EasyChair Preprint

№ 2813

Sentiment Analysis of English-Punjabi Code Mixed Social Media Content for Agriculture Domain

Mukhtiar Singh, Vishal Goyal and Sahil Raj

February 29, 2020

Sentiment Analysis of English-Punjabi Code Mixed Social Media Content for Agriculture Domain

Mukhtiar Singh,
PhD Research Scholar,
Department of Computer Science,
Punjabi University, Patiala, India.
mukhtairrai73@gmail.com

Vishal Goyal,
Professor,
Department of Computer Science,
Punjabi University, Patiala, India.
vishal.pup@gmail.com

Sahil Raj,
Assistant Professor,
School of Management and Studies,
Punjabi University, Patiala, India.
dr.sahilraj47@gmail.com

***Abstract*— In India, more than 70% of the population is dependent on agriculture. Since the independence of India, the people involved in agriculture mostly stay in rural areas. The government has taken numerous efforts for the improvement of conditions of farmers. Still the condition is not improved to acceptable rate. Currently, it has been easy to extract the reviews of farmers from micro-blogging websites. Since decades, a trend has been seen that multilingual speakers often switch between more than one languages to express themselves on social media networks. Multiple languages are mixed with different rules of grammars, which in itself is the challenging task. In this paper, the authors have extracted the agriculture-related comments having code-mixing property with English-Punjabi mixed content. Further, the performed language identification, normalization, and creation of English-Punjabi code-mixed dictionary. After that, we have tested various models trained on English-Punjabi code mixed data using Support Vector Machine and Naive Bayes techniques for sentiment analysis, tested the pipeline for unigram predictive model. Later experimented for n-gram and performance was found to be better in our implemented model.**

Keywords*— *Sentiment Analysis, Social Media, Support Vector Machines, Naive Bayes, Language Identification, Agriculture.

I. INTRODUCTION

Sentiment analysis is opinion mining. It has become a popular research area. We have an envious amount of text available. The textual information is divided into two parts, (1) Facts and (2) Opinions. Facts are a subjective expressions about entities and the subjective expression which describes people's feelings towards entity. Opinions are helpful for decision making in any business, such as seed, fertilizer and agriculture machinery industries plan production and marketing activities based on crop production [1]. There are two major regions, which are helpful for the farmer as well as the government for decision-making.

1) It provides farmers with the historical crop yield record with a forecast reducing the risk management.

2) It helps the government in framing crop insurance policies and policies for a supply chain operation.

Some of the factors on which crop production business agriculture is dependent are soil, climate, cultivation, irrigation, fertilizers, temperature, rainfall, harvesting, pesticide weeds, and other factors. Nowadays, the increasing importance of sentiment analysis corresponds with the development of social media. As a result, for analysis of the data, there is a huge volume of opinionated data recorded in digital form. English has been mostly dominated language in Web 2.0, but on microblogging websites, its dominance is receding. It has been noticed that on social media platforms about half of the messages were in the English language while other the half was in the user's mother tongue language. The web 2.0 allows users to write content freely on the internet through its different platforms such as blogs, social networks, microblogs, or forums with the purpose to provide, share and use information. There are various views of people in fields, such as Agriculture, Culture, Politics, Sports, Education, Entertainment, Health, Religion, Technology, Tourism etc. According to the 2010 census, users of English and Punjabi as their first language are 5.52% and 2.83% respectively of the total world population. From which we observed that, especially in Punjab state, around 2.57% of the total population express their opinion in English-Punjabi mixed on the social media platform. In our work, we have used the English-Punjabi code mixed data, which includes various comments from Facebook, Twitter, and YouTube. It is noticed that, on the social media platform, the text is informal. Moreover, the language is quite complex and context relevant, especially when people are expressing their views and opinions. For more effective and accurate sentiment analysis, it is very essential to understand the context and topic-specific terminology. In sentiment analysis process, the system classifies and categorizes the opinions and reviews expressed in content as positive or negative by applying the classification approaches with machine learning algorithms that extract and evaluates the point of view expressed.

To create structured data, the system performs two tasks:

1. Analysis of opinions and reviews, by assigning each opinion a quantitative value and a relevant degree of accuracy.
2. Include this metadata information to the index.

Lot of work has been done by researchers in the area of phrase level and sentence level sentiment analysis classification and on analyzing blog posts [2, 3]. For a lot of work has been done in English for sentiment analysis, whereas, very few experiments have been done on the Punjabi Language, but the amount of work done on English-Punjabi code mixed data is not much. In cases, we noticed that the user uses multilingual content to express their opinion. For solving this problem widely used method is to convert the Pun-English language transliterate into the Punjabi language. To best of our knowledge, no work has been done on sentiment identification of English-Punjabi code mixed data. Many challenges have been observed in this scenario. The most common ones are present in grammatical errors as well as ambiguity in the multilingual text.

In this paper, we have been having experimented different machine learning algorithms. When trained on English-Punjabi code mixed data. On a whole, we have performed two experiments. In the first experiment, we tested the English-Punjabi agriculture comments dataset, where the support vector machine used to obtain better results. For the second experiment, we, trained text data using the naive bayes machine-learning algorithm. Finally, we extracted features from the code-mixed trained dataset and again tested by using support vector machine, which support vector machine technique performed better as compared to naive bayes algorithm.

The remaining of this paper organized as follows: Section II gives an overview of the background and related work. Section III provides an outline of data collection. Section IV gives the proposed approach of sentiment analysis, in which, the pre-processing and feature extraction built for classification is described. Last section V gives the implementation and results related to the agriculture domain. The conclusion and future work is drawn in section VI.

II. BACKGROUND

System is developed for exploring Combined Multi-level model in Document Sentiment Analysis [4]. They have purposed a novel combination model based on phrase and sentence level analyses and they extracted the features. They used the different features for sentiment analysis. Word N-Grams, POS tag, linguistic analysis, negation terms, degree modifiers (very, much), transitional word (but) and the dependency relationship (asymmetric binary relationship and obtained. They have marked the number of positive words as WordPosNum and negative words as WordNegNum respectively. Sum of positive and negative words was marked as WordSubNum. They have also analyzed sentence-level sentiment analysis features.The applications of data mining techniques are explored in the field of agriculture [1]. Historical crop yield information is important for supply chain operation of companies engaged in industries that use agricultural products as raw material. Livestock, food, animal feed, chemical, poultry, fertilizer pesticides, seed, paper, and many other industries use agricultural products as for gradient in their production processes. An accurate estimate of crop size and risk helps these companies in planning supply chain decision [5]. The data selection is studied in [6] which retrieves the data relevant to the analysis from the various data locations. Data pre-processing is the process of data cleaning and data integration is done. Data integration is multiple data sources and combined in a common source. Data Mining is the critical step in which the best techniques are applied to extract potentially useful patterns. One has been careful about the data mining technique to be used. The applications are in [7] of data mining in agriculture. They have collected the agriculture data and stored it in an organized form, and their integration enables the creation of an agricultural information system. Data mining technology provides user-oriented access to new and hidden patterns in data, from which knowledge is generated which can facilitate decision making in agricultural organizations. Mucherino in [8] have studied about data mining in agriculture. They have discovered problematic wine fermentations at the early stages of the process and briefly describe other problems in the field. The k-means algorithm was used. They described the recent developments on this problem, and in particular new studies where bi-clustering techniques are employed for identifying the compounds of wine that are most likely the cause of problematic fermentations. The second problem they are predicting yield production. Wood in [5] evaluated the different types of data mining techniques on data sets. The K- means algorithm is able to partition the sample in clusters. Cunningham, S. in [9] have discussed WEKA process model for analyzing data and application construction process is illustrated through a case study in the agricultural domain. The result indicated that subjective attributes for mushroom grading may not be useful in practice. Mirjankar, N. in [10] have tried to analyze different data mining procedures related to agriculture field. They show a brief thought of a percentage of the broadly utilized data mining systems. Data mining through better management and data analysis can assist agricultural organizations to achieve greater profit. Palepu, R. B. in [11] presented the role of data mining in perspective of soil analysis in the field of agriculture. Fetanat, H. in [12] analyzed data with regression techniques, which showed the effect of chlorophyll content on the number of flowers. They have also analyzed agricultural data using different data mining methods. Valsamidis, S in [13] have outlined the challenges and opportunities of blogs for agriculture. They have used RapidMiner software for opinion mining. This framework can be used as baselines for opinion mining tasks. Kalpana, R. in [14] studied to search out appropriate data prediction capabilities. Majumdar, J. in [15] studied on the analysis of the agriculture data and finding optimal parameters to maximize crop production using PAM, CLARA, DBSCAN, and Multiple Linear Regression data mining techniques. The analyses of clustering quality metrics, DBSCAN give better clustering quality than PAM. CLARA and CLARA gives better clustering quality than the PAM. The proposed work can also be extended to analyze the soil and other factors for the crop and to increase crop production under the different climatic conditions.

Opinions are very important in the agriculture field because we always want to know opinions about any government services or conditions of the farmer. Local and central governments should also know public reviews about their policies or services. Therefore, there is a need for opinion mining. Some techniques are required to process the data.

There are several types of applications of data mining techniques in the field of agriculture, related to reviews through people or farmer. Data mining techniques are used to find the records and extract the information from repositories (dataset). The information is hidden in the dataset. Data mining techniques can be used to extract meaningful information and transfer useful knowledge. Data is divided into two groups (1) Classification (2) Clustering. Basically, classification is supervised learning techniques and clustering is considered as an unsupervised classification process [16]. A large number of clustering algorithms have been developed for a different purpose. By classification imperatively need training sets to identify similar features. The main difference is Clustering algorithm, which is mainly linear or nonlinear, whereas classification consists of linear classification i.e. support vector machine , naive bayes techniques, NN (neural networks), KNN (K- Nearest neighbors), Kernel estimation, decision trees. An attempt is made for opinion mining for checking the polarity values in terms of positive and negative values for Roman Language [17]. The authors concluded that opinions words are adjective words and are related under positive and negative context. Feature based clustering is done for words which are dependent on context in [18]. In this research, k nearest neighbor classifier is used for checking the polarity of given context words. The polarity of ambiguous for context words is discussed in [19]. This research highlighted ways to determine polarity values for words which are ambiguous in nature. Mining of huge amounts of web data is done in [20]. This research made use of back propagation neural network algorithm for such classification of words in given data.

III. DATA COLLECTION

For conducting this experiment, we collected the English-Punjabi code mixed data from micro-blogging sites from 13, December 2017 to June 31, 2018 (around 6 months). The data was collected using Twitter API, Facebook Graph API, and YouTube. The comments were collected on the agricultural domain. Total 95800 comments were collected after cleaning the data and records have been put in the repository. Table 1 shows the corpus statistics and data collected for the system. Data collection of relevant words per website is 30% from Twitter, 70% from YouTube and 45% from Facebook.

Websites	**Total Links**	**Average Comments per Link (appx.)**	**Relevant Comments to this work per Link**
https://twitter.com/	50	153	30%
https://www.youtube.com/?gl=IN	296	200	70%
https://www.facebook.com/	193	150	45%

Table 1. English-Punjabi Code mixed data collection.

IV. PROPOSED APPROACH OF SENTIMENT ANALYSIS

The process of sentiment analysis can be divided into main three parts: (A) Pre-Processing the text data (B) Feature identification and extraction (C) Sentiment classification such as positive, negative or neutral. These steps have been discussed below.

A. Pre-Processing of the Data.

User-generated content on the web is rarely present in a form usable for learning. It is very important to normalize the text by applying some pre-processing steps to make efficient data. In notepad++, using regular expressions, we used the set of pre-processing steps to remove the unrequired symbols i.e. Hashtags, questions marks etc.

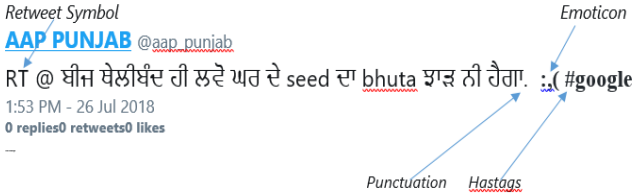

Figure 1. Example of a tweet with various features.

Data cleaning is a major step after data collection. A lot of unwanted information comes with the data, which is not related to specific operations. As social media, the text is often non- traditional and informal in nature. It is required to remove the punctuations from the text, remove multiple character repetitions, remove hashtags, web links, remove extra spaces in the Punjabi language etc. Therefore, it has to be filtered and cleaned for further processing i.e. accuracy, completeness, and consistency for sentiment analysis. The data cleaning for collected tweets is done with the help of python's Scikit Library. Table 2 shows the data cleaning operation, which is required for sentiment analysis.

Language	Original Comment	After Cleaning
Pun-English	bilkul shi keha....food nu kadi waste ni krna chaida....salute paji☺	bilkul shi keha food nu kadi waste ni krna chaida salute paji
English+ Punjabi	ਵੀਰ ਨੂੰ ਅੱਜ ਸੇਕ ਲੱਗਾ, ਇਕ ਦਿਨ ਘਾਟਾ ਪਿਆ।ਓਸ kisan ਬਾਰੇ ਸੋਚਿਆ ਜਿਹੜਾ ਵਰਿਆ ਸਾਲਾਂ ਤੋਂ ਇੰਨਾ ਘਾਟਾ ਖਾ ਕੇ ਵੀ ਜਿਓਦਾ ਹੈ so bad.	ਵੀਰ ਨੂੰ ਅੱਜ ਸੇਕ ਲੱਗਾ ਇਕ ਦਿਨ ਘਾਟਾ ਪਿਆ ਓਸ kisan ਬਾਰੇ ਸੋਚਿਆ ਜਿਹੜਾ ਵਰਿਆ ਸਾਲਾਂ ਤੋਂ ਇੰਨਾ ਘਾਟਾ ਖਾ ਕੇ ਵੀ ਜਿਓਦਾ ਹੈ so bad

Table 2. Showing the Data Cleaning

Hashtags are special characters or symbols. These are mostly used in naming subjects that are currently in trending

topics like #iPad, #news. Using regular expression, we removed special symbols.

Hyperlinks are links to other websites that are also common in comments. Twitter shortens them using its in house URL shortening service for example https://newsroom.fb.com/products/. The point of view the classification of text, any type of URL is not important. To detected the URL and removed hyperlinks using the regular expression.

Type of Problems	Original Comment	Regular Expression	Result
Hashtag	RT@ ਬੀਜ ਥੇਲੀਬੰਦ ਹੀ ਲਵੋ ਘਰ ਦੇ seed ਦਾ bahutaaaa ਝਾੜ ਨੀ ਹੇਗਾI https://www.soylent.com/#/	#(\w+)	RT@ ਬੀਜ ਥੇਲੀਬੰਦ ਹੀ ਲਵੋ ਘਰ ਦੇ seed ਦਾ bahutaaaa ਝਾੜ ਨੀ ਹੇਗਾI https://www.soylent.com//
Hyperlinks	RT@ ਬੀਜ ਥੇਲੀਬੰਦ ਹੀ ਲਵੋ ਘਰ ਦੇ seed ਦਾ bahutaaaa ਝਾੜ ਨੀ ਹੇਗਾI https://www.soylent.com//	(http|https|ftp)://[a-zA-Z0-9\\./]+	RT@ਬੀਜ ਥੇਲੀਬੰਦ ਹੀ ਲਵੋ ਘਰ ਦੇ seed ਦਾ bahutaaaa ਝਾੜ ਨੀ ਹੇਗਾI
Language Identification	RT@ ਬੀਜ ਥੇਲੀਬੰਦ ਹੀ ਲਵੋ ਘਰ ਦੇ seed ਦਾ bahutaaaa ਝਾੜ ਨੀ ਹੇਗਾI	[^ं-৭ওo-9A-Za-z.।?!\r\n]+	RTਬੀਜ ਥੇਲੀਬੰਦ ਹੀ ਲਵੋ ਘਰ ਦੇ seed ਦਾ bahutaaaa ਝਾੜ ਨੀ ਹੇਗਾ
Multiple Character Repetitions	RTਬੀਜ ਥੇਲੀਬੰਦ ਹੀ ਲਵੋ ਘਰ ਦੇ seed ਦਾ bahutaaaa ਝਾੜ ਨੀ ਹੇਗਾI	(.)\1{1,}	RTਬੀਜ ਥੇਲੀਬੰਦ ਹੀ ਲਵੋ ਘਰ ਦੇ seed ਦਾ bahuta ਝਾੜ ਨੀ ਹੇਗਾ

Table 3. Examples for problems solved by regular expressions.

Language Identification is used for extracting relevant data from the data repository. They are expressing their opinion using our mother language or state language. So it very hard to extract the required data. Therefore, we explored the relevant data to location wise on social networking sites where English-Punjabi code mixed data has been used. Extracted has been the relevant data using regular expression.

Multiple Character Repetitions is mostly used on microblogging websites where the text is written in informal languages like “I am very happyyyyyyyyy”, “We won, yaaaahhhhhhhh!”. Certain characters are repeated many times. This type of abbreviations is very hard to deal because these are not matched to any dictionary. We replaced repeating more than twice as two characters by using regular Expression for Multiple Character Repetitions.

Emoticons are very prevalent throughout the web. For the purpose of this research, all the emoticons are removed.
Punctuations and unwanted data like stop words; question mark can also provide information about the sentiments of the text. We replace any word boundary by a list of relevant punctuations present at that point. We remove any single quotes that might exist in the text. We tried to make this text as noise free as possible.
Table 3 shows the various problems solved by regular expressions in terms of few examples.

Abbreviations are also mostly used while users comment in slang words. These words are usually important for sentiment analysis but not be easy to analyze the sentiments.

We used an abbreviations list to normalize all these words. For example, ‘nyc’ was replaced by nice.

B. Feature Extraction

For creating a supervised learning model, features are very important for getting better results. The features, which are mentioned below:

a) The Number of word matches with English-Punjabi sentiment words (EPSW): We have collected a list of positive and negative word list from the repository list for sentiment classification. By manual, we expanded the Punjabi positive and negative list of words and split every sentence into N-gram. It contains 11077 positive words and 13762 negative words.
b) The Number of Slangs Words or ill words: We have also collected slangs word e.x. ‘bitch’,’ਸਾਲਾ’ etc., based on online resources. Also, used to list find out all the words, which have been used with negative sentiment analysis.
c) The Number of Character Repetitions: It is observed that most of the users used to repeat number of character of consonants or vowels to stress their opinion. On microblogging websites, user typed the text like ‘happyyyyyyy’, ‘gteatttttttt’.This type of characters often indicative of opinion mining.

d) N-Grams: N-gram denoted the continuous sequence of n numbers of items from the given text. We used the unigrams, bigrams, and trigrams. N-Gram plays an important role in context capture. For example, the Punjabi word 'ਚੰਗਾ (whose meaning is 'good' in English) is a positive word with the negative word prior to it; such as 'ਨਹੀ' (whose meaning is 'not' in English) makes it is a negative word.

Example: (ਇਹਨਾ ਕੋਲ ਕੋਈ)pun (agriculture PM)eng (ਨਹੀ ਖੇਤੀ ਨੀਤੀ ਕਿਥੋ ਹੋਊ)pun
Unigrams-[{ਇਹਨਾ}, {ਕੋਲ},{ਕੋਈ},{agriculture},{PM},{ਨਹੀ },{ਖੇਤੀ},{ਨੀਤੀ},
{ਕਿਥੋ}, {ਹੋਊ}]
Bigrams- [{ਇਹਨਾ ਕੋਲ}, {ਕੋਈ agriculture},{PM ਨਹੀ },{ਖੇਤੀ ਨੀਤੀ},{ਕਿਥੋ ਹੋਊ}]
Trigrams- [{ਇਹਨਾ ਕੋਲ ਕੋਈ},{agriculture PM ਨਹੀ },{ਖੇਤੀ ਨੀਤੀ ਕਿਥੋ},{ਹੋਊ}]

C. Classification of Sentiment

In this paper, we has used English-Punjabi code mixed social media data has been collected from Facebook, YouTube and, Twitter. After that, pre-processing of data normalize irregular words. Using regular expressions to remove the noise of text data, stop words and translate the abbreviations to regular words to make meaningful text data. Total 95800 comments collected from the social media network, related to the agricultural field. Then, split the data into positive and negative polarity as sentence wise and store to positive sentences in positive list and negative sentences in the negative list. Those sentences that are not sensed in positive or negative, store to neutral list. When the system has been trained, the bag of words has been calculated in the training phase of the system. The system takes input sentences in English-Punjabi language. Using dataset 70% of comments are selected for the training phase and 30% comments for the testing phase. In the training phase, the system is trained to classify and analyze the English-Punjabi code mixed text sentence. The training data for English text classification is available on the net, but for English-Punjabi code mixed data is not available. For training the systems, required corpus has been gathered the data where English-Punjabi code mixed data available. In the training phase, words frequency has been calculated for both the training data corpus by the following equation:

$$\text{POL-POSITIVE} = \sum_{i=0}^{n} POSITIVE\_SCORE_i$$

$$\text{POL-NEGATIVE} = \sum_{i=0}^{n} NEGATIVE\_SCORE_i$$

Finding Polarity Score

By using this step, system check the polarity score such as Positive and Negative words in the comments, Some of the words do not affect the polarity of the score like Country name, Village name, State name etc. such words has made and named as "BAG OF WORDS". Therefore, if these words occur in the comments then their polarity score is ignored. System analyses positive as well as negative polarity score using the formula as given below.

$$\text{POL-POSITIVE} = \sum_{i=0}^{n} POSITIVE\_SCORE_i$$

$$\text{POL-NEGATIVE} = \sum_{i=0}^{n} NEGATIVE\_SCORE_i$$

Where POL_POSITIVE is the positive polarity score of the comments, POL_NEGATIVE is the negative polarity score of the comments and n is the total number of words in the comments, POSITIVE_SCOREi is the positive score of the word currently processed and NEGATIVE_SOCREi is the negative score of the word currently processed.

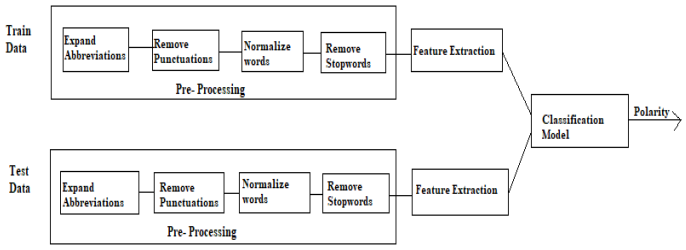

Figure 2. Training and Testing Process.

V. IMPLEMENTATION AND RESULTS

The results depict the opinion of people related to agriculture. The opinions are divided into three categories positive, negative and neutral. While performing experiments for sentiment analysis in English-Punjabi code mixed data has been extended. The results for sentiment analysis using Support Vector Machine and Naive Bayes Technique are shown in Figure 3 and plotted in Figure 4.

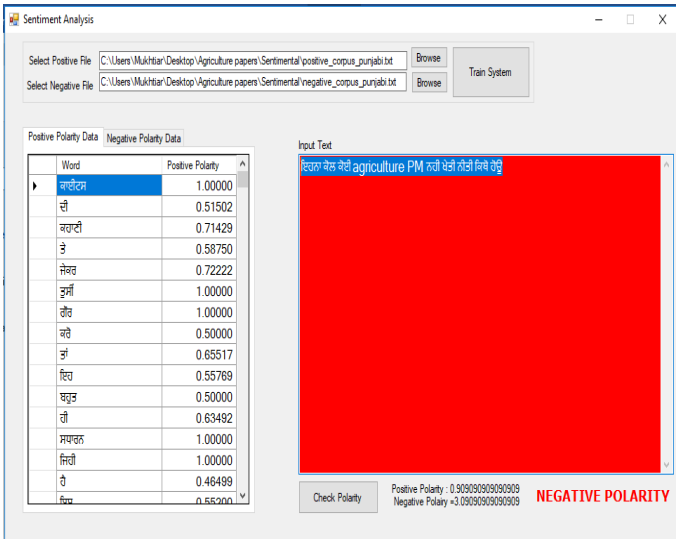

Figure 3. Result of Sentiment Analysis.

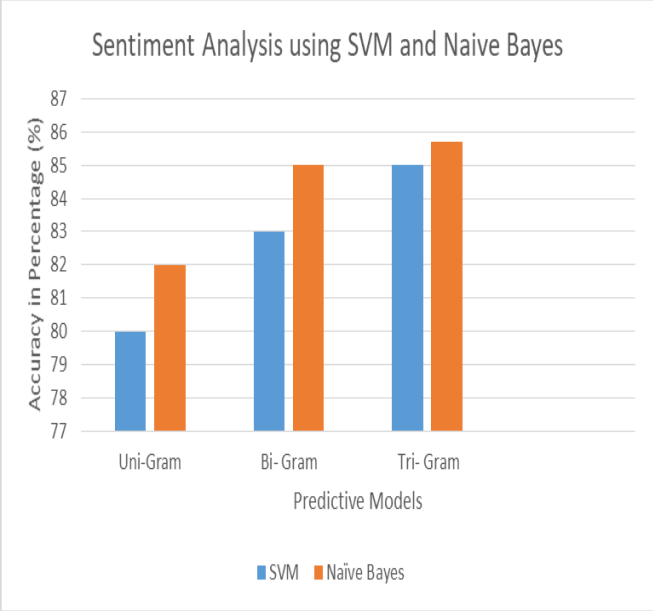

Figure 4. Sentiment analysis using Support Vector Machine and Naive Bayes technique**.**

VI. CONCLUSION AND FUTURE WORK

In this paper, we analyzed the sentiment analysis in the field of agriculture. We created a sentiment classification using labeled dataset. The agriculture data has been extracted from social media and has been performed sentiment analysis of English-Punjabi code mixed data has been using as Support Vector Machine and Naive Bayes Technique. This research work process first tested the pipeline for unigram predictive model and accuracy achieved. Later the process has been repeated for n-grams and performance enhanced, which is marginally better in comparison to the unigram model. This work can be extended in the future by checking the sentiments of emoticons. The challenge for including these emoticons while using the availability of special symbols. To identify the emoticons and replace them with a single word is a very challenging task.

REFERENCES

[1] Veenadhari, S., Bharat Misra, and C. D. Singh. "Data mining techniques for predicting crop productivity—A review article." IJCST, Vol. 2, Issue 1, 2011, pp 90-100.

[2] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005.

[3] Mishne, Gilad. "Experiments with mood classification in blog posts." Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access. Vol. 19. 2005.

[4] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR), Vol. 31, Issue 3, 1999, pp 264-323.

[5] Wood, Brennon A., et al. "Agricultural science in the wild: A social network analysis of farmer knowledge exchange." PloS one, Vol. 9, Issue 8, 2014, pp 105203.

[6] Bhagawati, Kaushik, et al. "Application and Scope of Data Mining in Agriculture." International Journal of Advanced Engineering Research and Science, Vol 3, Issue 7.

[7] Milovic, B., and V. Radojevic. "Application of data mining in agriculture." Bulgarian Journal of Agricultural Science, Vol. 21, Issue 1, 2015, pp 26-34.

[8] Mucherino, Antonio, Petraq Papajorgji, and Panos M. Pardalos. "A survey of data mining techniques applied to agriculture." Operational Research, Vol. 9, Issue 2, 2009, pp. 121-140.

[9] Cunningham, Sally Jo, and Geoffrey Holmes. "Developing innovative applications in agriculture using data mining." The proceedings of the Southeast Asia regional computer confederation conference. 1999, pp. 25-29.

[10] Mirjankar, Namita, and Smitha Hiremath. "Application of Data Mining In Agriculture Field." International Journal of Computer Engineering and Applications, iCCSTAR-2016, Special Issue 2016.

[11] Palepu, R. B., and R. R. Muley. "An analysis of agricultural soils by using data mining techniques." International Journal of Engineering Science and Computing October 2017), pp 15167-15177.

[12] Fetanat, Hooman, Leila Mortazavifar, and Narsis Zarshenas. "The Application of Data Mining Techniques in Agricultural Science." Ciência e Natura, Vol. 37, 2015, pp. 108-116.

[13] Valsamidis, Stavros, et al. "A framework for opinion mining in blogs for agriculture." Procedia Technology, Vol. 8, 2013, pp. 264-274.

[14] Kalpana, R., N. Shanthi, and S. Arumugam. "A survey on data mining techniques in agriculture." International Journal of Advances in Computer Science and Technology, Vol. 3, Issue 8, 2014, pp. 426-431.

[15] Majumdar, Jharna, Sneha Naraseeyappa, and Shilpa Ankalaki. "Analysis of agriculture data using data mining techniques: application of big data." Journal of Big Data, Vol. 4, Issue 1, 2017, pp. 20.

[16] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer, Berlin, Heidelberg, 2006, pp. 25-71.

[17] Rathi, Shivam, Shashi Shekhar, and Dilip Kumar Sharma. "Opinion mining classification based on extension of opinion mining phrases." Proceedings of International Conference on ICT for Sustainable Development. Springer, Singapore, 2016.

[18] Garg, Sonal, and Dilip Kumar Sharma. "Feature based clustering considering context dependent words." 2015 1st International Conference on Next Generation Computing Technologies (NGCT). IEEE, 2015.

[19] Garg, Sonal, and Dilip Kumar Sharma. "Sentiment Classification of Context Dependent Words." Proceedings of International Conference on ICT for Sustainable Development. Springer, Singapore, 2016.

[20] Kaur, Sukhpal, and Er Mamoon Rashid. "Web news mining using Back Propagation Neural Network and clustering using K-Means algorithm in big data." Indian Journal of Science and Technology, Vol. 9, Issue 41, 2016.