IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 3, Ser. I (May – June 2019), PP 77-81
www.iosrjournals.org

Sentiment Analysis of Hindi Language data for Agriculture Domain

Sandeep Rai
(Assistant Professor, Computer Science and Engineering, Technocrats Institute Technology (Excellence), Bhopal, India)

Abstract: *Presently, it has become relatively easy to gather feedback from farmers through micro-blogging websites. Over the years, a trend has emerged where individuals proficient in multiple languages often switch between them to express themselves on social media platforms. In this study, the authors have collected comments related to agriculture that exhibit code-mixing, specifically incorporating Hindi content. They performed language identification, normalization, and created a Hindi code-mixed dictionary. They subsequently tested various models trained on Hindi code-mixed data using LSTM, CNN and Naive Bayes techniques for sentiment analysis, finding improved results with their implemented model. Online media has become a prominent platform for expressing interests and criticizing organizations and government policies. Every internet user has the freedom to express their views and share their sentiments on such platforms. Agriculture is the primary livelihood for at least 70% of the Indian population. To express their discontent, these communities have utilized multiple languages, making sentiment analysis on such platforms a challenging task. Focus of the study lies on the accuracy and performance of the agriculture dataset for predicting sentiment on the test dataset.*

Keywords: *Sentiment Analysis, Hindi, LSTM, Naive Bayes, CNN, Agriculture.*

I. Introduction

Presently, it has become relatively easy to gather feedback from farmers through micro-blogging websites. Over the years, a trend has emerged where individuals proficient in multiple languages often switch between them to express themselves on social media platforms. Mixing different languages with distinct grammatical rules poses a significant challenge in itself. In this study, the authors have collected comments related to agriculture that exhibit code-mixing, specifically incorporating Hindi content. They performed language identification, normalization, and created a Hindi code-mixed dictionary. They subsequently tested various models trained on Hindi code-mixed data using Support Vector Machine and Naive Bayes techniques for sentiment analysis. They also evaluated the performance of a unigram predictive model and experimented with n-gram models, finding improved results with their implemented model. Online media has become a prominent platform for expressing interests and criticizing organizations and government policies. Every internet user has the freedom to express their views and share their sentiments on such platforms. In India, the government introduced three acts related to farmers, which have faced opposition from the farming community and other associated groups. These acts have raised concerns about their implementation, given that agriculture is the primary livelihood for at least 70% of the Indian population.

To express their discontent, these communities have utilized multiple languages, making sentiment analysis on such platforms a challenging task. The focus of the study lies on the accuracy and performance of the agriculture dataset for predicting sentiment on the test dataset.

Nowadays, online media platforms like Facebook, Twitter, and LinkedIn have become popular means of communication (Yang et al., 2013) (Fazil & Abulaish, 2018). These platforms are used by people from various age and professional groups to express their views and opinions about different products and organizational policies, as well as government actions.

India holds the second position globally in terms of agricultural production (Top Agricultural Producing Countries, n.d.), and more than 70% of the population is engaged in the agriculture sector. Although agriculture economists and consultants have largely supported these reforms within the agricultural sector, sentiment analysis of expert groups and farmers on social media platforms plays a crucial role in research and development.

In the field of sentiment analysis, researchers have predominantly focused on analyzing content written in a single language until recent years (Dashtipour et al., 2016). However, there has been a shift in researchers' interest towards mixed-code textual data written in two or more languages.

In this study, the authors extracted textual data from social media related to agriculture domain. Natural language processing techniques were used to facilitate decision-making processes across different domains (Ghosh, 2009). The research proposes a novel textual data analysis tool for sentiment analysis specifically related to agriculture domain.

The key aspects of our research are as follows:

The authors created a dictionary of English and Hindi code-mixed language using comments collected from online sources related to the agriculture domain. The dictionary was developed after an extensive cleaning process.

Our approach involves using statistical tools for sentiment analysis on the generated dictionary of English and Hindi code-mixed language. Parameters such as accuracy and F1-score are employed to evaluate the reliability of our proposed system.

The paper is divided into several sections. The first section presents the conceptual framework of sentiment analysis, while the 2nd section discusses related research work. The 3rd section explains our proposed methodology, followed by the results and experiments in the 4th section. Finally, the 5th section concludes the paper.

II. Background

Data mining plays a crucial role in extracting valuable patterns from data by employing the most suitable techniques. The choice of data mining techniques requires careful consideration, as evidenced in (Milovic et al., 2015). , which discusses the applications of data mining in agriculture. The authors have collected agricultural data and organized it to establish an agricultural information system. By utilizing data mining technology, users gain access to new and previously unseen patterns in the data, thereby generating knowledge that aids decision-making in agricultural organizations. Another study conducted by Mucherino focuses on data mining in agriculture (Mucherino et al., 2009). The author explores a system for exploring Combined Multi-level model in Document Sentiment Analysis (Jain et al., 2015). They propose a novel combination model based on phrase and sentence-level analyses, extracting relevant features for sentiment analysis. These features include word n-grams, POS tags, linguistic analysis, negation terms, degree modifiers (e.g., "very" and "much"), transitional words (e.g., "but"), and dependency relationships. The positive words are denoted as WordPosNum, negative words as WordNegNum, and the sum of positive and negative words as WordSubNum. Additionally, sentence-level sentiment analysis features are analyzed.

Rural areas in India particularly depend on agriculture (Arora, 2013). The importance of sentiment analysis has increased in the present scenario due to the rapid growth of social media (Beigi et al., 2016). Analyzing the views and opinions of people from different online platforms has become an urgent requirement. People have the freedom to express their thoughts and openly share content on the internet. Data mining techniques are also applied in the agricultural field (Veenadhari et al., 2011). Historical crop yield information holds importance for supply chain operations in industries that utilize agricultural products as raw materials. Various industries such as livestock, food, animal feed, chemicals, poultry, fertilizers, pesticides, seed, paper, and others incorporate agricultural products into their production processes. Accurate estimations of crop size and risk assist these companies in making supply chain decisions (Wood et al., 2011). The selection of relevant data for analysis is explored in (Bhagawati et al., 2011). , where data is retrieved from various sources, undergoes data cleaning in the data pre-processing stage, and is integrated from multiple sources into a common database during the data integration phase.

It has been observed that individuals who express their views and emotions on online social media platforms tend to utilize multiple languages (Lantz-Andersson, 2018). Pertaining to the agricultural domain, Fetanat et al. (2015) developed an agricultural information system by collecting and organizing agricultural datasets. Through data mining, patterns are extracted to generate a knowledge base that aids decision-making in agricultural corporations. Mittal and Agarwal (2013) analyzed different data mining techniques utilized in the agriculture domain, highlighting the potential for increased profits through effective data analysis. The importance of data mining for soil analysis was emphasized by Palepu and Muley (2017). Fetanat et al. (2015) employed regression techniques to analyze data and highlighted the impact of chlorophyll on flower coloration, utilizing agricultural data for their analysis (Fetanat et al., 2015).

III. Methodology

A) Data collection

In our experimentation, the data is collected from online platforms. The collected data is related to the agriculture domain. We selected English-Hindi code mixed data from online platforms. A total of 10362 comments have been collected. After the collection of data, the cleaning process is performed for the generation of a dictionary for better results of the proposed approach.

B) Data cleaning

Cleaning of data plays a vital role after cleaning of data. A chunk of undesirable information is always present within the data. So it becomes significant to apply preprocessing to the text so that undesirable data can be removed. Besides pre-processing step is used to remove the un-required tokens such as #tags, punctuation's, repetitions of the characters, URLs, spaces between words, emotions. In the proposed system following pre-processing steps are applied for the removal of special symbols.

- Emoticons removal: The emoticons, such as are very prevalent on micro-blogging sites all the emoticons were removed.

- Punctuation removal: In the proposed research, a punctuation mark also has removed. If exist in the text, it has removed any single quotes.

- Abbreviation removal: Abbreviations include mostly slang's. These types of words are very useful and important for sentiment analysis but it is added to the complexity of analysing sentiments. To normalize all these words the proposed method has used an abbreviations list. For example u ("you"), y ("why"), etc.

- Hashtags removal: These are the special symbols that are widely used for subject naming, like #iPad, #news.

- URL's and user references (identified by tokens "https" and "@") are removed for analyzing the text of the sentence.

- Hyperlinks These are the links to some other web-pages. These are found in the comments also. To analysing the sentiments of the farmers have removed such type of URLs from the collected data-set.

C) Data pre-processing

Tokenization is the process of breaking down a word in such a way so that computers can understand text into words. It separates a piece of text data into smaller meaningful units called tokens. White space and punctuation can be used to separate individual tokens of a sentence (Mittal & Agarwal, 2013). In the proposed system a code mixed statement of tokens is given as an input to the language Identifier.

IV. Results and Experiments

In this paper, we used Hindi code mixed online platform dataset in which total 10362 reviews related to the agricultural domain are there. Afterward, the data undergoes pre-processing and is split into positive and negative polarity on a sentence level. Positive sentences are stored in the positive list, while negative sentences are stored in the negative list. During the training phase, the system calculates the bag of words representation. The system accepts input sentences in English-Hindi language. For training and testing, a dataset is utilized, with 80% of comments assigned for training and 20% for testing. During the training phase, the system is trained to classify and analyze English-Hindi code-mixed text sentences. While training data for English text classification can be found online, obtaining training data for English-Hindi code-mixed text is more challenging. Consequently, a corpus containing the required data with English-Hindi code-mixed text is collected to train the system effectively.

We performed comparison between two classifiers, namely NB (Naive Bayes) and CNN (Convolutional Neural Network), using the HDAD (Hindi Data of Agriculture Domain) dataset. The findings indicate that NB achieves a higher level of accuracy compared to CNN. However, it is important to note that NB exhibits over fitting issues when applied to the HDAD dataset.

In contrast, LSTM (Long Short-Term Memory) demonstrates superior performance, surpassing CNN with an accuracy of 71% in sentiment analysis specifically tailored for the HDAD dataset, which focuses on Hindi language data within the agriculture domain. Precision: 0.72, Recall: 0.88, F1-Score: 0.84.

V. Conclusion

In this paper, the comparison between NB (Naive Bayes) and CNN (Convolutional Neural Network) classifiers on the HDAD (Hindi Data of Agriculture Domain) dataset reveals that NB achieves higher accuracy than CNN. However, it should be noted that NB suffers from the issue of over fitting in the HDAD dataset. On the other hand, LSTM (Long Short-Term Memory) outperforms CNN with 71% accuracy in sentiment analysis for the HDAD dataset, specifically focusing on Hindi language data in the agriculture domain.

Moving forward, there are several avenues for future research in this field. Firstly, due to Hindi being a scarce resource language, thorough data cleaning becomes crucial. Additionally, since the data labeling is automatically done based on tweet reactions, it is important to recognize that this labeling process is imperfect. Conducting human re-labeling of the entire dataset would greatly enhance its quality and reliability.

Moreover, it is worth exploring whether data cleaning techniques applied to the HDAD or other scarce resource language datasets can lead to improved model performance. Understanding the impact of data cleaning on enhancing the accuracy and effectiveness of sentiment analysis models in such datasets is a valuable area of investigation.

Lastly, experimenting with hybrid model architectures could provide insights into whether they can enhance the performance of traditional models on HDAD or other scarce resource language datasets. By combining different techniques and approaches, there is potential for achieving superior results in sentiment analysis for these challenging language datasets.

References

[1]. 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.

[2]. 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.

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

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

[5]. 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

[6]. 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.

[7]. Arora, V. (2013). Agricultural Policies in India: Retrospect and Prospect. Agricultural Economics Research Review, 26(2), 135–157. http://www.indianjournals.com/ijor.aspx?target=ijor:aerr&volume=26&issue=2&article=001&type=pdf

[8]. Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. Studies in Computational Intelligence, 639(January), 313–340. https://doi.org/10.1007/978-3-319-30319-2_13

[9]. Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y. A., Gelbukh, A., & Zhou, Q. (2016). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. Cognitive Computation, 8(4), 757–771. https://doi.org/10.1007/s12559-016-9415-7

[10]. Fazil, M., & Abulaish, M. (2018). A Hybrid Approach for Detecting Automated Spammers in Twitter. IEEE Transactions on Information Forensics and Security, 13(11), 2707–2719. https://doi.org/10.1109/TIFS.2018.2825958

[11]. Fetanat, H., Mortazavifar, L., & Zarshenas, N. (2015). The Application of Data Mining Techniques in Agricultural Science. Ciência e Natura, 37, 108. https://doi.org/10.5902/2179460x20760

[12]. Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge university press.

[13]. Ghosh, S. (2009). Application of natural language processing (NLP) techniques in e-governance. E-Government Development and Diffusion: Inhibitors and Facilitators of Digital Democracy, 122–132. https://doi.org/10.4018/978-1-60566-713-3.ch008

[14]. Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. Processing, 1–6.

[15]. Hubert, R. B., Estevez, E., Maguitman, A., & Janowski, T.

[16]. (2018). Examining government-citizen interactions on twitter using visual and sentiment analysis. ACM International Conference Proceeding Series. https://doi.org/10.1145/3209281.3209356

[17]. Hürlimann, M., Davis, B., Cortis, K., Freitas, A., Handschuh, S.,Fernández, S. (2016). A twitter sentiment gold standard for the Brexit referendum. ACM International Conference Proceeding Series, 13-14-Sept, 193–196. https://doi.org/10.1145/2993318.2993350

[18]. Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., Godfray, H. C. J., Herrero, M., Howitt, R. E., Janssen, S., Keating, B. A., Munoz-Carpena, R., Porter, C. H., Rosenzweig, C., & Wheeler, T. R. (2017). Brief history of agricultural systems modeling. Agricultural Systems, 155, 240–254. https://doi.org/10.1016/j.agsy.2016.05.014

[19]. Lantz-Andersson, A. (2018). Language play in a second language: Social media as contexts for emerging Sociopragmatic competence. Education and Information Technologies, 23(2), 705–724. https://doi.org/10.1007/s10639-017-9631-0

[20]. Majumdar, J., Naraseeyappa, S., & Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: application of big data. Journal of Big Data, 4(1). https://doi.org/10.1186/s40537-017-0077-4

[21]. Mittal, N., & Agarwal, B. (2013). Sentiment Analysis of Hindi Review based on Negation and Discourse Relation. Sixth International Joint Conference on Natural Language Processing, October, 57–62. http://www.aclweb.org/website/old_anthology/W/W13/W13-43.pdf#page=57

[22]. Mohammed Rushdi-Saleh, M. T. M.-V. (2011). OCA: Opinion Corpus for Arabic. Journal of the American Society for Information Science and Technology, 64(July), 1852–1863. https://doi.org/10.1002/asi

[23]. Mountassir, A., Benbrahim, H., & Berrada, I. (2013). Sentiment classification on arabic corpora: A preliminary cross-study. Document Numerique, 16(1), 73–96. https://doi.org/10.3166/DN.16.1.73-96

[24]. Palepu, R. B., & Muley, R. R. (2017). An Analysis of Agricultural Soils by using Data Mining Techniques. 7(10), 1–7.

[25]. Shoukry, A., & Rafea, A. (2012). Sentence-level Arabic sentiment analysis. Proceedings of the 2012 International Conference on Collaboration Technologies and Systems, CTS 2012, 546–550. https://doi.org/10.1109/CTS.2012.6261103

[26]. Yang, C., Harkreader, R., & Gu, G. (2013). Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Transactions on Information Forensics and Security, 8(8), 1280–1293.

[27]. Zavattaro, S. M., French, P. E., & Mohanty, S. D. (2015). A sentiment analysis of U.S. local government tweets: The connection between tone and citizen involvement. Government Information Quarterly, 32(3), 333–341.