REVIEW AN ASPECT BASED SENTIMENT ANALYSIS ON MIXED-INDIC SOCIAL MEDIA TEXT

Sevak Tarjani*1, Dr.Sanjay Singh Bhadoria *2
**1(School Of Computer Science & IT, Devi Ahilya Vishwavidyalaya, Indore (M.P.))*
**2 (Department of computer Science & Application, Dr.A.P.J Abdul kalam University, Indore(M.P.))*
*tarjani_85@yahoo.co.in*1, sanjaybhadoria@aku.ac.in*2*

Abstract -*It is critical to conduct sentiment analysis and opinion mining in English, Hindi, and Gujrati. For the EN-HI-GJ mixed language dataset, we will investigate how negation and discourse relations can be handled efficiently to improve sentiment analysis performance. The proposed method makes use of a social media corpus that has been compiled in a mix of English, Hindi, and Gujrati. An algorithm with negation and discourse relations has shown to improve sentiment analysis performance in experiments. It is possible to expand the dataset in the future in order to get better and more general results. Word Sense Disambiguation (WSD) and morphological variants can be incorporated into this research to improve accuracy for words with dual nature.*

Keywords: *Sentiment Analysis, Multi-Linguality, Polarity, Opinion*

1. INTRODUCTION

The massive flow of data across multiple social media platforms provides us with valuable information. This information is available in the form of online journals, comments, reviews, and other forms of expression. People nowadays prefer to share their opinions about anything and everything on social media platforms, whether it's about an item, a person, or a location. The task of processing such a large amount of data, as well as data from previously conducted studies, becomes time-consuming. This enormous amount of social information opens up a plethora of possibilities for the smooth operation of advertising campaigns and allows for the measurement of the campaign's impact on the efficient decision-support network. Two decades ago, there were no communal sites like the ones that are currently at their most popular in terms of usage. However, at the moment, social media sites are being overrun by people sharing their perspectives. There has been an increase and a decrease in the usage of various websites that are widely used by people, respectively. The number of requests for the use of different websites is represented by a ranking. Face book and Twitter are the social media platforms that share the most information. As a result, it is considered a difficult task to extract human-oriented information from the vast amount of data available on the internet as it has appeared. It also gets in the way of making the best decision possible, as previously stated. The exponential growth of data on the internet has increased the urgency with which it is necessary to remove insignificant data in order to extract information. It is possible to evaluate an individual's experience on the basis of resource-rare languages and resource-rich languages, and this evaluation is optional. Its domain is distinct from that of a small, individual thing anywhere in the world. The requirement for improved procedures in the field of sentiment generation arises as a result of the need for automation in the processing of such large amounts of data. Researchers in South Africa have discovered a nexus point between online networking and human-developed data, and

they are collaborating on projects in the country. Without question, we were approached by sentiment analysis through various circles of our daily lives, regardless of our level of comprehension. It has an impact on our purchasing habits, work, and other aspects of our lives. In a decision support system, sentiment analysis is an integral part of the process.

2. LITERATURE REVIEW

The ability of a person to make decisions is greatly influenced by the emotions associated with a particular situation. Traditionally, people would consult friends and users for their opinions on a product before making a purchase decision, or they would consult political forums before deciding who to vote for in the upcoming elections, according to tradition. As a result of the geographical limitations, the suggestions for the same were restricted due to the limited reach of the people. However, with the introduction of the internet, the reach has increased by orders of magnitude, and one can now not only obtain hundreds and thousands of reviews, but also collect them from all over the world. As we mentioned in chapter 1, social media sites have a large number of users who are active on them. Because of the large number of users, the internet is overburdened with massive amounts of data. This information is extremely valuable for sentiment analysis. Traditionally, the primary goal of Sentiment Analysis was to convert the signals of positive and negative emotions into binary forms, i.e. positive and negative emotions. Similarly to how humans experienced the technological advancement in machine learning, there was a concurrent increase in the level of granularity, with the primary motivation of researchers being to check the document for polarity. When the analysis was performed at the document level, this was the result. Later on, however, it became more focused on the sentence level (only subjective sentences were taken into consideration), and it is now even more concentrated.

R. Bhargava et al[1]In this study, we aimed to construct a system for extracting feelings from coding mixed sentences composed of English and combinations of four additional Indian languages (Tamil, Telugu, Hindi and Bengali). Due to the problem's complexity, the approach is broken into two stages: language identification and sentiment mining (see Figure 1). When compared to the baseline acquired from machine translated sentences in English, the evaluated findings are roughly 8% more exact. When it comes to identifying other languages, as well as unusual foreign or superfluous terms, the suggested approach is sufficiently versatile and robust to handle them.

D. R. Sharma et al[2]As a result of the social media revolution on the internet's world wide web, a new dimension of language mixing and the accompanying language processing has become visible. Numerous messengers, programmes, and social networking websites that accept several languages enable us to publish our messages and opinions. The aim of supervised machine learning is to map a text to one of a collection of taught languages, which is accomplished by mapping the text to a unique language from the collection of learned languages. Individuals routinely converse in a multitude of languages. It is critical to identify a native language from mixed data since it provides useful information about a person's background. The present study investigates many Supervised Learning Methods, including the K-Nearest Neighbor (KNN) approach and the character level

n-gram method, for the goal of categorizing web content written in English, Bengali, Assamese, Marathi, and Hindi.

Guha, R., et al[3]To address this issue in the present paper, we have developed and implemented an entirely new FS algorithm, which we have dubbed Hybrid Swarm and Gravitation-based FS (HSGFS). This algorithm has been applied to three feature vectors that have only recently been introduced into the literature: the Distance-Hough Transform (DHT), the Histogram of Oriented Gradients (HOG), and the Modified log-Gabor (MLG) filter Transform.

Sundar, A., et al[4] In the field of natural language processing, the task of hope speech detection has gained popularity due to the need for increased positive reinforcement online during the COVID-19 pandemic, which necessitated the development of new algorithms. Hope speech detection is concerned with identifying texts in social media comments that have the potential to elicit positive emotions in people from them.

Bains, J.K., et al[5] The computation of correct features is a critical step in the development of text recognition systems that are both efficient and accurate when measured against benchmarks. The offline text does not contain any dynamic information about the writing order or the nature of the trajectories of the strokes. Using the technique of recovery of drawing order, it is possible to retrace the trajectory of a stroke.

3. PROPOSED METHODOLOGY

The research work presented here makes a significant contribution to the field of sentiment analysis. It is primarily concerned with the extraction of hidden sentiments from social data that is available on the internet today. As a result of the concept of the internet, we are introduced to social data and methods. This avoids the difficulties associated with manually gathering opinions from a large population. Social data is available in a variety of formats, such as online reviews, comments, Twitter messages about elections, blogs for any event, or movie reviews, among other things. The primary task of sentiment analysis is to categories the opinions expressed in social data into three categories: positive, negative, and neutral opinions. Throughout the past several decades, sentiment analysis has emerged as a popular topic in the scientific community. Using IBM SPSS [2], for example, you can refine a product or any service based on the opinions of users about that product or service. LexisNexis [3] also uses news media to analyze consumers' content in order to determine their perception of a brand. There are numerous other algorithms for sentiment analysis that are being traded by a variety of organizations in addition to these. The following questions are addressed in this research work, which is motivated by the observations made above: The research work is concerned with how to deal with text that contains slang, emoticons, misspell words, and other such elements. Textual data must be normalized in order to be processed effectively, which means that emoticons must be converted to sent icons and slang must be converted to lexical words. The purpose of this study is to investigate the impact of normalization on the performance of classifiers. The datasets used in the experiment have been annotated. This research also takes into account the fact that each emoticon has a unique meaning that differs from the meanings of the others. This study extends the investigation of the task of sentiment analysis from textual data by incorporating the concept of macaronic content into the analysis. Content that assumes the indulgence

of more than one language in a single document is referred to as macaroni content. Manual annotation of the dataset in preparation for the application of various data-driven approaches used in sentiment analysis techniques. The methodologies for carrying out multilingual sentiment analysis from textual data as well as macaronic data are presented in this research work, which addresses some of the fundamental questions of sentiment analysis and provides answers to them. The paper also includes a proposal for an algorithm for temporal sentiment analysis. As a result, the final results are less reliant on outdated reviews and are more independent. According to the findings of the study, the proposed system has the potential to broaden the simplistic temporal perspective of the sentiment analyzer. Tempo Sentiscore is a new term that is introduced in this piece. With the help of explicit and implicit time variants, this captures the temporality of sentiment generation. Tempo-sentiscore is represented by an empirically derived equation, which can be used to investigate its influence on the star rating. It aids in the comprehension of temporality in the context of the generation of sent score.

Multi-linguality The internet is the most technologically advanced and widely used medium of communication today. In order to ensure that the submitted text is read by a growing number of people, writers frequently choose English as the medium of communication, recognizing that the vast majority of their readers speak the same native language. As an alternative, the writer can write the reviews in both English and his native language, if that is more convenient for him. As a result, not only does the effort required to write a document increase, but so does the effort required to maintain it. only 39.4 percent of internet users are fluent in the English language. The rest of the world's population communicates over the internet in either their native language or any other language that has a greater number of supporters in the existing systems. The following sections discuss various aspects of dealing with multilingualism in textual data:

1. The use of translation and transliteration: Previously, there had been very little work done on multilingual data sets. Because there were no high-quality translation techniques available at the time, the vast majority of SA tasks were performed on bilingual lexicons that were manually constructed. The research in South Africa has been significantly improved as a result of the use of advanced techniques such as supervised translation and transliteration.

2. Expansion of self-regulating thesaurus for use in multiple languages: In order to be able to work in multiple languages, it was discovered that parallel lexicons were required. Due to the frequent introduction of new words into natural language, it was virtually impossible to compile a dictionary that contained all of the words. It became necessary, as a result, to have an expansion of the current lexicon in order to work with multi-lingual data.

3. Collocation: A linguistic term can be translated in a number of different ways in other languages, which is known as collocation. It makes it extremely difficult to locate bilingual collocation correspondence. The collocation method entails translating a word based on the meaning of the word being translated.

4. Sentiment analysis based on implicit and explicit nature: The use of translation approaches made it simple to extract entities explicitly as well as sentiment from the data. Finding implicit entities,

on the other hand, continues to be a challenging task. It was discovered that a parallel corpus was required in order to perform SA implicitly.

5. Multilinguism associated with sentences: Any sentence that contains the base language is easier to deal with than a sentence that does not. The majority of foreign language words are considered stop words in sentences that contain multi-lingual words, which is often a valuable task for extracting opinions when a sentence contains multi-lingual words. Every approach to translation and transliteration currently available is incapable of detecting words without an explicit knowledge of the language in which they are spoken.

4. CONCLUSION

The overall goal of our research work was to establish a clear connection between various aspects of sentiment generation in order to better understand them. Due to the vast amount of information available on social media, a reliable senti-score generation methodology should be developed and implemented. This chapter summarizes all of the research that has been done throughout this thesis, as well as the conclusions reached. The first section of this chapter is devoted to a summary of the work. This section summarizes the overall study's main findings, which can be further connected to current research for the improvement of decision support systems in subsequent sections.

Reference

[1]. R. Bhargava, Y. Sharma and S. Sharma, "Sentiment analysis for mixed script Indic sentences," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 524-529, doi: 10.1109/ICACCI.2016.7732099.

[2]. D. R. Sharma and S. Bhaskaran, "Categorization of Multilingual Text on Languages of Indic Script," 2019 Global Conference for Advancement in Technology(GCAT), 2019, pp. 1-4, doi: 10.1109/GCAT47503.2019.8978281.

[3]. Guha, R., Ghosh, M., Singh, P.K. et al. A Hybrid Swarm and Gravitation-based feature selection algorithm for handwritten Indic script classification problem. Complex Intell. Syst. 7, 823–839 (2021). https://doi.org/10.1007/s40747-020-00237-1

[4].Sundar, A., Ramakrishnan, A., Balaji, A. et al. Hope Speech Detection for Dravidian Languages Using Cross-Lingual Embeddings with Stacked Encoder Architecture. SN COMPUT. SCI. 3, 67 (2022). https://doi.org/10.1007/s42979-021-00943-8

[5].Bains, J.K., Singh, S. & Sharma, A. Dynamic features based stroke recognition system for signboard images of Gurmukhi text. Multimed Tools Appl 80, 665–689 (2021). https://doi.org/10.1007/s11042-020-09653-4

[6].Gandhi, S. The Persian Writings on Vedānta Attributed to BanwālīdāsWalī. J Indian Philos 48, 79–99 (2020). https://doi.org/10.1007/s10781-019-09415-z

[7].Akhil, K.K., Rajimol, R. & Anoop, V.S. Parts-of-Speech tagging for Malayalam using deep learning techniques. Int. j. inf. tecnol. 12, 741–748 (2020). https://doi.org/10.1007/s41870-020-00491-z

[8].Khan, G.F. Social media-based systems: an emerging area of information systems research and practice. Scientometrics 95, 159–

180 (2013). https://doi.org/10.1007/s11192-012-0831-5

[9]. Peng, KL., Lin, MC. & Baum, T. The constructing model of culinary creativity: an approach of mixed methods. Qual Quant 47, 2687–2707 (2013). https://doi.org/10.1007/s11135-012-9680-9

[10]. Rutenfrans-Stupar, M., Hanique, N., Van Regenmortel, T. et al. The Importance of Self-Mastery in Enhancing Quality of Life and Social Participation of Individuals Experiencing Homelessness: Results of a Mixed-Method Study. Soc Indic Res 148, 491–515 (2020). https://doi.org/10.1007/s11205-019-02211-y

[11]. Zhou, L., Lin, H. & Lin, YC. Education, Intelligence, and Well-Being: Evidence from a Semiparametric Latent Variable Transformation Model for Multiple Outcomes of Mixed Types. Soc Indic Res 125, 1011–1033 (2016). https://doi.org/10.1007/s11205-015-0865-1

[12]. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, LSM '11, pages 30{38, Stroudsburg, PA, USA, 2011. ISBN 978-1-932432-96-1.

[13]. N. R. Prabhugaonkar, A. Nagvenkar, and R. Karmali, "IndoWordNet Application Programming Interfaces," 2012.

[14]. L. Gohil and D. Patel, "A sentiment analysis of Gujarati text using Gujarati SentiWordNet," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 9, pp. 2290-2293, 2019, doi: 10.35940/ijitee.I8443.078919.

[15]. Mukesh Yadav and Varunakshi Bhojane. Semi-supervised mix-hindi sentiment analysis using neural network.In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pages 309–314. IEEE, 2019.

[16]. Sonali Rajesh Shah , Abhishek Kaushik , Sentiment Analysis On Indian Indigenous Languages: A Review On Multilingual Opinion Mining, November 2019. DOI:10.20944/preprints201911.0338.v1

[17]. Somnath Banerjee, Alapan Kuila, Aniruddha Roy Sudip K. Naskar, Paolo Rosso and Sivaji Bandyopadhyay. A Hybrid Approach for Transliterated Word-Level Language Identification: CRF with Post-Processing Heuristics. Proceedings of the Forum for Information Retrieval Evaluation 2014, pp. 54–59.

[18]. Das A. and Bandyopadhyay S. Subjectivity Detection in English and Bengali: A CRF-based Approach. In the Proceeding of ICON 2009.

[19]. Das, A and Bandyopadhyay, S. (2010) SentiWordNet for Indian Languages. Proceedings of the 8th Workshop on Asian Language Resources (ALR), Beijing, 21-22 August 2010, 1-8.