A SURVEY OF SENTIMENT CLASSIFICATION TECHNIQUES USED FOR INDIAN REGIONAL LANGUAGES

Pooja Pandey and Sharvari Govilkar
Department of Computer Engineering, University of Mumbai, PIIT, New Panvel, India

ABSTRACT

Sentiment Analysis is a natural language processing task that extracts sentiment from various text forms and classifies them according to positive, negative or neutral polarity. It analyzes emotions, feelings, and the attitude of a speaker or a writer towards a context. This paper gives comparative study of various sentiment classification techniques and also discusses in detail two main categories of sentiment classification techniques these are machine based and lexicon based. The paper also presents challenges associated with sentiment analysis along with lexical resources available.

KEYWORDS

NLP, sentiment, sentiment analysis, classification techniques, challenges, lexical resources, features, machine learning, lexicon based.

1. INTRODUCTION

Sentiment Analysis (SA) is a natural language processing task that deals with finding orientation of opinion in a piece of text with respect to a topic [1]. It deals with analyzing emotions, feelings, and the attitude of a speaker or a writer from a given piece of text. Sentiment Analysis involves capturing of user’s behaviour, likes and dislikes of an individual from the text. The main goal behind sentiment analysis is to identify sentiment associated with the text by extracting sentimental context from the text.

The purpose of sentiment analysis is to determine the attitude or inclination of a communicator through the contextual polarity of their speaking or writing. Their attitude may be reflected in their own judgment, emotional state of the subject, or the state of any emotional communication they are using to affect a reader or listener. It is trying to determine a person’s state of mind on the subject they are communicating about. This information can be mined from various data sources like: texts, tweets, blogs, social media, news articles, product comments.

There are different classification levels in SA: document-level, sentence-level and aspect-level. Document-level SA aims to classify an opinion of the whole document as expressing a positive or negative sentiment. Sentence-level SA aims to classify sentiment expressed in each sentence which involves identifying whether sentence is subjective or objective. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities which is done by identifying the entities and their aspects for instance researchers need a tool to generate summaries for deciding whether to read the entire document or not and for summarizing information searched by user on internet. News groups can use multi document summarization to cluster the information from different media and summa.
The paper presents a detail survey of various sentiment classification techniques. Related work done and past literature is discussed in section 2. Baseline algorithm is defined in section 3 along with challenges associated in performing sentiment analysis. Two main categories of sentiment classification techniques which are machine based SA and lexicon based SA are discussed in detail in section 4 along with the comparison of each method cited in Indian regional languages. Finally, section 5 concludes the paper.

2. LITERATURE SURVEY

In this section we cite the relevant past literature of research work done in the field of sentiment analysis for Indian languages.

Amitava Das and Bandopadhya developed SentiWordNet for Bengali language, which is an automatically constructed lexical resource in which WordNet synset are assigned positive and a negative score. SentiWordNet and Subjectivity Word List are used to generate merged sentiment lexicon in which duplicate words are removed. Bengali SentiWordNet is created by applying word level lexical-transfer, using an English-Bengali dictionary, on content available in SentiWordNet. [4]

Authors have [1] developed a modified approach to identify the sentiments associated with Hindi content by handling negation and discourse relation. They updated the existing Hindi SentiWordNet (HSWN) by when specific sentiment words where not found in existing HSWN by extracting same meaning word from English SentiWordNet. Through handling of negation and discourse associated with text, their proposed algorithm achieved approximately 80% accuracy on classification of Hindi reviews.

Amandeep Kaur and Vishal Gupta Proposed Algorithm for Sentiment Analysis for Punjabi Text. Their used the Hindi WordNet to develop the Subjective Lexicon for the Punjabi language.[8] They are using three popular methods used for the generation of subjective lexicon- Use of Bi-Lingual Dictionary, Machine Translation, Use of Word net. Then they devise an Algorithm Combining the unigram method and simple scoring method which provides better efficiency. The overall efficiency of the proposed algorithm is 54.2%.

Aditya Joshi and Pushpak Bhattacharyya [2] proposed a fallback strategy for finding sentiment associated with Hindi language. Machine Translation, In-language and Resource Based SA where the three approach proposed for SA in Hindi. Through WordNet linking, words present in English SentiWordNet were replaced by similar Hindi words to construct Hindi SentiWordNet (HSWN). To determine the polarity of the opinion associated in text SVM classifier was used to perform In-language SA. In a machine translation based method Google translator was used to translate Hindi corpus into English and resulted corpus was used as input to the classifier to determine polarity. In resource based SA approach the synset corresponds to the English SentiWordNet is used in the corresponding synset in Hindi to build the SentiWordNet (H-SWN) for Hindi. 78.14 was the best accuracy achieved through in-language sentiment analysis for Hindi documents.

Kishorjit and Sivaji Bandyopadhyay proposed verb based Manipuri sentiment analysis. They are using the conditional random field (CRF) approach. It is an unsupervised approach where the system learns by giving some training and can be used for testing other texts. Then they processed text for part of speech tagging using CRF. Here With the help of POS tagger the verbs of each sentence are identified and the modified lexicon of verbs is used to notify the polarity of the sentiment in the sentence, because the sentiment of the sentence is highly dependent on the verbs. Their proposed algorithm achieved approximately 75% accuracy on sentiment analysis of Manipuri. [9]
Authors have used self learning neural network which takes linguistic and part of speech emotive features as input, for detection of sentiment associated in Tamil content. The primary inputs to the neural network are; first the outputs of domain classifier which retrieves noun and verbs and term domain frequencies, second the two schmaltzy analyzers which are Negation Scorer and Flow scorer which assigns score to each document based on the pleasantness of the words in it, and lastly noun, verb and urichol the three taggers. Unsupervised learning was selected and Hebbian learning was incorporated since emotion recognition has to deal with emotional intelligence. Two dimensional animated face generator is used to show resultant emotion which is identified by assigning weights for features based on their affective influence [7].

Das and Bandopadhya [5] used different strategies to predict the sentiment of a word in the given text. In one of the strategies they annotated the words with their associated polarity manually. In another strategy, to determine the polarity of the text, Bi-Lingual dictionary for English and Indian Languages was used. In next strategy synonym and antonym relations of words in WordNet are used, to determine the polarity. Final strategy used learning from pre-annotated corpora to find the polarity of the text.

Authors have used lexicon based approach for extracting sentiment from Urdu text. Sentiment-annotated lexicon based approach for analyzing sentiment based on SentiUnits. SentiUnits are the expressions made of one or more words, which carry the sentiment information of the whole sentence. Shallow parsing is used for identification and extraction of SentiUnits from the given text. Two types of SentiUnits were used i.e. Single adjective phrase and multiple adjective phrases. Adjectives, modifiers and orientation are the three attributes associated with SentiUnits. Process of sentiment analysis is composed of three phases; pre-processing where normalization and segmentation of text was performed then next phase involves use of shallow parsing for extracting SentiUnits and finally extracted SentiUnits are compared with lexicon and their polarities are calculated for classification as positive or negative and overall polarity is obtained by combining polarity’s.[6]

Authors have built a robust sense based classifier [3] which is a supervised document level sentiment classifier on basis of semantic space based on WordNet senses. Through the use of similarity metrics unknown synset in the training set was replace by similar synset in test set. Words in the corpus were annotated with their senses using combinations of manual sense annotation and automatic iterative WSD. In synset replacement algorithm a synset encountered in a test document is not found in the training corpus, it is replaced by one of the synsets present in the training corpus. The substitute synset is determined on the basis of its similarity with the synset in the test document. The synset that is replaced is referred to as an unseen synset as it is not known to the trained model.

Authors have proposes Cross-Lingual Sentiment Analysis(CLSA) [10] using WordNet senses as features for supervised sentiment classification where machine translation is not available for translation between specific languages. Concept of linked WordNet is used which bridge the gap between those two languages. WordNet of Hindi and Marathi was developed using an expansion approach having same synset identifier The words in the corresponding synsets represent translations of each other in specific contexts. Words of the training as well as the test corpus were mapped to their WordNet synset identifiers. A classification model is learnt on the training corpus and tested on the test corpus leads to a new corpora which is represented in the common feature space i.e. sense space. Accuracy of 72% and 84% for Hindi and Marathi was obtained for sentiment classification of these languages.
3. SENTIMENT ANALYSIS

Sentiment analysis is computational study of emotions, opinions and mainly the sentiment expressed in the text by user. Sentiment analysis is a challenging task due to many challenges which are associated while processing natural language. Any sentiment analysis system needs first to extract feature i.e. sentimental words or phrases from the given text and then using suitable text classifier overall sentiment associated with the text is extracted.

3.1. Challenges for sentiment analysis

3.1.1. Contextual Information

Identifying the context of the text becomes an important challenge to address in SA. Behaviour/use of the word changes in a great aspect based on the context.

Ex-1 The journey was long.
Ex-2 Seminar was long.
Ex-3 Battery life of Nexus 5 is long.

In all the above 3 examples, meaning of long is same it indicates the duration or passage of time. In Ex-1 and Ex-2 “long” indicates bored hence a Negative expression whereas in Ex 3 “long” indicates efficiency hence a Positive expression. In Ex 3 “long” indicates efficiency hence a Positive expression.

3.1.2. Sarcasm Detection

Sarcasm involves statement and a remark which is usually indirect taunt towards any object or an appraisal in a negative way. Detecting sarcasm is a tough task for humans and equally harder for machine. Some examples of sarcasm: Ex- Amazing presentation by Mr. X, I won’t ever attend such presentation again.

3.1.3. Word Sense Disambiguation

Word sense disambiguation (WSD) is the problem of determining in which sense a word having a number of distinct senses is used in a given sentence. The same word can have multiple meanings, and based on the sense of its usage the polarity of the word also changes. [19] For example, the word "cold" has several senses and may refer to a disease, a temperature sensation, or an environmental condition. The specific sense intended is determined by the textual context in which an instance of the ambiguous word appears. In "I am taking aspirin for my cold" the disease sense is intended, in "Let's go inside, I'm cold" the temperature sensation sense is meant, while "It's cold today, only 2 degrees", implies the environmental condition sense.

3.1.4. Word Order:

Word order plays a vital role in deciding the polarity of a text, in the text same set of words with slight variations and changes in the word order affect the polarity aspect. For example “X is efficient than Y” conveys the exact opposite sentiment from “Y is efficient than X”.

3.1.5. Identify subjective portions of text:

The same word can be treated as subjective in one context and objective in some other. This makes it difficult to identify the subjective (sentiment-bearing) portions of text. Consider following examples:

Ex-1 the language of the author was very crude.
Ex-2 Crude oil is extracted from the sea beds.
3.1.6. Indirect negation of sentiment

Negation of sentiment [17] is assigned to words like no, not, never, etc. But there are certain words tend to reverse the sentiment polarity implicitly. For example the sentence, “This movie avoids all predictable and boring drama found in most of the bollywood movies.” The negative sentiment associated with words predictable and boring is reversed by associating word avoid with those words.

3.1.7. Entity Recognition

Same entity is not seen for all the texts in a document. When multiple entities are being mentioned about in a single document, the overall document polarity does not make much sense. We need to separate out the text about a particular entity and then analyze its sentiment. Ex- I hate Heat, but I like ICE.

3.2. Features for sentiment analysis

Converting a piece of text to a feature vector is the basic step in any data driven approach to SA. Some commonly used features used in Sentiment Analysis are:

3.2.1. Term Presence vs. Term Frequency:

Term frequency is mostly used in different Information Retrieval (IR) and Text Classification tasks. But Pang-Lee [11] found that term presence is more important to Sentiment analysis than term frequency as for term presence binary-valued feature vectors are used, and those vectors the entries merely indicate whether a term occurs indicated as value 1 or do not occur indicated by value 0. Polarity based classification is very useful for SA as overall sentiment may not usually be highlighted through repeated use of the same term an also occurrences of rare word can have more sentimental value compared to frequently repeated word.

3.2.2. Term Position:

More sentimental value is given to certain words based on their position in a sentence or a document. Generally words appearing in the 1st few sentences and last few sentences in a text are given more weightage than those appearing elsewhere in the document.

3.2.3. Parts of speech (POS):

Part-of-speech (POS) information is commonly exploited in sentiment analysis which deals with finding adjectives, adverbs in the text as they are important indicators of sentiment in a given text. Amongst all parts of speech like nouns, verbs, adjectives, etc most important POS is considered as adjective to represent sentimental feature or simply subjective text. Adverbs when occurs along adjectives improves the probability of finding exact sentiment of a given text.

3.2.4. Topic-Oriented Features:

Interactions between topic and sentiment play an important role in extracting sentiment from the content. For example, in a hypothetical article on Rebook, the sentences “Reebok reports that profits rose” and “Target reports that profits rose” could indicate completely different types of news regarding the subject of the document. Topic information can also be incorporated into features set as they also help in specifying sentiment.
3.2.5. Opinion words and phrases:

There are words like; good or bad, like or hate, happy or sad which are mostly used to express opinions towards an object. Also sometimes there are phrases which expresses opinions without using an of the opinion words in the phrase. For example: It cost me my full month salary.

3.3. Lexicon resources for sentiment lexicon

Lexicon resources provide [17] a polarity value of associated with words. Lexical resources can be developed either manually by humans or automatically by training machines. Manually built lexical resources tend to be more accurate but time consuming task, automatically built resources can attain much higher coverage in less time. WordNet is a lexical hierarchical database with nodes represented by word meaning instead of word itself, and relationships between different synonym synsets representing the edges between the nodes. Multiple words can be assigned to single node and a same word might be present in more than one node at a time.

3.3.1. Manually built Lexical Resources

General Inquirer consists of 11000+ words compiled from two different sources with roughly 2000 positive and 2000 negative words. Each word is tagged either as positive, negative or none. The precision with which Inquirer tags a word is very high; because of its low coverage the recall rate is very low. Bing- Liu Lexicon consists of around 2000 positive and 4000 negative words which are frequently seen in social media. Bing-Liu lexicon has advantage over General Inquirer in terms of coverage as it contains more sentiment-bearing words but has a low precision as less overall words. Subjectivity Lexicon is developed from multiple sources and sentiment which are manually tagged and contains 8000+ sentiment words, each tagged as either positive or negative. Problem with manually-tagged sentiment lexicons is that there is no concept of like mildly positive or strongly negative to classify in deep.

3.3.2. SentiWordNet

SentiWordNet is a lexical resource [17] which is an extension of WordNet. It appends sentiment information to every synset present in the WordNet. Positive, Negative and Objective score are tagged to each synset where there combined score sum up to 1.0 overall for a synset. Here synsets sentiment are not fixed to a single category i.e. the same synset has non-zero score for both the classes as it is positive in particular context and negative in another context.

3.3.3. SentiWS

SentimentWortschatz or SentiWS [15] is a German-language resource for sentiment analysis. It contains list of positive and negative sentimental words which are assigned weighted between \([-1; 1]\) and also part of speech tag and their inflections are associated. Currently SentiWS contains 1,650 negative and 1,818 positive words in the list. The list contains adjectives and adverbs which explicitly expresses a sentiment with addition of nouns and verbs.

3.3.4. WordNet Affect

WordNet Affect is a lexical resource [12] that represents the affective content of synsets by dividing them into affective categories. Thus, it gives more affective information as compared to SentiWordNet and is used when analysis to be done is with respect to emotions like sad, anger, happy, joy and others.
3.4. Baseline Algorithm

Baseline algorithm for sentiment analysis consists of tokenization, feature extraction and sentiment classification [18] using classifier such as SVM, Naive Base, etc.

Figure 1. Baseline Algorithm

We first have to tokenize the sentence which involves breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. It is done by segmenting text by splitting it by spaces and punctuation marks, and forms a bag of words. Care should be taken so that short forms such as “don’t”, “I’ll”, “she’d” will remain as one word.

Then important features are identified. Features such as Terms presence and frequency, Parts of speech (POS), Opinion words and phrases, Negations are identified. We need to take care of negations, since they will reverse polarities and decide whether we want to use only adjective, adjectives plus adverbs or simply all the words as features. Lexicon-based or statistical feature selection methods can be used to select features from documents which treat document as Bag of Words (BOW) or string. Stemming and removal of stop-words are mostly common feature selection step.

After we’ve tokenized and decided which features to use we need to classify the sentiment. Is it good or bad? Classification can be done with different algorithms. For example: Naïve Bayes, Support Vector Machines, or Max Entropy. Lexical resources like dictionary, WordNet, SentiWordNet are uses by these classifier algorithms. This technique attempts to determine whether a text is objective or subjective and whether a subjective text contains positive or negative sentiments. The system automatically collects, cluster, categorizes, and summarizes news from several sites on the web on a daily basis. A summarization machine can be viewed as a system which accepts either a single document or multiple documents or a query as an input and produces an abstract or extract summary.

4. SENTIMENT CLASSIFICATION TECHNIQUES

Sentiment classification is a task under Sentiment Analysis (SA) that tags text as positive, negative or neutral automatically. Thus, a sentiment classifier tags the sentence ‘the movie is entertaining and totally worth your money!’ in a movie review as positive with respect to the movie. On the other hand, a sentence ‘The movie is so boring that I was dozing away through the second half.’ is labelled as negative. Finally, ‘The movie is directed by Nolan’ is labelled as neutral [2]. There are two main techniques for sentiment classification: machine learning based and lexicon based. Better performance can be obtained by combining these two methods.
4.1. Machine Learning Approach

The machine learning method uses several learning algorithms to determine the sentiment by training on a known dataset. Training and a test set are the two document sets which are mostly needed for machine learning based techniques. Training set is used by classifier to understand the different characteristics associated with documents, and to check the overall performance of the classifier test set are used.

4.1.1. Supervised learning

The supervised learning methods depend on the existence of labelled training documents. Supervised learning process: two Steps; Learning (training): Learn a model using the training data testing: Test the model using unseen test data to assess the model accuracy. There are different types of supervised classifiers like: Rule-based Classifiers, Decision Tree Classifiers, Linear Classifiers and Probabilistic Classifiers.

a. Probabilistic Classifier

Mixture models for classification are used by Probabilistic classifiers where it assumes that each class of the content is a component of the particular mixture. Each mixture component can be referred as generative model which provides the probability of sampling a particular term for that component.

1. Naive Bayes Classifier (NB)

Naive Bayes classification model computes the posterior probability of a class is computed in Naive Bayes Classifier which is based on the way words are distributed in the particular document. The positions of the word in the document are not considered for classification in this model as it uses BOWs feature extraction technique. Bayes Theorem is used to predict the probability where given feature set belongs to a particular label of the content.

\[
P(\text{label} | \text{features}) = \frac{P(\text{label}) \times P(\text{features} | \text{label})}{P(\text{features})}
\]

\(P(\text{label})\) signifies the prior probability of a label. \(P(\text{features} | \text{label})\) signifies the prior probability that a particular feature set is being classified as a label. \(P(\text{features})\) specifies the prior probability that a given feature set has occurred in the process. On basis of Naive assumption i.e. all features are independent; the equation can be rewritten as:

\[
P(\text{label} | \text{features}) = \frac{P(\text{label}) \times P(f1 | \text{label}) \times \ldots \times P(fn | \text{label})}{P(\text{features})}
\]

2. Bayesian Network (BN)

Bayesian Network model is a form of directed acyclic graph in which nodes represent random variables and edges represent conditional dependencies in the graph. Complete joint probability distribution (JPD) is specified for the model as it is reckoned as complete model for the variables and their relationships. BN is not frequently used for text mining as computation complexity is very expensive.
3. Maximum Entropy Classifier (ME)

The MaxentClassifier also known as a conditional exponential classifier uses encoding to create encoded vectors by converting labelled feature sets. This encoded vector are used to calculate weights associated with each feature which can then be combined to determine the most likely label for a feature set. Main parameters for this classifier is a set of \( X \) \{weights\}, which is used to merge the joint features which are generated from a feature-set by an \( X \) \{encoding\}. Every \( C \) \{(featureset, label)\} pair is mapped to a vector through encoding.

B. Linear classifiers

Probability of a particular classification is based on a linear combination of features and their weights. A linear classifier determines which class an object belongs by making a classification decision based on the value of a linear combination of the characteristics of the objects. Feature vector are used to represent object characteristics. Different linear classifiers are available

1. Support Vector Machines Classifiers (SVM)

The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. Due to sparse nature of text the text data are well suited for SVM classification. In text data some of the features are irrelevant, but most of them seem to be correlated with one another and can be easily organized into linearly separable categories. The basic SVM has no dependency on probability as it predicts which two possible class forms an output for a given input.

2. Neural Network (NN)

Neural Network simply consists of neurons which are arranged in layers and convert a given input vector into meaningful output. Each neuron processes an input by applying nonlinear function to it and the output is passed to the next layer for further process. Mostly neural networks are designed as feed-forward network. Signals passing from one neuron to another are assigned different weights and during the training phase these weights are adjusted so that neural network adapts to a particular problem to solve. Multi-layer neural networks are used for non-linear boundaries. These multiple layers are used to create multiple piecewise linear boundaries, which are mainly used to approximate enclosed regions which belongs to a particular class of output. Here training process is more difficult as there is a need of back-propagation of errors over several layers.

C. Decision tree classifier

In decision tree classifier training data set ate hierarchical decomposed where to divide the data the condition on the attribute value is used. The condition or predicate of the attribute signifies the presence or absence of one or more words associated with it. Data space is divide recursively until certain minimum numbers of records which are used for the purpose of classification are available at leaf node.

D. Rule based classifiers

The data space is modeled with a set of different rules in rule based classifiers where the left side signifies a condition on the feature set expressed in disjunctive normal form while the right hand side represents the class label. During training phase all the rules are constructed based on different criteria used to generate rules. Support and confidence are commonly used criteria
4.1.2. Unsupervised learning

Unsupervised learning deals with finding hidden structure in unlabeled data set. There is no error or reward signal to evaluate a potential solution as examples given to the learner are unlabeled. Unsupervised learning methods are useful when there are documents to classify which are unlabeled. Nearest neighbor (KNN) is unsupervised machine learning algorithm in which objects are classified based on the majority of its nearest neighbor of the object. The class which is assigned to the object is based among its most k nearest neighbor’s object. Objects are classified based on their similarities to objects in the training data in this algorithm. Selection process is based on either majority voting or distance weighted voting.

4.2. Lexicon-based approach

The lexicon-based approach involves calculating sentiment polarity for a review using the semantic orientation of words or sentences in the review. The semantic orientation is a measure of subjectivity and opinion in text. Sentiment lexicon contains lists of words and expressions used to express people’s subjective feelings and opinions. For example, start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative. The lexicon based techniques to Sentiment analysis is unsupervised learning [13] because it does not require prior training in order to classify the data.

Manual construction, corpus-based methods and dictionary-based methods are the methods through which sentiment lexicon are constructed. The manual construction of sentiment lexicon is a difficult as it involves humans to manually assign polarities to sentimental words and it’s a time-consuming task. Dictionary based method is an iterative technique which is initially constructed manually by selecting small set of sentimental word and this set then iteratively grows by adding the synonyms and antonyms from the WordNet. This iterative process continues till no new words are reaming to be added to the seed list. The dictionary based approach have a limitation is that it can’t find opinion words with domain specific orientations. Corpus based techniques rely on syntactic patterns in large corpora. Corpus-based methods can produce opinion words with relatively high accuracy. Most of these corpus based methods need very large labelled training data but it helps to easily find domain specific opinion words and orientations of this words towards a context.

Following tables presents a comparison of sentiment classification techniques cited in Indian regional languages.
Table 1. Sentiment Analysis techniques used in Hindi text.

<table>
<thead>
<tr>
<th>Type</th>
<th>Hindi [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>250 Hindi Movie Reviews and English movie reviews</td>
</tr>
</tbody>
</table>

**Classification Technique**

<table>
<thead>
<tr>
<th>Features</th>
<th>In-Language using SVM</th>
<th>Machine Translation Based using SVM</th>
<th>Resource-based using SentiWord list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term Frequency</td>
<td>Term presence</td>
<td>TF-IDF</td>
<td>TF-IDF</td>
</tr>
<tr>
<td>Accuracy</td>
<td>74.57%</td>
<td>72.57%</td>
<td>78.14%</td>
</tr>
</tbody>
</table>

**Benefit**

MT-based systems give superior classification performance as compared to majority-based systems based on lexical resources.

**Limit**

The error of the machine translation system affects the performance of MT-based SA.

Table 2. Sentiment Analysis techniques used in Hindi text.

<table>
<thead>
<tr>
<th>Type</th>
<th>Hindi [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Hindi Movie Reviews</td>
</tr>
</tbody>
</table>

**Classification Technique**

<table>
<thead>
<tr>
<th>Features</th>
<th>Improved HindiSentiWordNet (HSWN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved HSWN</td>
<td>Improved HSWN + negation</td>
</tr>
<tr>
<td>Improved HSWN + Discourse</td>
<td>Improved HSWN + negation+</td>
</tr>
<tr>
<td>Accuracy</td>
<td>69.78%</td>
</tr>
</tbody>
</table>

**Benefit**

Increases the coverage of HindiSentiWordNet (HSWN)

**Limit**

Low accuracy for words which have dual nature.
Table 3. Sentiment Analysis techniques used in Manipuri text.

<table>
<thead>
<tr>
<th>Type</th>
<th>Manipuri [9]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Manipuri newspaper, total 2,75000 words</td>
</tr>
<tr>
<td>Classification Technique</td>
<td>Conditional Random Field (CRF).</td>
</tr>
<tr>
<td>Features</td>
<td>Part Of Speech(POS)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Recall:72.10, Precision:78.14, F-Score:75.0</td>
</tr>
<tr>
<td>Benefit</td>
<td>Model is easy to interpret</td>
</tr>
<tr>
<td>Limit</td>
<td>More methods and algorithms are to be search and implemented in order to improve the accuracy</td>
</tr>
</tbody>
</table>

Table 4. Sentiment Analysis techniques used in Punjabi text.

<table>
<thead>
<tr>
<th>Type</th>
<th>Punjabi [8]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Documents written in Punjabi language</td>
</tr>
<tr>
<td>Classification Technique</td>
<td>Subjective lexicon</td>
</tr>
<tr>
<td>Features</td>
<td>Part Of Speech(POS)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Recall:70, Precision:78, F-Score:67</td>
</tr>
<tr>
<td>Benefit</td>
<td>Better accuracy</td>
</tr>
<tr>
<td>Limit</td>
<td>Performance is low. Lexicon developed for Hindi language has limited coverage</td>
</tr>
</tbody>
</table>

Table 5. Sentiment Analysis techniques used in Tamil text.

<table>
<thead>
<tr>
<th>Type</th>
<th>Tamil [7]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Tamil News Text</td>
</tr>
<tr>
<td>Classification Technique</td>
<td>Neural network</td>
</tr>
<tr>
<td>Features</td>
<td>Part Of Speech(POS)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Precision: 60</td>
</tr>
<tr>
<td>Benefit</td>
<td>The emotion is identified by assigning weights for features based on their affective influence.</td>
</tr>
<tr>
<td>Limit</td>
<td>Local minima and over fitting</td>
</tr>
</tbody>
</table>
Table 6. Sentiment Analysis techniques used in Urdu text.

<table>
<thead>
<tr>
<th>Type</th>
<th>Urdu[6]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Movie</td>
</tr>
<tr>
<td></td>
<td>Product</td>
</tr>
<tr>
<td>Classification</td>
<td>Lexicon based using SentiUnits</td>
</tr>
<tr>
<td>Technique</td>
<td></td>
</tr>
<tr>
<td>Features</td>
<td>Part Of Speech(POS)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>72%</td>
</tr>
<tr>
<td></td>
<td>78%</td>
</tr>
<tr>
<td>Benefit</td>
<td>Achieving better results by using SentiUnits.</td>
</tr>
<tr>
<td>Limit</td>
<td>The SentiUnits with adjectives made by postposition combined with nouns, cause errors and hence, an improved algorithm is required.</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

Sentiment analysis has lead to determine the attitude or inclination of a communicator through the contextual polarity of their speaking or writing. Sentiments can be mined from texts, tweets, blogs, social media, news articles, comments or from any source of information.

Sentiment Analysis has been quite popular and has lead to building of better products, understanding user’s opinion, executing and managing of business decisions. People rely and make decisions based on reviews and opinions. This research area has provided more importance to the mass opinion instead of word-of-mouth.

Large amount of work in sentiment analysis has been done in English language, as English is a global language, but there is a need to perform sentiment analysis in other languages also. Large amount of other languages contents are available on the Web which needs to be mined to determine the sentiment.

ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks to the head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.
REFERENCES


[18] https://class.coursera.org/nlp/lecture/145


Authors

Pooja Pandey is currently a graduate student pursuing masters in Computer Engineering at PIIT, New Panvel, and University of Mumbai, India. She has received her B.E in Computer Engineering from University of RGPV. She is having 2 years of past experience in the field of teaching. Her areas of interest are Natural Language processing, theory of computation and ethical hacking.

Sharvari Govilkar is Associate professor in Computer Engineering Department, at PIIT, New Panvel, and University of Mumbai, India. She has received her M.E in Computer Engineering from University of Mumbai. Currently she is pursuing her PhD in Information Technology from University of Mumbai. She is having 17 years of experience in teaching. Her areas of interest are text mining, Natural language processing, Compiler Design & Information retrieval etc.