International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Sentiment Analysis on Punjabi News Articles Using SVM

Gagandeep Kaur[1], Kamaldeep Kaur[2]

[1]M.Tech. Student, Guru Nanak Dev Engineering College, Ludhiana
[2]Professor, Guru Nanak Dev Engineering College, Ludhiana

Abstract: ***Sentiment analysis is a field of Natural Language Processing and it is the most trending field of research. In the process of text mining that is used to find out people's opinion about a particular product, topic and predicting market trends or outcomes of elections, detecting and classifying sentiments from the text. 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. This work basically focuses on mining sentiments and analyzing them for the Punjabi language. With the increase in the amount of information being communicated via regional languages like Punjabi, comes a promising opportunity of mining this information. Nowadays, it is a new trend to read online news in a daily practice. People's opinion tends to be changed as per they read news content. The news content that they read normally about the negative content regarding various things for example rapes, corruption, thefts etc. Reading such negative news is spreading negativity around the society. So there is need to classify the positive and negative news content for creating a positive environment because if they read positive they think positive. Support Vector Machine approach is used by proposed system to classify the content into different categories of news like crime, entertainment, politics, sports, and weather and then finding its polarity. 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: Natural Language Processing, Sentiment Analysis, Punjabi Language, Machine Learning, Support Vector Machine, News Articles

1. 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. 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.

1.1 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.

1.2 Sentiment Analysis on Punjabi Language

There are 100+ million speakers of Punjabi language spread all over the world. The coverage area of Punjabi language is also growing day by day via the internet. The web pages may contain a huge amount of an important data regarding any corporate data and government data in government websites.

The resources, approaches, and tools to do successful research in Punjabi language are very limited so in this field, research scope is high [5]. 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 [6].

Sentiment analysis or opinion mining deals with finding sentiments or polarity of sentiments from a piece of input text regarding any entities. It is a process of analyzing emotions, feelings, opinion, and the attitude of a person that expressed in the form a given input of text information/data. Sentiment Analysis studies the behavior of the individual and their likes/dislikes of an individual from the input text information. The main aim of sentiment analysis is to identify sentiment associated with the text by extracting sentimental context from the text [7]. To finding the attitude, state of mind, and emotions of individuals through the contextual analysis and the way of the polarity of their speaking or writing sentiment analysis is used. The attitude

can be reflected by their own judgment, the emotional statement of the opinion, or the statement of any emotional conversation that they used to influence a reader or listener. To determine an individual's state of mind about the opinion that they are communicated sentiment analysis is used. The huge amount of data can be fetched from various data resources like texts, twitter data, Facebook data, blogs, social media, news articles, product reviews etc [7].

1.3 Approaches for Text Classification Task

Machine learning based classification need training datasets and test datasets. Training datasets are prepared to learn the different characteristics of data. Test datasets are used to validate the automatic classifier performance. A various machine learning approaches have been available to classify content. Different kind of machine learning approaches is Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) that have been gained great success in text domain. Some another mostly used machine learning approaches in the natural language processing fields are winnow classifier, K-nearest neighborhood, N-gram model, ID3, C5, centroid classifier [8].

Naive Bayes is a simple machine learning approach. It is based on probabilistic classifier approach. It uses Bayes theorem due to strong (naive) independence assumptions. The reason behind the success of Naïve Bayes approach is that it's working quality for text domain because the evidence is vocabularies/dictionaries or words/terms appearing in texts or documents, and the size of the vocabularies/dictionaries are commonly in the range of thousands. The large size of the corpus, evidence, and vocabularies will make this model best and work best for text classification problems. The Naive Bayes approach has widely used for content present in Indian Language for document classification [8].

A statistical classification approach is a support vector machine approach. It used to classify the training data into two different classes and makes decisions usually depends upon the support vectors that are used to choosing only effective elements of the training datasets. A support vector machine is a discriminative classification approach and it is considered as the best text classification approach. The K-nearest neighbor well known with short form KNN is a typical example-based classification approach that does not makes explicit decisions and declarative representation of the sentiments categories but it relies on the categories labels that allocated in the training documents similar to the test documents for comparison. For example, give a test document as an input to the system to finds the k nearest neighbors from the training documents and getting optimal classification results [8].

All supervised machine learning approaches provide a reasonable amount of accuracy but it can be obtained subjective only if it fulfills the requirement that validates the test data that should be similar to training data. When moving towards supervised machine learning approaches text domain classification to another specific domain it would require the collection of annotated data for the new specified domain and retained data for the classification process. This dependency on annotated training datasets is one of the major shortcomings of all supervised machine learning approaches [8].

Unsupervised learning machine learning approach is used to build a set of fully annotated data that helps to train a classifier with the use of machine learning approaches. This classification approach is then used to separate novel incoming data [8]. Indian Languages belongs to three language families: indo Aryan, indo dravidian, and Sino Tibetan. Indo Aryan consists of languages like Punjabi, Bengali, Hindi, Urdu, Oriya; Indo-Dravidian consists of languages like Telugu, Kannada, Tamil, Malayalam; and Sino-Tibetan consists of languages like Manipuri, Meithei, and Himalayish [9].

To separate subjective words from objective words statistical approaches like Naïve Bayes and Support Vector Machine are used. The Urdu language used a specific pre-processor for features selections. The Urdu language is a rich morphological language so the classification process is more complex in this language. According to results that will show accuracy, and the performance of support vector machines much better than Naïve Bayes classification approach. For Bengali language, n-gram classification approach is used and to analyze the performance of the classification process Prothom-Alo news corpus is proposed. According to the results when the value of n increases from 1 to 3, accordingly the performance of the text classification also increases, but when the value of n increases from 3 to 4 performances will decreases [9].

A very few works have been done in the field of Punjabi text classification. Text domain based classification is done in this language. The classification is done only sports categories. For this classification task two new approaches ontology based classification, and hybrid approach are proposed for Punjabi text classification. The conclusion of this experiment shows that ontology based classification and hybrid approach 85% provides better results [9].

2. Related Works

A 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 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. [10] 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 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. [11] 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. [12] 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. [13] proposed a system to 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 Nave 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. [3] presented 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. [14] there is not much quantity of research work done for 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. [15] 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. [16] 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. [16] 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. [17] 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. [18] 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. [5] 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. [19] 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.

3. Methodology

In this thesis work, 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.

In proposed system, to select the common features such as news and further news categories like politics, sports, entertainment etc Support Vector Machine approach is used. And then classify news articles into the positive or negative news. For doing this a support vector machine learning algorithm to used to classify news articles into different categories of news and then classifies news into the positive or negative news.

Support vector machine is a classification approach that receives input information during its training process and then building a hypothesis function for systems input and output that used for predicting future results. Support vector machine is a supervised machine learning approach with various learning methods. Support vector machine algorithm is used to analyze data and it uses recognize patterns and regression methods for classification. A support vector

machine is based on natural learning instead of heuristics or analogies techniques and the results are based on statistical learning techniques. Support vector machine performs an implicit embedment of data into a high-frequency feature space. It used nonlinear rules, geometry, and linear algebra for classification in input spaces.

And the next step to achieve further objectives to classify news articles into the positive or negative news are as follow:

3.1 Tokenization

Given input as form of text, documents to the system, tokenization is a process of breaking text up into small unit of pieces known as tokens and deleting some unique characters such as punctuation marks and special symbols at the same time. 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 (':', '*', ',', ':', '|', '/', '[', ']', '{', '}', '(', ')', '^', '&', '+', '?', '=', '<', '<', '।', '.', '?', '_', '-', '\', '%', '"', '!', '@', '#', '$', '%', '1', '2', '3', '4', '5', '6', '7', '8', '9', '0' etc.) from the Punjabi text documents.

3.2 Stop Word Removal

A stoplist is a list of commonly used words. Stop words usually language specific, although every language may contain many stop words. Natural language processing systems and many search engines having a variety of stop words depend upon languages, or it can be used a single multilingual stop-list. By the ignorance of functional words, the meaning of terms can be more clear, when processing the natural language content. Hence it is a process of removing words which do not support for information analysis task and appear too often in the documents. To remove these words from the text documents, 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, ਸੈ, ਨ:, ਸੀ, ਨਾਹੀ, ਫਿਰ, ਇਉਂ, ਉਸੇ, ਰੇ, ਸੋ, ਇਹੁ, ਕਿਸ, ਵਲ.

3.3 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 for 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 3.1 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 3.1: Punjabi Language Noun/Proper Name Suffix

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

3.4 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.

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. Punjabi words may be inflected or uninflected. Inflection is usually a suffix, which represents grammatical equation such as number, person, tense etc. The tagset consists of 38 Coarse-grained tags. Table 3.2 shows the Punjabi POS tagset used for the proposed system. The number of tags used for a language depends upon the length of the tag which further depends upon the amount of information that represents using a tag. e.g. if just basic word class is to be represented by each word then the length of the tag will be 2, 3 or 4. Various approaches available for tagging, for example, HMM tagging, constraint grammar tagging, and transmation-based tagging.

Table 3.2: 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

3.5 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.

3.6 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. Punjabi corpus prepared for the Punjabi language e.g. list of sports relates terms are:

ਖੇਡ, ਕ੍ਰਿਕੇਟ, ਹਾਕੀ, ਫੁੱਟਬਾਲ, ਕਬੱਡੀ, ਰਾਸ਼ਟਰੀ, ਇਨਾਮ, ਜਿੱਤ, ਨੁਕਸਾਨ, ਖੇਡਣਾ, ਗੇਮ, ਗੋਲ, ਖਿਡਾਰੀ, ਸਨਮਾਨ, ਪ੍ਰਦਰਸ਼ਨ, ਮੈਚ, ਦੌੜ, ਜੰਪ, ਟੀਚਾ, ਸਾਈਕਲ, ਮੈਡਲ, ਜੇਤੂ, ਕੋਚ, ਕੱਪ, ਬਾਹਰ etc.

And similarly, crime, entertainment, politics, and weather related terms by deleting non-relevant contents from the documents.

3.7 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 proposed system used supervised learning approach for text classification. Each word corresponding to features takes its weight value.

3.8 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.

4. 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 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 existing system. To enhance the accuracy many features can be constructed. Based on the experiments following results are concluded:

Table 4.1: 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 4.2: Overall Systems Performance

Precision	88.5%
Recall	89.3%
Accuracy	90%

Table 4.3: 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%

5. Conclusion and Future Scope

The proposed sentiment analysis system is used for the domain (crime, sports, entertainment, politics, and weather) based classification of Punjabi text documents. This is the proposed system for the classification of Punjabi text documents where Python language is used for the implementation of the system. As not much work has been

done for Punjabi Language so in this approach an initiation is done to develop a corpus for the Punjabi language by creating news based training datasets in Punjabi that consist of class related terms. E.g. crime class consists of words like:

ਡਰ, ਹਨੇਰ, ਕਤਲ, ਖੂਨ, ਦੁਰਘਟਨਾ, ਉਦਾਸ, ਰੋਵੋ, ਡਰ, ਜੀਵਨ, ਖੁਦਕੁਸ਼ੀ, ਪੁਲਿਸ, ਜਾਂਚ, ਭ੍ਰਿਸ਼ਟਾਚਾਰ, ਬੰਦੂਕ, ਗੋਲੀ, ਤਾਕਤ, ਚੋਰ, ਚੋਰੀ, ਪੈਸਾ, ਕੱਟੋ etc.

Labeled documents are used to classify the documents. The results obtained using support vector machine approach is satisfactory, and the precision, recall, and accuracy of the system is 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. It is used to find out negative and positive articles and delivered only good news contents after classifying news articles using sentiment analysis using support vector machine approach and creating a positive environment. This will help spread positivity around society and would allow people to think positive because if they read positive they think positive.

The Punjabi databases like WordNet have also not been used in the existing works. In future, to do experiments with the more focused approach, tools, techniques and other heuristics to develop a subjective lexicon for the Punjabi language, this does not utilize it, WordNet but a proposed algorithm. The same approaches will be implemented to translate English to Punjabi languages. Then there is need to explore and dig in depth regarding the task of sentiment classification for the web text and also to improve it. Once sufficient data is available for experiments, various machine-learning techniques can be easily tested and applied to learn from the text data more effectively.

In context to Indian Languages, earlier work done for sentiment analysis has been on Bengali and Hindi, rest all the other languages are unexplored. By the help of deep study of existing research papers, it has been found that Punjabi is unexplored language. The things that will be included in the study of these research papers will consist of algorithm, tools, approaches which have not been implemented yet but proposed.

The present work is to classify the Punjabi content into sports, politics, entertainment, crime, and weather documents only. Therefore, in future, it can be expanded for others domains too, that means classification can be performed on another level of sentiment analysis.

Many applications need the content of text so that way needs more research work in context based sentiment analysis. Overall Sentiment Classification used in various applications of Sentiment Analysis like the market, political, equality value based and box office prediction etc. but, a lot of work still remains to be done for Indian content. Indian content still needs more research work in the field of sentiment classification. Large amount Punjabi contents are available over the Web which needs to be mined to determine the sentiment.

References

[1] Pooja Pandey and Sharvari Govilkar, "A SURVEY OF SENTIMENT CLASSIFICATION TECHNIQUES USED FOR INDIAN REGIONAL LANGUAGES," International Journal on Computational Science & Applications, vol. 5, no. 2, pp. 13-26, April 2015.
[2] Alexandra Balahur and Guillaume 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, no. 10, pp. 428–432, July 2015.
[3] Navneet Garg and Deepali, "Movie Review Mining in Punjabi," International Journal of Application or Innovation in Engineering & Management (IJAIEM), vol. 2, no. 12, pp. 372-375, December 2013.
[4] Anu 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.
[5] Ubeeka Jain and Amandeep 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.
[6] Amitava Das and Sivaji Bandyopadhyay, "Opinion-Polarity Identification in Bengali," The International Arab Journal of Information Technology, vol. 2, no. 9, pp. 169-182, May 2010.
[7] Pooja Pandey and Sharvari Govilkar, "A SURVEY OF SENTIMENT CLASSIFICATION TECHNIQUES USED FOR INDIAN REGIONAL LANGUAGES ," International Journal on Computational Science & Applications, vol. 5, no. 2, pp. 13-14, April 2015.
[8] Jasleen Kaur, kumar Jatinder, and R. SAINI, "A Study of Text Classification Natural Language Processing Algorithms for Indian Languages," VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY, vol. 4, no. 9, pp. 162-167, July 2015.
[9] Jasleen 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.
[10] Rudy Prabowo and Mike Thelwall, "Sentiment Analysis: A Combined Approach," International Journal of Computer Applications & Information Technology, vol. 4, no. 8, pp. 143-157, July 2009.
[11] Bandyopadhyay, Amitava Das, and Sivaji, "Opinion-Polarity Identification in Bengali," International Journal of Computer Science, vol. 10, no. 5, pp. 169-182, April 2010.
[12] K. Mouthami and Arzu Baloglu, "Sentiment Analysis and Classification Based On Textual Reviews," in International Conference on Internet and Web Applications and Services, chicago, 2010, pp. 10-17.
[13] Neethu, Christos, Troussas, and A. Agarwal, "Sentiment Analysis in Twitter using Machine Learning Techniques," Computer Science & Engineering: An International Journal, vol. 4, no. 2, p. 3038, June 2011.

[14] Amandeep Kaur and Vishal Gupta, "Proposed Algorithm of Sentiment Analysis for Punjabi Text," JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, vol. 6, no. 4, pp. 180-183, may 2014.
[15] Bandyopadhyay, Kishorjit, and Sivaji, "Verb Based Manipuri Sentiment Aanalysis," Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 2, pp. 180-183, April 2014.
[16] Walaa Medhat, Ahmed Hassan, and Hoda Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 1, no. 5, pp. 1093-1113, April 2014.
[17] Jasmine Bhaskar, Sruthi K, and Prema Nedungadi, "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining," in International Conference on Information and Communication Technologies, Amritapuri, 2015, pp. 635-643.
[18] Kumar Ravi and Vadlamani Ravi, "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.
[19] V.K. Singh and R. Rani, "Sentiment Analysis of Movie Reviews A new Feature-based Heuristic for Aspect-level Sentiment Classification," International Journal of Advanced Computer Research, vol. 9, no. 5, pp. 10-39, April 2015.