC IIIT Hyderabad
Hyderabad Telangana India, Hyderabad Telangana India
ramarohitreddy.g@research.iiit.ac.in, radhika.mamidi@iiit.ac.in

Abstract

Understanding the polarity or sentiment of a text is an important task in many application scenarios. Sentiment Analysis of a text can be used to answer various questions such as election prediction, favouredness towards any product etc. But the sentiment analysis task becomes challenging when it comes to low resource languages because the basis of learning sentiment classifiers are annotated datasets and annotated datasets for non-English texts hardly exists. So for the development of sentiment classifiers in Telugu, we have created corpora "Sentiraama" for different domains like movie reviews, song lyrics, product reviews and book reviews in Telugu language with the text written in Telugu script. In this paper, we describe the process of creating the corpora and assigning polarities to them. After the creation of corpora, we trained the classifiers that yields good classification results. Typically a sentiment classifier is trained using data from the same domain it is intended to be tested on. But there may not be sufficient data available in the same domain and additionally using data from multiple sources and domains may help in creating a more generalized sentiment classifier which can be applied to multiple domains. So to create this generalized classifier, we used the sentiment data from the above corpus from different domains. We first tested the performance of sentiment analysis models built using single data source for both in-domain and cross-domain classification. Later, we built sentiment model using data samples from multiple domains and then tested the performance of the models based on their classification. Finally, we compared all the three approaches based on the performance of the models and discussed the best approach for sentiment analysis.

Keywords: Sentiment analysis, in-domain, cross-domain, machine learning, reviews, support vector machine, bernoulli naive bayes.

1. Introduction

With the rapid increase of textual content on the internet, efficient text processing is very important for various applications. With the advancement of machine learning approaches, the issue of processing can be addressed to a decent level (Hirschberg and Manning, 2015). Automated sentiment analysis is one of the important research topics. For example: Sentiment analysis is very useful in social media monitoring as it allows us to gain an overview of the wider public opinion. Most of the sentiment analysis approaches use supervised machine learning algorithms or expert-defined lexicons.

The automated sentiment analysis is a challenging task because of the natural language processing overheads like intentions of the author and the sentiment of the text can change depending on the situation. It becomes a more challenging task for non-English texts because of the unavailability of annotated data sets. In this paper, we explain how to create suitable corpora for supervised machine learning approaches i.e how to extract data from various sources and annotate them using an appropriate method.

We created a corpus "Sentiraama" for multiple domains like movie reviews, song lyrics, product reviews and book reviews in Telugu. Telugu is an agglutinative Dravidian language spoken widely in India. It is the third most popular language in India after Hindi and Bengali. According to Ethnologue[1] list of most spoken languages worldwide, Telugu ranks fifteenth in the list, and a total of 85 million Telugu native speakers exist across the world. After creating the data, we performed sentiment analysis on the data available using supervised machine learning approaches in three different ways: **In-domain, Cross-domain, Generalized** i.e classifier is trained using data from multiple domains. We, then presented the performance of the classifier models created and discussed which approach would be better to increase the accuracy or performance. To our knowledge, this is the first work in Telugu sentiment analysis at document level.

The paper is structured as follows. In section 2, we presented the related work , in section 3, we discussed the challenges in creating corpus, in section 4, we described the creation of corpus in detail, in section 5, we discussed the statistics and experiences in building classifiers and finally, we presented concluding remarks in section 6 and discussed the ideas for further improvement of automated sentiment analysis.

[1]https://www.ethnologue.com/statistics/size

2. Related Work

2.1 Sentiment Analysis in English

Sentiment analysis systems have been applied to many different kinds of texts including customer reviews (Liu, 2015; Hu and Liu, 2004; McGuinness and Ferguson, 2004), newspaper headlines (Bellegarda, 2013), blogs (Neviarouskaya et al.,2011), novels (Boucouvalas, 2002), emails (Mohammad and Yang, 2013). Often these systems have to cater to the specific needs of the text such as formality versus informality, length of utterances, etc.

2.2 Sentiment Analysis in Telugu

Sentiment analysis of Telugu social media texts has several challenges. Telugu is a morphologically complex

language. Very little work is done on sentiment analysis in Telugu. Sentiment analysis systems have been applied to different kinds of Telugu texts including song lyrics (Abburi et al., 2016), News (Mukku et al., 2016; Naidu et al., 2018).

3. Corpus Requirements and Challenges in Creating Corpus

To train a sentiment classifier for texts, an appropriate dataset is required. Here we present a method for its creation.

Firstly, we identified the scenarios as follows:

(1) We discussed the task of building a sentiment classifier to analyze the sentiment of Telugu songs from its lyrics. The classifier should support various movie songs. The sentiment classification of song lyrics is especially challenging because songs may not contain any of the subjectivity clues in a general subjectivity lexicon, yet express positive or negative emotions.

(2) We discussed the task of building sentiment classifiers to analyze reviews in their respective domains like movie, product and book reviews. The main challenges in this are the usage of colloquial language and a large number of spelling mistakes and non grammatical constructions of sentences in the reviews.

We observe that in each domain the requirements differ a lot with respect to sentiment analysis.

(1) Songs generally reflect a person's emotion at a particular situation in a movie and the lyrics of the song play a key role in carrying that emotion. Automated sentiment analysis should be able to classify the emotion of the person in that situation. The corpus is annotated at document level with two sentiment labels: positive and negative. The sentiment annotations reflect how the emotion of a song is percieved by the people.

(2)Where as in the case of reviews of objects in multiple domains like movies, books and products, the common thing in all of these is the user opinion. We decided to create a corpus with two sentiment labels: positive and negative in each domain because each review is about the user liking or disliking the object. The corpus is annotated at document level.

4. Creating the Telugu Corpus

In order to apply a machine learning approach, a corpus matching the requirements of the scenario is needed. Since there is no corpus available, we created a new corpus "Sentiraama". We identified the necessary steps for the annotation process which include:

1) Deciding the sentiment definition and formulating detailed instructions for annotation.

2) Deciding classification type and document source.

3) Annotating the documents based on the formulated instructions.

In the following paragraphs, we present our sentiment definition and describe the corpus creation procedure.

4.1 Sentiment Definition and Annotation Procedure

Two annotators annotated all the dataset items in all the domains i.e song lyrics, movie reviews, product reviews and book reviews using a 2-value scale, distinguishing between positive and negative based on the specific procedure to each domain as mentioned below. After multiple meetings and discussions with them, a kappa score of 0.9 is achieved.

4.1.1 Telugu Song Lyrics

In order to gather a dataset of unique song lyrics, we mined song lyrics from two websites viz. a2zsonglyrics24.blogspot.com and telugulyrics.org. After mining the lyrics, we cleaned them of html tags and other extraneous text. This created a dataset of 339 Telugu song lyrics. Due to lack of resources and not many lyrics in Telugu script are available, we could mine very less songs. For each song in our dataset, annotators first went through the lyrics and annotated them. But annotating only based on lyrics would be misleading as it depends on the situational context in the movie. So, annotators learned the situation of the song in the movie and corrected the wrong annotations. Finally, to make the dataset completely error free, we cross checked the annotations by keeping the count of number of positive and negative subjectivity clues that occured in the top tags of the songs.

Corpus Statistics

As shown in Figure 1, the amount of positive songs present in Telugu are higher than those of negative songs. So when we collected the data , it got reflected. We were able to get 230 positive songs and 109 negative songs whose lyrics are present in Telugu script as shown in Table 1.

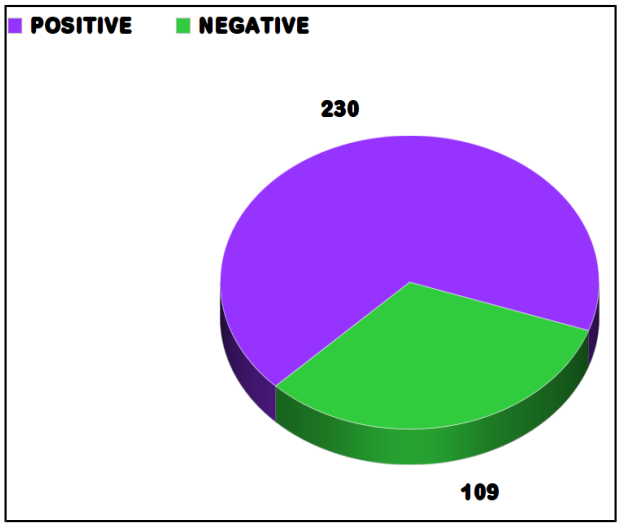

Figure 1: Pie diagram of song lyrics data.

Songs	Documents	Sentences	Words
Positive	230	4500	22500
Negative	109	2180	10900
Total	339	6680	33400

Table 1: Songs Corpus Statistics

Movie_Reviews	Documents	Sentences	Words
Positive	136	6041	60410
Negative	131	3959	39590
Total	267	10000	100000

Table 2: Movie_Reviews Corpus Statistics

4.1.2 Movie Reviews

Sentiment analyis in movie review is important because it helps in understanding how a movie is percieved by the viewer. But the very same challenge i.e the low availability of Telugu movie reviews arises even here. We tried to scrape most of the data available from different sites like tupaki.com , telugu.samayam.com and created a dataset with 267 movie reviews of more than a total of 10000 sentences. All the datasets were annotated in the following way:

The movies rated above 2.5 out of 5 by the reviewer are annotated as positive and less than 2.5 as negative. When the movie is rated as 2.5 in the review, we annotated it based on the last one line summarisation by the reviewer as it removes ambiguity. A recheck is made by going through the entire review and its annotation. Surprisingly, we found that no errors were found in the annotation made through this method.

Corpus Statistics

Here we got nearly the same amount of positive and negative reviews but the number of reviews were less. We were able to get 136 positive reviews and 131 negative reviews. The entire data consists of 10000 sentences. Though the number of reviews were less, it has decent amount of sentences. The advantage of having equal amount of data of both the types is that it prevents classification system from learning biases inherent in the dataset. Table 2 shows the corpus statistics.

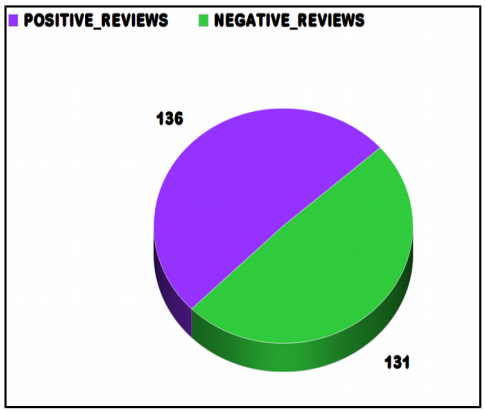

Figure 2: Pie diagram of movie reviews data.

4.1.3 Book Reviews and Product Reviews

The method of creation and procedure of annotation is the same for both of these i.e book reviews and product reviews. All the dataset items were annotated using a 2-value scale, distinguishing between positive and negative reviews. In all these cases, a review is marked as positive if a person or user who reviews is satisfied with it else it is marked as negative. But the main problem here is data gathering because we could find only 20-30 reviews written in Telugu language. A classifier model cannot be build with a such small data, so we decided to translate reviews from English to Telugu in order to get more data. The major challenge is that we should not lose the emotion conveyed by the user while translating. We tried using Google translate for the translation process but after translation the result was very bad as both the meaning and emotion conveyed were lost. So the reviews were manually translated by two experienced translators.

As shown in Figure 3, the translation was carried out in the following way: First the data was normalised and all the spelling mistakes were corrected. All the abbrevations were expanded so that the classifier could treat both abbrevation and its expansion as the same word. The numbers were retained as in Roman script. The translators were instructed to be faithful to the original text as much as possible and retain the same sentimental value. First they got the exact meaning and emotion of the review. Later they generate the review in Telugu such that it fits Telugu grammar and syntax and also carries same emotion. Though this process is accurate, it is time consuming.

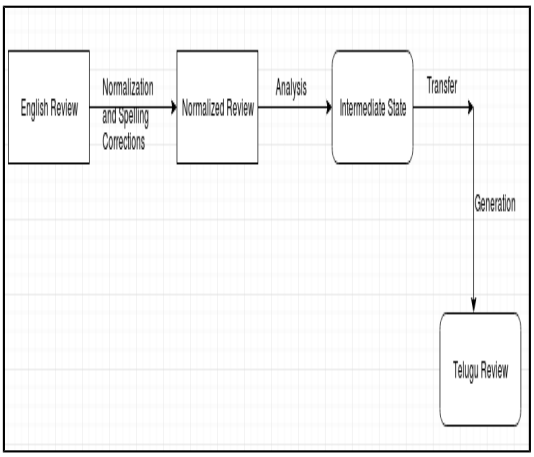

Figure 3: Translation Procedure.

In this way we were able to get 200 book reviews and 200 product reviews in each of which half of them are positive and rest of the half is negative. The annotation procedure here is going through the entire review and if the person is satisfied with the product and rated it above 2.5 out of 5 then review is annotated as positive. If the person is not satisfied with the product and rated it below 2.5 then the review is annotated as negative.

Corpus Statistics

It is necessary to equalize the dataset so that there is an equal number of positive and negative reviews. This prevents classification systems from learning biases inherent in the dataset. As shown in Figure 4, our dataset consists of 100 positive and 100 negative reviews in both the domains i.e book reviews and product reviews. Table 3 shows corpus statistics of books and Table 4 shows that of products.

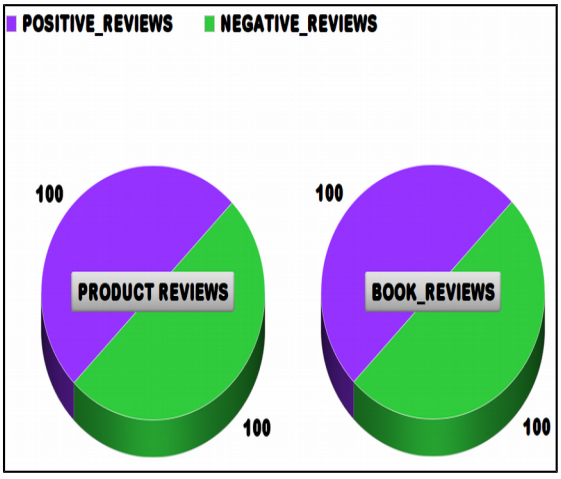

Figure 4: Pie diagrams of book reviews and product reviews.

Book_Review	Documents	Sentences	Words
Positive	100	1340	7001
Negative	100	2000	8030
Total	200	3340	15031

Table 3: Book_Reviews Corpus Statistics

Product_Review	Documents	Sentences	Words
Positive	100	2052	17390
Negative	100	2305	20104
Total	200	4357	37494

Table 4: Product_Reviews Corpus Statistics

5. Building and Analyzing Classifiers

5.1 Classifiers

We employ Bernoulli Naive Bayes as it has been found to perform well in text-related domain (Rish Irina, 2001). We also employ SVM as it gives good accuracy (Joachims T. , 1998). All of these are implemented using the scikit learn toolkit[2]. We evaluated our model applying the 10-fold cross-validation.

In Naive Bayes for text classification, the instance is assigned to the class which has the highest conditional probability of P(C|X), where C is the sentiment and X is the set of words for that instance.

In SVM, each data item is plotted as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, we performed classification by finding the hyper-plane that differentiate the classes very well as shown in Figure 5.

[2]http://scikit-learn.org/stable/supervised_learning.html

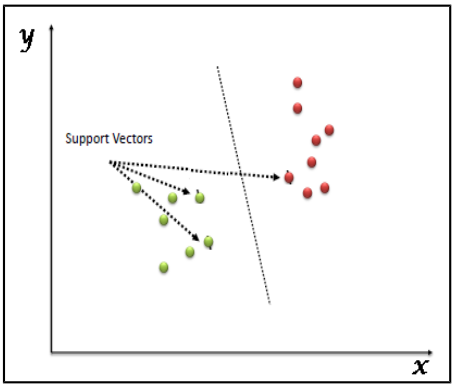

Figure 5: Support Vector Machines model.

5.2 Experimental Setup

We created four different datasets belonging to different domains. Now we use these datasets for training the classifier. We used Scikit learn[3] framework to learn a classifier based on these datasets. We first performed the in-domain sentiment analysis i.e both the training and the testing data will be from the same domain and then we evaluated the performance of the classifiers of each domain. Later we performed a cross-domain sentiment analysis i.e the training and the testing data are from different domains and evaluated the performance of the classifiers.

Finally we created a generalized classifier which is trained on data samples from all the domains and evaluated its performance.

5.3 Results

5.3.1 Single Source Data Sets(Same domain or in-domain)

Table 5 shows the results which we achieved using 10-fold cross-validation on the corpus using Bernoulli Naive Bayes classifier. Table 6 shows the results, we achieved using 10-fold cross-validation on the corpus using support vector machine.

The performance of both the classifiers was good with almost same accuracy with slight variations in precision and recall.

Thus when trained and evaluated on same domain the classifiers performed well but the problem arises when data is very sparse because the training data for classifier would be low which would affect accuracy. So we tried other approaches namely cross-domain and generalized approach.

[3]http://scikit-learn.org/stable/supervised_learning.html

Dataset	Precision	Recall	f1_score
Song Lyrics	Pos-100% Neg-75%	Pos-70% Neg-100%	84.03%
Movie Reviews	Pos-86% Neg-82%	Pos-77% Neg-89%	83.25%
Product Reviews	Pos-85% Neg-85%	Pos-85% Neg-85%	85%
Book Reviews	Pos-84% Neg-84%	Pos-84% Neg-84%	84%

Table 5: Results achieved by Bernoulli Naive Bayes classifier learned on the song lyrics, movies, products and book reviews evaluated using 10-fold cross-validation.

Dataset	Precision	Recall	f1_score
Song Lyrics	Pos-82% Neg-88%	Pos-90% Neg-78%	84.25%
Movie Reviews	Pos-88% Neg-79%	Pos-80% Neg-87%	83.3%
Product Reviews	Pos-91% Neg-80%	Pos-77% Neg-92%	84.499%
Book Reviews	Pos-88% Neg-81%	Pos-79% Neg-89%	84.034%

Table 6: Results achieved by Support Vector Machine classifier learned on the song lyrics, movies, products and book reviews evaluated using 10-fold cross-validation.

5.3.2 Cross-domain approach

In this method, we trained the classifies using datasets from one domain and tested its performance on other domains. Tables 7 shows the results, we achieved in cross-domain approach using Bernoulli Naive Bayes classifier on the corpus.

Training Dataset	Test Dataset	f1_score
Song lyrics	Movie reviews	49.5%
Song lyrics	Product reviews	53%
Song lyrics	Book reviews	50%
Movie reviews	Song lyrics	51.1%
Movie reviews	Product reviews	63.4%
Movie reviews	Book reviews	67.4%
Product reviews	Movie reviews	64.94%
Product reviews	Song lyrics	51.3%
Product reviews	Book reviews	76.8%
Book reviews	Movie reviews	61.32%
Book reviews	Product reviews	72.6%
Book reviews	Song lyrics	47%

Table 7: Results achieved by Bernoulli Naive Bayes classifier learned using cross-domain approach.

Training Dataset	Test Dataset	f1_score
Song lyrics	Movie reviews	52.2%
Song lyrics	Product reviews	60.9%
Song lyrics	Book reviews	56.7%
Movie reviews	Song lyrics	50%
Movie reviews	Product reviews	52.2%
Movie reviews	Book reviews	45.5%
Product reviews	Movie reviews	51.5%
Product reviews	Song lyrics	54.1%
Product reviews	Book reviews	50.5%
Book reviews	Movie reviews	46.6%
Book reviews	Product reviews	61.9%
Book reviews	Song lyrics	49%

Table 8: Results achieved by Support Vector Machine classifier learned using cross-domain approach.

We can clearly see in Table 7 and 8, a huge drop in f1 scores. Performing cross-domain analysis with classifiers trained performed significantly worse than in-domain classification.

5.3.3 Generalized Approach:

This section presents results for classifiers trained from a combination of datasets from all the domains. The results for using the combined approach can be found in Table 9. The dataset size refers to the number of instances found in the combined dataset. We used three training datasets namely

Case-1)Training set contains 80% of dataset from each domain.

Case-2)Training set contains 80% of dataset from all domains except the domain which is to be tested on, only 50% of that dataset is taken for training and other part will be used as test data.

Case-3)Training set contains 80% of dataset from all domains except the domain which is to be tested on, only 20% of that dataset is taken for training and other part will be used as test data.

Using the combined data set yields similar performance to using data set from same domain. But we can see a small

increase in accuracies in generalized approach.This is for case one. Where as from case-2 and case-3 we can say that even if large data from a domain is not available to train, we can generate good results using this generalized classifier. But the f1score would be little low compared to in-domain approach yet it is much better than that of cross-domain approach. So a generalized classifier here even solves the problem of sparsity to certain extent.

	Case1- 80%	Case2- 50%	Case3- 20%
Movie Reviews	Precision-86 Recall-86% f1_score-86%	Precision-84 Recall-83% f1_score-83.497%	Precision-81 Recall-81% f1_score-81%
Song lyrics	Precision-86 Recall-85% f1_score-85.497%	Precision-70 Recall-72% f1_score-70.98%	Precision-66 Recall-68% f1_score-66.985%
Product Reviews	Precision-87 Recall-87% f1_score-87%	Precision-82 Recall-81% f1_score-81.497%	Precision-79 Recall-78% f1_score-78.497%
Book Reviews	Precision-87 Recall-86% f1_score-86.497%	Precision-77 Recall-77% f1_score-77%	Precision-76 Recall-76% f1_score-76%

Table 9: Results achieved by Support Vector Machine classifier learned using Generalized approach.

5.3.4 Comparision

From the above results a commonality which we observe is that generalized classifier is the top group for evaluating various domain datasets, unlike classifiers trained from single domain, which performs better only in its domain and gives worst results for other domains. Generalized classifier is significantly better than classifier generated using cross-domain data sets and also better than in-domain classifier. Also it would gives little less but better accuracies when trained on small data than that of large data. So one classifier for all domains in a language would be a good idea than building a classifier for each domain.

6. Conclusion and Future Work

In this study, we presented a method of resource creation for sentiment analyis and a formal procedure to annotate them. We also set out to determine the performance of multi-domain sentiment analysis using data from all the domains available. We performed in-domain , cross-domain and generalized approaches and evaluated classifier's performance in each of these. We found out that using generalized (multi-source) sentiment classification would yield better results than that of in-domain and cross-domain classification.

6.1 Future Work

Future work will involve extending our methodology for sentiment classification from document level to aspect level and entity level. This would be very useful in understanding which features of products are liked by the users and which are not liked by them. This would result in the production of better products.

Using the dataset ''Sentiraama'' that we created, many machine learning algorithms and lexicon based sentiment analysis can be applied and performance can be measured and also improvements can be made.

We are also planning to work on political bias in Telugu newspapers based on this work of domain adaptation.

7. Acknowledgements

We would like to thank the annotators and translators.
This work is supported by KCIS project no: LTRC-CPH-KCIS-78

8. Bibliographical References

Julia-Hirschberg and Christopher D. Manning (July 2015) Advances in natural langage processing Science **349** (6425), 261-266. [doi: 10.1126/science.aaa8685]

Sandeep Sricharan Mukku, Nurendra Choudhary, and Radhika Mamidi. 2016. Enhanced sentiment classification of telugu text using ML techniques. In Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016) colocated with 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), New York City, USA, July 10, 2016., pages 29–34.

Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu, and Ramesh Kumar Mohapatra. 2018. Text summarization with automatic keyword extraction in telugu e-newspapers. In Smart Computing and Informatics, pages 555–564. Springer.

Harika Abburi, Eswar Sai Akhil Akkireddy, Suryakanth Gangashetti, and Radhika Mamidi. 2016. Multimodal sentiment analysis of telugu songs. In Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016) co-located with 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), New York City, USA, July 10, 2016., pages 48–52.

Thorsten Joachims. (April 21 - 23, 1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceeding ECML '98 Proceedings of the 10th European Conference on Machine Learning, Pages 137-142.

K. Mallareddy, Dr. (2012). Evolution of Telugu Language Teaching and Challenges to Present Curricular Trends. IOSR Journal of Humanities and Social Science. 5. 33-36. 10.9790/0837-0513336.

Heredia, Brian & Khoshgoftaar, Taghi & Prusa, Joseph & Crawford, Michael. (2016). Integrating Multiple Data Sources to Enhance Sentiment Prediction. 285-291. 10.1109/CIC.2016.046.

Lommatzsch, Andreas & Bütow, Florian & Ploch, Danuta & Albayrak, Sahin. (2017). Towards the Automatic Sentiment Analysis of German News and Forum Documents. 18-33. 10.1007/978-3-319-60447-3_2.

Rish, Irina. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell. 3. .

Bing Liu. 2015. Sentiment Analysis - Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In (McGuinness and Ferguson, 2004), pages 755–760.

Deborah L. McGuinness and George Ferguson, editors. 2004. Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA. AAAI Press / The MIT Press.

Jerome R. Bellegarda. 2013. Data-driven analysis of emotion in text using latent affective folding and embedding. Computational Intelligence, 29(3):506–526.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2011. Affect analysis model: novel rule based approach to affect sensing from text. Natural Language Engineering, 17(1):95–135.

Anthony C Boucouvalas. 2002. Real time text-to-emotion engine for expressive internet communications. In Proceedings of International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP-2002).

Saif M. Mohammad and Tony Yang. 2013. Tracking sentiment in mail: How genders differ on emotional axes. CoRR, abs/1309.6347.

9. Language Resource References

The corpus ‘‘Sentiraama’’ was created at KCIS, LTRC, IIIT-Hyderabad. It will be available at ‘‘https://ltrc.iiit.ac.in/showfile.php?filename=downloads/sentiraama/’’.