International Journal of Scientific Research in Computer Science and Engineering
Vol.5, Issue.6, pp.39-46, December (2017)

Research Paper

E-ISSN: 2320-7639

Sentiment Detection from Punjabi Text using Support Vector Machine

Gagandeep Kaur[1*], Kamaldeep Kaur[2]

[1*]Computer Science and Engineering, Guru Nanak Dev Engineering College, I.K. Gujral Punjab Technical University, Ludhiana, Punjab
[2] Computer Science and Engineering, Guru Nanak Dev Engineering College, I.K. Gujral Punjab Technical University, Ludhiana, Punjab

[*]*Corresponding Author: ggndeep946@gmail.com, Phone No.: 9592533843*

Available online at: www.isroset.org

Received: 10/Nov/2017, Revised: 21/Nov/2017, Accepted: 20/Dec/2017, Published: 31/Dec/2017

***Abstract*—** This paper focuses on sentiment analysis on Punjabi News Articles using Support Vector Machine. Sentiment analysis is a field of Natural Language Processing and it is the most trending field of research. Sentiment analysis on Punjabi language is to be performed because of increasing amount of Punjabi content over the web, provides an important aspect for the researchers, organizations, and governments to analyze the user-generated content and get the useful information from it. With the increase in the amount of information being communicated via regional languages like Punjabi, comes a promising opportunity of mining this information. Support Vector Machine approach is used by proposed system to classify the content. Support Vector Machine is a supervised machine learning approach that is be used for classification and regression problems. However, it is mostly used for classification problems. So there is a need to analyze the Punjabi language content and get better understanding of Punjabi text. The work focuses on detecting positive or negative sentiment from Punjabi content. The results of the proposed system depict remarkable accuracy. The accuracy of sentiment analysis on Punjabi news articles using Support vector machine is found to be 90%.

***Keywords*—** Sentiment Analysis, Natural Language Processing, Punjabi News Articles, Support Vector Machine

I. INTRODUCTION

Sentiment Analysis is an important application of Natural Language Processing. It is a process of finding people's opinion, attitude, views and detecting emotions towards any entity. The entity can be individuals, products, events or topics. These topics are mostly covered by reviews. Sentiment analysis is a process of identifying the sentiment expressed in the form of text and then analyzing it to get beneficial information [1]. Document-level, sentence-level, and aspect-level are three main classification levels. In document-level sentiment analysis aims to classify an input document as a positive or negative sentiment. The whole document is considered as an input unit in document-level sentiment analysis. Sentence-level sentiment analysis has an aim to classify sentiments expressed in each sentence and then identifying the sentence as subjective or objective. If the sentence is subjective, sentence-level sentiment analysis will determine whether the sentence contains positive or negative sentiments. Aspect-level sentiment analysis aims to classify the sentiments with respect to the specific aspects of entities and then identifying the opinion regarding entities and their aspects [2].

A. Sentiment Analysis

In Natural Language Processing, the field of sentiment analysis is a computational task for automatically detecting and classifying sentiment from text, document or from sentences to finding it's ''polarity'' or ''orientation''. The polarity of the documents can be positive, negative or neutral. Some more fine-grained in polarity including very positive, very negative or intensity levels from 1 to 5 scales have also been considered [3].

The resources, approaches, and tools to do successful research in Punjabi language are very limited so in this field, research scope is high [4]. Punjabi is an Indo Aryan language. It is the language of 130 million individuals over all around the world, that makes it the 12th most commonly spoken the language in the world. It is gaining the 9th position from commonly spoken languages in India. Punjabi as a cultural language is increasing day by day in the Indian subcontinent and its all credit goes to Bollywood. Mostly all Bollywood movies have Punjabi mixed vocabulary in their scripts, use Punjabi remix music, and few songs fully sung in Punjabi. In these days movies are incomplete without Punjabi music [5].

B. Needs of Sentiment Analysis

a) Industry Evolution

Only the useful amount of data is required in the industry as compared to the set of the complete unstructured form of the data. However, the sentiment analysis was done for extracting the important feature from the text that will be needed for benefits purpose of industry. Sentimental Analysis will provide a great opportunity to the industries for providing value to their gain value and audience for themselves. Any of the industries with the business to the consumer will get benefit from this whether it is restaurants, entertainment, hospitality, mobile customer, retail or being travel.

b) Research Demands

Another important reason that stands behind the growth of SA deals with the demand for research in evaluation, appraisals, opinion and their classification. Present solutions for the purpose of sentiment analysis and opinion mining are rapidly evolving, specifically by decreasing the amount of human effort that will be required to classify the comments. Also, the research theme that will be based in the long-established disciplines of computer science like as text mining, machine learning, natural language processing and artificial intelligence, voting advice applications, automated content analysis, etc.

c) Decision Making

Every person who stores information on the blogs, various web applications, and the web social media, social websites for getting the relevant information you need a particular method that can be used to analyze data and consequently return some of the useful results. It is going to be very difficult for the company to conduct the survey that will be on the regular basis so that there the need to analyze the data and locate the best of the products that will be based on user's opinions, reviews, and advice. The reviews and the opinions also help the people to take important decisions helping them in research and business areas.

d) Understanding Contextual

As human language is getting very complex day by day so it has become difficult for the machine to be able to understand the human language that can be expressed in the slangs, misspelling, nuances and the cultural variation. Thus, there will be a need for a system that will make better understanding between the human and the machine language.

e) Internet Marketing

Another important reason behind the increase in the demand for sentimental analysis is the marketing done via the internet by the business and companies organization. Now they regularly monitor the opinion of the user about their brand, product, or event on a blog or the social post that would help to make beneficial decisions.

C. Applications of Sentiment Analysis

Sentiment analysis has many applications in the NLP field. Due to the increase of social media data also increases demands of sentiment analysis. Some of the major applications are mentioned as following:

a) Word of Mouth (WOM)

Word of Mouth (WOM) is a task by which the information is given from one person to another person. It would help the people to make the decisions. Word of Mouth has given the information about the opinions, attitudes, reactions of consumers about the related business, services and the products or even the ones that can be shared with more than one person. As the online review blogs, sites, social networking sites have provided a lot of opinions; it has helped in the process of decision-making so much easier for the user.

b) Voice of Voters

Each of the political parties usually spent a major chunk of the amount of money for the aim of their party or for influencing the voters. Thus if the politicians know the people opinions, reviews, suggestions, these can be done with more effect. This is how the process of Sentimental analysis helps on the news analysts alongside.

c) Online Commerce

There is a vast number of websites related to e-commerce. The majority of them had the policy of getting the feedback from its users and customers. After getting information from various areas like service and quality details of the users of company users experience about features, product, and any suggestions. These details and reviews have been collected by company and conversion of data into the geographical form with the updates of the recent online commerce websites who use these current techniques.

d) Voice of Market (VOM)

Whenever a product is to be launched by a specific company, the customers would know about the product ratings, reviews and detailed descriptions of it. Sentiment Analysis can help in analyzing marketing, advertising and for making new strategies for promoting the product. It provides the customer an opportunity to choose the best among them all.

e) Brand Reputation Management (BRM)

Sentiment analysis would help to determine how would be a company's brand, service and the service or product that would be perceived by the online community. Brand Reputation Management will be concerned about the management of the reputation of the market. It has focused on the company and product rather than a customer.

f) Government

Sentiment Analysis has helped the administration to provide various services to the public by analyzing their needs. Fair results have to be generated by analyzing the negative and positive points of government. One of the interesting problems which can be taken up is applying this method in the multi-lingual country like the India where the content of the generating mixture of the different languages (e.g. Bengali-English) is a very common practice.

II. RELATED WORK

Sentiment analysis for the Punjabi language is a new treading research field as a number of systems is available for many other languages but for the Punjabi language not much research work has been done. The following research papers provide in-depth knowledge regarding this topic that would help for providing a better understanding of existing systems:

Prabowo et al. [6] used rule-based, support vector machine and hybrid approaches for finding the polarity of text. In rule-based approach, rules are used as an antecedent and if/then the relationship is used by its associated consequent. A consequent represent a sentiment or opinion in form of positive or negative terms. There are a number of the rule-based classifications algorithms like SBC, GIBC, IRBC, and RBC. Full form of GIBC is General Inquirer Based Classifier. It has 3673 pre-classification rules 1599 are positives and 2074 are negatives. These rules are applied to separate document content in form of positive or negative. IRBC is another rule-based classification algorithm. In this algorithm, 2nd rule sets are built by replacement of each and every proper noun that is in each document having sentences with '?' or '#' assigning to a set of antecedents, and each antecedent having an assigned sentiment for classification. SBC stands for Statistics Based Classifier.

Das et al. [7] developed a system to find the polarity of news content in the Bengali language with the help of support vector machine. They proposed Bengali Senti-WordNet. They collect the data for experiments from online Bengali news websites. They classify news corpus into two type's type 1 and type 2. Type 1 contains news content that goals to objectively represent factual information/data categories whereas type 2 contain opinionative content that contains editorial, letters, forum and editor categories related data. They proposed a classification approach to identify the sentences which contain opinionative terms. If any document having any opinionative terms and having theme phrases then they leveled or mark a sentence as subjective. They used SVM (Support Vector Machine) approach for features extract from the sentence. They used POS tagger for extraction of sentiment oriented terms from sentences. Sentiment oriented terms from the sentence are contained mainly adverbs, adjectives, verbs, noun. They also built a functional word list. Functional words list is mainly higher frequency words and that generally having very less opinionative information.

Natural Language Processing approaches, FCA (Formal Concept Analysis) based on ontology, and support vector machine is used to classifying reviews into positive, negative or neutral reviews. Opinions play an important role when an individual wants to buy any software they make a decision based on opinions of other-regarding any software. Mouthami et.al. [8] proposed a mining application for finding movies reviews. Due to increased growth of web-oriented data and use of social media applications, individual start to share their feelings, experiences, reviews, and opinions about any products or services over the internet.

Agarwal et. al. [9] proposed a system for identifying and classifying sentiments expressed in form of text. Now day's social media is producing a huge quantity of sentiment oriented rich data in the form of blog posts, Facebook data, Twitter data, online news articles etc. This user-generated, web-oriented data may contain very useful information that helps for finding the sentiments of the crowd data or getting useful information from unstructured data. Twitter sentiment analysis is different from general sentiment analysis because the involvement of slang words and use of creative writing makes it difficult for analyzing. The two strategies are used for sentiment analysis on text domain that's knowledge base approach and machine learning approach. They proposed a system to analyze the Twitter data to finding the reviews regarding electronic products like laptops, mobiles etc, with the use of machine learning approach. A new vector feature presented by them to identify the polarity of twitter data as positive or negative and detecting opinion oriented terms regarding products. Classification techniques were used by them is Naive Bayes classifier, SVM classifier, Max entropy classifier, and Ensemble classifier. Various symbolic and machine learning approaches are available to classifying sentiments or opinion from the documents. Machine learning approaches are very simple and better than various symbolic approaches. These kinds of approaches can be applied to Twitter data for analysis. There are some kinds of issues associated with it while deals with classifying emotional words (terms) from tweets that contain multiple words (terms). There are also difficulties to deals with misspellings, use of creative writing, and slang words. For dealing with issues, an optimal feature vector is developed by them for performing feature selection in two steps after pre-processing. Twitter specific features are selected in the first step. Then it entered into the feature vector. And at the last steps, these features deleted from tweets. Then again feature selection is performed on normal text.

Garg et al. [10] presented a modified algorithm for Punjabi. The algorithm developed using Naive Bayes approach. And this algorithm is applied to the Punjabi language for finding movies review. On reviews domain, authors concluded that in the Punjabi language the main issues while performing sentiment analysis is the reviews span over more than two sentences. In some situations when a review data may include more than two sentences and among them, but also a few sentences includes opposite sentiment. The algorithm takes the text as a input, and calculating the probability, If trigram detected in training datasets then the weight value is positive or negative, if the weight value is neutral value then it splits into bigram, if bigram to be in training datasets then

the weight value is considered as positive or negative value, if again the weight value is the neutral value then further it splits into unigrams, if unigrams detected in training datasets then the weight value is positive or negative value but if the weight value is the neutral value then discard it.

According to Kaur et al. [11], there is not much quantity of research work done in Indian Languages. Sentiment analysis in the Punjabi Language can be used in many businesses, social purpose applications like automatic summarization; opinion mining, machine translation, question answering systems etc. and increases their accuracy. The work focuses on resolving sentiment analysis for the Punjabi language. They proposed a model for resolving sentiment and an experiment is performed to measure the accuracy of the system. These popular approaches like a subjective lexicon, machine translation, Wordnet, and Bi-Lingual dictionary used by them. Then they combined the unigram approach and simple scoring approach for better efficiency. 54.2% is the overall efficiency of the proposed approach.

Bandyopadhyay et al. [12] proposed sentiment analysis system for Manipuri language and to achieve this they used unsupervised conditional random field (CRF) approach. In unsupervised learning, the system learns from itself for this purpose system they prepare some training datasets and that can be applied to other text for testing. Then they develop part of speech tagger algorithm for text domain using CRF. Due to the help of part of speech tagger, the verbs from each document are identifying and then the modified version of the lexicon is used to calculate the polarity of the sentiments that expressed in documents because the sentiments normally depend upon the verbs in the sentences. 75% accuracy on sentiment analysis of Manipuri language has achieved by their proposed algorithm approximately.

According to the survey of Medhat et al. [13], the data used in sentiment analysis mostly comes from product reviews in the overall counts. Other kinds of data mostly come from social media data that used the most frequent data over the past few years. They used support vector machine approach to identifying reviews. SVM algorithm was used by Medhat et al. [13] as a sentiment polarity classifier to find the polarity of text. They proposed an application that identifying the compact numeric summarization of documents for microblog platforms. They classifying and selecting the topics that included into the documents and also associated with the user's queries by using SVM. Twitter REST API is used to access public data for their research work. They effectively proved that their systems can analyze market intelligence (MI). That helps to support decision makers for establishment, development and monitoring systems to examine external opinions of users for different domains of a business in real time and it would help for developing better services.

A hybrid approach introduced by Bhaskar et al. [14] to analyze both speech and the corresponding text content in order to identify the speaker's emotion. In their work authors use different text features like frequency of word counts, the polarity of the post, the size of the post, PMI, and special symbols like punctuation marks etc. Pointwise Mutual Information (PMI) gives a numerical score value for terms based on its relationship with other specific domains. They used support vector machine classification approach. To find the speaker's emotion they performed transcribed audio recordings and to find the writers emotions they performed text. They use audio conversation of the users for their research work from a call center. The authors used a different kind of feature extraction algorithms for their research work. They used unsupervised and supervised machine learning approaches for further text/speech classification process. Authors proposed three approaches for analyzing speech that are the statistical approach, the semantic approach, and a hybrid statistical-semantic approach. Term-Frequency, Inverse-Document Frequency (TF-IDF) were used as the features in this approaches to find the values of the word from the text.

Preliminary steps of sentiment analysis according to Kumar et al. [15] are data acquisition and data preprocessing. Data collection is highly subjective to the type of analysis needed to perform, data format, and the type of media. Some blogging sites are available like Twitter, Sina-Weibo, and Facebook etc. for providing their Application Programming Interface (API) that helps for collecting public data from their websites. Twitter provides their Twitter REST API. Raw data collected from various different sources so preprocessed needs to be performed before performing a full analysis process. Some initial preprocessing steps are tokenization, stop word removal, stemming, parts of speech tagging, and feature selection and classification. Tokenization is a process to break a sentence into words and other meaningful tokens by removing punctuation marks, phrases, and symbols. Stop words are those common words that used in text document many times but do not involve in the analysis. Stemming is a task to get the root value of a word while ignoring another part of speech of the word. Part of speech tagging is used to assign different parts of speech to the word in the text documents. It is a process of assigning parts of speech to words that help the machine to understand human language.

Preliminary steps of sentiment analysis according to Kumar et al. [16] are data acquisition and data preprocessing. Data collection is highly subjective to the type of analysis needed to perform, data format, and the type of media. Some blogging sites are available like Twitter, Sina-Weibo, and Facebook etc. for providing their Application Programming Interface (API) that helps for collecting public data from their websites. Twitter provides their Twitter REST API.

Raw data collected from various different sources so preprocessed needs to be performed before performing a full analysis process. Some initial preprocessing steps are tokenization, stop word removal, stemming, parts of speech tagging, and feature selection and classification. Tokenization is a process to break a sentence into words and other meaningful tokens by removing punctuation marks, phrases, and symbols. Stop words are those common words that used in text document many times but do not involve in the analysis. Stemming is a task to get the root value of a word while ignoring another part of speech of the word. Part of speech tagging is used to assign different parts of speech to the word in the text documents. It is a process of assigning parts of speech to words that help the machine to understand human language.

The hybrid concept of maximum entropy and support vector machine is proposed by Jain et al. [17] states that support vector machine is used to assign the values in vector matrix. Support vector machine used linear regression model for calculation. A threshold value is used by vector matrix to decide the weight value of emotion. Depending upon this weight values emotion is classified into the predefined classes that are joy, anger, fear, surprise, sadness, and disgust.

Singh et.al. [18] presents research work for sentiment analysis on aspect-level for finding movie reviews using feature-based heuristic approach. They have modified a perspective-arranged plan that works for examinations of the text-based audits of a film and dole out it an assumption name on every viewpoint. The values are then collected on every angle from different audits and evaluation matrices produced on all performance parameters. They use a Senti WordNet-based approach. And two different distinctive etymological gimmick choices including modifiers, qualifiers, verbs and n-gram characteristic extraction are used. They have likewise utilized Senti WordNet approach to register the record level assumption for every motion films looked into and contrasted the results and then results got utilizing Alchemy API. The slant views of motion films are additionally contrasted and the archived aspect-level notion results. The results got to demonstrate that their plan to delivers a more exact, efficient and centered supposition views than the basic record level estimation investigation.

III. METHODOLOGY

A piece of end-to-end research on sentiment labeling, using supervised learning techniques is performed. This will involve preprocessing corpora, making choices about features extraction to include in-text representations, training classifiers and evaluating performance.

A. Tokenization

A token is a small unit of characters, and a sequennncly grouped characters that makes a sensible sentence and it useful for semantic processing. Remove punctuation marks, special symbols, numerical values (':', '*', ',', ':', '|', '/', '[', ']', '{', '}', '(', ')', '^', '&', '+', '?', '=', '<', '<', 'ا', '.', '?', '_', '-', '\', '%', "'", '!', '@', '#', '$', '%', '1', '2', '3', '4', '5', '6', '7', '8', '9', '0' etc.) from the Punjabi text documents.

B. Stop word removal

To remove these words from the text documents, the proposed system has built a list of Punjabi stop words, which has been manually analyzed and identified stop words.
Some common Punjabi stop words are as follow:
ਦੇ, ਦੀ, ਵਿਚ, ਦਾ, ਨੂੰ, ਹੈ, ਹੀ, ਹੇ, ਕੇ, ਉਸ, ਨਹੀਂ, ਤੇ, ਉਹ, ਤੋਂ, ਨਾਲ, ਹੋ, ਇਹ, ਭੀ, ਨੇ, ਕਰ, ਜਿਸ, ਇਸ, ਆਪਣੇ, ਜੋ, ਮੈਂ, ਕੋਈ, ਵਾਲਾ, ਆਪ, ਤੂੰ, ਕਰਦਾ, ਕਿ, ਉਹਨਾਂ, ਜੀ, ਤਾਂ, ਕਰਨ, ਸਭ, ਜਾ, ਰਹਿੰਦਾ, ਵਾਲੇ, ਵਾਲਾ, ਹਨ, ਹੈ, ਹੋਰ, ਪਰ, ਜੇ, ਕੀ, ਜਾਂਦੇ, ਅਤੇ, ਕਿਸੇ, ਨਾਹ, ਹੋਇਆ, ਰਿਹਾ, ਜਾਂਦੀ, ਮਿਲ, ਉਤੇ, ਹੁੰਦਾ, ਤੇਰੇ, ਰਹਾਉ, ਆ, ਹੋਏ, ਦੂਰ, ਬਿਨਾ, ਪੈਦਾ, ਲੈਂਦਾ, ਮੈਨੂੰ, ਕਾ, ਦੇਂਦਾ, ਲਈ, ਕਿਰਪਾ, ਦੇਣ, ਹਰ, ਰਹਿੰਦੇ, ਮੇਰਾ, ਜੀਵਾਂ, ਪੈ, ਹਰੇਕ, ਤੇਰੀ, ਤੇਰਾ, ਕਰਦੇ, ਆਪਣਾ, ਸਕਦਾ, ਜਦੋਂ, ਬਣ, ਕਰਿ, ਹੋਈ, ਦੀਆਂ, ਥਾਂ, ਆਪਣੀ, ਕੁਝ, ਪੈਂਦਾ, ਵਾਲੀ, ਵੇਲੇ, ਆਪੇ, ਆਦਿਕ, ਵਾਸਤੇ, ਇਹਨਾਂ, ਕਦੇ, ਮਨੁ, ਹੋਇ, ਰਹੇ, ਉਹੀ, ਰਹਿ, ਮੇਰੀ, ਵਿਚੋਂ, ਤਾ, ਪਾਇਆ, ਕੀਤਾ, ਲੈ, ਪਾ, ਸਾਰੀ, ਕਈ, ਲਿਆ, ਦਿੱਤਾ, ਤਰ੍ਹਾਂ, ਕੰਮ, ਸਮਝ, ਆਪਿ, ਜਿਵੇਂ, ਉੱਤੇ, ਤਦੋਂ, ਕੋ, ਨਾ, ਹਾਂ, ਮੈ, ਨੰ:, ਸੀ, ਨਾਹੀ, ਫਿਰ, ਇਉਂ, ਉਸੇ, ਰੇ, ਸੇ, ਇਹੁ, ਕਿਸ, ਵਲ.

C. Stemming

Stemming is a task to reduce a derived word from their root value or stem value. This is simple and fast kind of approach. Stemming from proper names and nouns in the Punjabi language is proposed to get the root/stem value of Punjabi words. For depth analysis possible noun and proper name, suffixes have been mentioned in Table I and the various Punjabi rules for word proper names and noun stemming have been produced. Proper Names and nouns are used to deciding the need for sentences for analysis. E.g. ਲੜਿਕਆਂ-ਲੜਕਾ, ਲੜਿਕਓ-ਲੜਕਾ, ਭਾਸ਼ਾਈ-ਭਾਸ਼ਾ etc.

Table 1. Punjabi Language Noun/Proper Name Suffix

ੀ ਆਂ	ਿੀਆਂ	ਿੀਆ	ੀ ੀਂ
ੀ ਏ	ੀ	ੀ ਓ	ਿੀਓ
ੀ ਆ	ਈ	ਵਾਂ	ਿੀਉਂ
ਈਆ	ਜ	ਜ਼	ਸ

D. Part of speech tagging

POS tagging is a task that used for allotment of correct tags to the word from a number of available tags. Here the tag means grammatical information of the word. It is well known that a computer will understand the language and process the language if the meaning of each and every word of that language is known or well defined.

Punjabi words may be infected or uninfected. Inflection is usually a suffix, which represents grammatical equation such as number, person, tense etc. The tagset consists of 38 Coarse-grained tags. Table II shows the Punjabi POS tagset used for the proposed system.

In most of the natural language processing applications like grammar checking, sentence identification, phrase chunking etc. the computer required only grammatical information of the input text. This grammatical information is given in the form tags called part of speech tags. Here the parts of speech are different word classes in which a word lies like a noun, adjective, verb etc. A word can have more than one tags and it can occur in more than one-word class in different context.

Table 2. POS Tagset for Punjabi

MAIN CATEGORY	SUB CATEGORY	POS TAG
NOUN	COMMON	NN
NOUN	PROPER	NNP
NOUN	COMPOUND	NNC
NOUN	COMPOUND-PROPER	NNCP
PRONOUN	ALL-CATEGORIES	PRP
ADJECTIVE	ALL- CATEGORIES	JJ
VERB	MAIN	VB
VERB	FIRST PERSON	FP
VERB	PRESENT TENSE	PT
VERB	PAST TENSE	PAT
VERB	FUTURE TENSE	FT
VERB	AUXILIARY	VAUX
ADVERB	-	RB
CONJUNCTION	SUB-ORDINATE	CS
CONJUNCTION	CO-ORDINATE	CC
INTERJECTION	-	INJ

E. Transformation

The weight value of each and every word from the corpus is determined with the use of term frequency and inverse document frequency (TF-IDF). TF-IDF defines the weight values of each and every word in a document using the formula:-

$$wd = fw, d^* \log(|Doc| fw,Doc)$$

w denotes words in an individual document, Doc is a collection of documents, d is single document belongs to Doc, |Doc| is the size of the corpus, fw,d is a number of times word appears in a document, fw,Doc is a number of documents in which word appears in Doc. And the highest valued TF-IDF Punjabi words are candidates for feature selection in this step.

F. Feature selection

Feature Selection helps to build a classification more effective by deleting the quantity of content to be analyzed and selecting related features to be considered for the classification process. For Punjabi text classification, TF*IDF is used to extract the related features for example words that have less than 2 threshold value are not selected as features.

G. Classification

Text classification is used to classifying data into predefined classes and the classes can be positive and negative or it can be used some other classes based on their needs. The first step of machine learning approaches is transforming documents means converting a string format into a suitable string format. In the proposed system used supervised learning approach for text classification. Each word corresponding to features takes its weight value.

H. Sentiment analysis

Sentiment analysis is a task that changed unstructured data into beneficial information. When the analysis process has been done, the results are represented in the form of graphs for examples pie chart, line graphs, and bar chart.

IV. RESULTS AND DISCUSSION

The proposed sentiment analysis system used only support vector machine approach for the implementation, and the advantage of this method is that it can further applied to virtually any language in this world. There are 14.8% of errors occurring due to the absence of certain Punjabi noun words in noun morph, dictionary mistakes, and input text syntax mistakes. In the proposed system, stemmer implemented for Punjabi and it is a simplified version of stemmer. There is a very little influence of suffix stripping algorithm in this stemmer. There is a problem of over-stemming and under-stemming comes under suffix stripping approach. There is a need to do suffix substitution with suffix stripping to avoid the problem of over-stemming and under-stemming. The accuracy of proposed stemmer is 93%.

The proposed part of speech tagger shows an accuracy of 93-95% whereas existing system gives an accuracy of 86-88%. And the precision, recall, and accuracy of Punjabi language feature selection are 89.4%, 95.6%, 95.2% respectively. Support vector machine can be successfully applied to part of speech tagging for the Punjabi language. Support vector machine achieves high accuracy as compared to rule-based

and HMM approaches.

Support Vector Machine approach is used for classification. It is based on supervised learning. It may make errors. To compare different classifiers for deciding which approach better? The experiment is conducted on Punjabi news articles to evaluate its performance which is done using precision and recall. The comparison made between support vector machines towards Naïve Bayes classifier. Proposed approach has been used optimal values which classify the dataset with more accuracy than the existing system. To enhance the accuracy many features can be constructed. Based on the experiments following results are concluded:

Table 3. Systems Performance

STEPS	PRECISION	RECALL	ACCURACY
TOKENIZATION	100%	100%	100%
STOP WORD REMOVAL	84.6%	89.9%	90%
STEMMING	57.1%	80%	84.7%
PART OF SPEECH TAGGING	68.4%	84.7%	87.1%
TRANSFORMATION	89.2%	90%	90%
FEATURE SELECTION	88.6%	92.4%	89.1%
CLASSIFICATION	88.5%	89.3%	90%

Table 4. Overall Systems Performance

PRECISION	88.5%
RECALL	89.3%
ACCURACY	90%

Table 5. Comparison with Other Approaches

APPROACHES	ACCURACY
UNIGRAM	75.5%
BIGRAM	52.5%
TRIGRAM	60.5%
UNIGRAM+BIGRAM+TRIGRAM	54.5%
WEIGHTED RESULTS	51.5%
AVERAGE RESULTS	61.5%
SUPPORT VECTOR MACHINE	90%

V. CONCLUSION AND FUTURE SCOPE

The results obtained using support vector machine approach is satisfactory, and the precision, recall, and accuracy of the system are 88.5%, 89.3%, 90% respectively. The results are evidence of how a correctly implemented approach can help in making a text classifier. The proposed system and implementation is based on dynamic datasets of news corpus on a broad range of topics.

Large amount of work in sentiment analysis has been done in English language, because English is a global language, but there is a need to perform sentiment analysis in other languages Indian languages also. Large amount Punjabi contents are available over the Web which needs to be mined to determine the sentiment. The Punjabi databases like Word Net have also not been used in the existing works. In future, try to experiments with more focused approach, tools, techniques and other heuristics to develop subjective lexicon for the Punjabi language, which does not utilized it, Word Net but a proposed algorithm.

ACKNOWLEDGMENT

I am highly grateful to Dr. M.S. Saini, Director, Guru Nanak Dev Engineering College, Ludhiana, for providing this opportunity to carry out the present research work. I humbly acknowledge my gratitude to my Guide Er. Kamaldeep Kaur for helping me, answering my doubts all along and guiding me. Without the wise counsel and able guidance, it would have been impossible to complete the research work in this manner. I would also like to express a deep sense of gratitude and thanks profusely to the H.O.D. of department Dr. Parminder Singh and research committee members Dr. K.S. Mann, Dr. Akshay Girdhar, Dr. Manpreet Singh, Mr. Amanpreet Singh Brar, and Mr. Jasbir Singh Saini for providing necessary guidance and encouragement during my thesis work. I am also thankful to the other faculty members of Computer Science and Engineering Department, of Guru Nanak Dev Engineering College for their support through the course of this work. Special thanks to my parents and all the family members for their encouragement and tireless support for my academic pursuit. I would like to express my immense gratitude to God, the Almighty who is the most Gracious and Merciful, whose guidance enabled me to accomplish my research work.

REFERENCES

[1] P. Panday, "*A survey of Sentiment classification techniques used for Indian regional languages* ," International Journal on Computational Science and Applications, vol. 5, pp. 13-26, April 2015.

[2] A. Balahur and G. Jacquet, "*Sentiment analysis meets social media – Challenges and solutions of the field in view of the current information sharing context*," International Journal Of Engineering And Computer Science, vol. 3, pp. 428–432, July 2015.

3] J. Kaur and R. Saini, "*A study and analysis of opinion mining research in indo-aryan, dravidian and tibeto-burman language families*," International Journal of Data Mining and Emerging Technologies, vol. 4, no. 2, pp. 53-60, July 2014.

[4] A. Kaur and V. Gupta, "*Proposed algorithm of sentiment analysis for Punjabi text*," International Journal of Science and Technology, vol. 6, no. 4, pp. 180-183, May 2014.

[5] A. Sharma, "*Sentiment analyzer using Punjabi language*," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 9, pp. 5904-5905, September 2014.

[6] R. Prabowo and M. Thelwall, "*Sentiment analysis: a combined approach*," International Journal of Computer Applications & Information Technology, vol. 4, no. 8, pp. 143-157, July 2009.
[7] A. Das and S. Bandyopadhyay, "*Opinion-Polarity Identification in Bengali*," International Arab Journal of Information Technology, vol. 2, no. 9, pp. 169-182, May 2010.
[8] K. Mouthami and A. Baloglu, "*Sentiment analysis and classification based on textual reviews*," International Conference on Internet and Web Applications and Services, chicago, 2010, pp. 10-17.
[9] N.C. Troussas, and A. Agarwal, "*Sentiment analysis in twitter using machine learning techniques*," International Journal of Computer Science and Engineering, vol. 4, no. 2, p. 3038, June 2011.
[10] N. Garg "*Movie review mining in Punjabi*," International Journal of Application or Innovation in Engineering & Management, vol. 2, no. 12, pp. 372-375, December 2013.
[11] J. Kaur, J. Kumar, and R. SAINI, "*A study of text classification natural language processing algorithms for Indian languages*," International Journal of Computer Applications & Information Technology, vol. 4, no. 9, pp. 162-167, July 2015.
[12] A. Das, "*Opinion-polarity identification in Bengali*," International Journal of Computer Science, vol. 10, no. 5, pp. 169-182, April 2010.
[13] W. Medhat, A. Hassan, and H. Korashy, "*Sentiment analysis algorithms and applications: A survey*," Ain Shams Engineering Journal, vol. 1, no. 5, pp. 1093-1113, April 2014.
[14] J. Bhaskar, K. Sruthi, and P. Nedungadi, "*Hybrid approach for emotion classification of audio conversation based on text and speech mining,"* International Conference on Information and Communication Technologies, Amritapuri, pp. 635-643, April 2015.
[15] R. Kumar and R. Vadlamani , "*A survey on opinion mining and sentiment analysis: tasks, approaches and applications*," International Journal of Advanced Research in Computer and Communication Engineering, vol. 37, no. 11, pp. 14-46, June 2015.
[16] V. Kumar, "*Opinion mining and sentiment analysis*," International Journal of Advanced Research in Computer and Communication Engineering, vol. 39, no. 10, pp. 141-166, April 2015.
[17]U. Jain and A. Sandu, "*Emotion Detection from Punjabi Text using Hybrid Support Vector Machine and Maximum Entropy Algorithm*," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 11, p. 5, November 2015.
[18]V.K. Singh and R. Rani, "*Sentiment analysis of movie reviews for aspect-level sentiment classification*," International Journal of Advanced Computer Research, vol. 9, no. 5, pp. 10-39, April 2015.

Authors Profile

Er. Gagandeep Kaur pursed Bachelor of Computer Science and Engineering from Guru Nanak Dev Engineering College Ludhiana in 2015 and Master of Computer Science and Engineering from Guru Nanak Dev Engineering College Ludhiana in 2017.

Prof. Kamaldeep Kaur currently working as Assistant Professor in Department of Computer Science and Engineering in Guru Nanak Dev Engineering College Ludhiana.