A Study on Sentiment Analysis for Low-Resource Language with Emphasis on Khasi Language

Sufal Das

Department of Information Technology

North-Eastern Hill University, Shillong 793022 India

sufal.das@gmail.com

Abstract*: *Sentiment Analysis is an NLP task of finding the opinion and classifies the opin- ion expressed in a text according to its polarity (e.g., positive, negative or neutral). Low resource sentiment analysis refers to the task of perfor ming sentiment analysis on text data with limited annotated data available. This is a common scenario in many real-world applications, where annotating large amounts of text data can be time-consuming, expensive, or even impossible. To overcome this challenge, various methods have been proposed to perform sentiment analysis with limited annotated data, such as transfer learning, multi-task learning, unsupervised learning, and active learning. In this paper, we look into works done on low-resource language sentiment analysis, compare the approaches no these papers and compiling the success, challenges and pending issues on them. This paper gives an outline of how sentiment analysis is performed and presents a set of prerequisite before applying sentiment analysis on Khasi Text.

***Keywords*: Sentiment Analysis, Khasi Language, Sentiment Classification, Opinion Mining, Emotion Mining**

(Article history: Received: 3rd September 2023 and accepted 14th July 2024)

I. INTRODUCTION

Sentiment analysis is the process of analyzing if a block of text is positive, negative, or neutral. Sentiment analysis is the contextual mining of words that reveals the social sentiment. The purpose of sentiment analysis is to assess people's opinions in order to assist businesses grow. It focuses not only on polarity (positive, negative, and neutral), but also on emotions (happy, sad, angry, etc.). It employs a variety of Natural Language Processing methods, including Rule- based, Automatic, and Hybrid. The method of identifying the sentiment represented in a text document with minimal annotated data available for training is referred to as low resource sentiment analysis. This is a prevalent issue in many real-world applications where gathering and annotating massive amounts of text data can be expensive, time-consuming, or even impossible. To solve this issue, different strategies in the field of natural language processing have been developed to do sentiment analysis with little annotated data. Transfer learning, multi-task learning, un- supervised learning, and active learning are among the strategies that may be generically characterized. It is important to note that the effectiveness of these strategies for low- resource sentiment analysis is dependent on a variety of circumstances, including the task at hand, the nature of the data, and the availability of other sources of information. Furthermore, to attain optimal performance, these approaches may need domain-specific changes and fine-tuning.

II. ISSUES ON SENTIMENT ANALYSIS

Despite its widespread use and popularity, sentiment analysis is still a challenging task in NLP, and it faces several issues that limit its performance and accuracy. Some of the key issues in sentiment analysis are:

◦ Subjectivity: The nature of sentiment analysis is inherently subjective, as the sentiment expressed in a text document is not always a straightforward and objective representation of the writer's feelings. The sentiment expressed in a text document can depend on various factors, such as the writer's perspective, the context in which the text was produced, and the writer's cultural background.

◦ Ambiguity: In sentiment analysis, ambiguity can present a significant challenge, as words can often have multiple meanings and sentiments. This can make it difficult for sentiment analysis models to accurately determine the sentiment expressed in a text document.

◦ Sarcasm: Sarcasm is a common form of irony that can be very challenging for sentiment analysis algorithms to account for. It is challenging for sentiment analysis algorithms to precisely discern the sentiment portrayed in a text document since sarcasm is frequently employed to express the opposite of what is truly meant.

◦ Negation: Negation words like "not" and "never" are crucial in setting the sentiment a written document is expressing. It can be a challenge for sentiment analysis models to precisely discern the sentiment communicated in a text document since these phrases might flip the emotion of a remark.

◦ Idiomatic: Human language frequently contains idiom terms, which may be very challenging for sentiment analysis algorithms to account for. A phrase with a figurative meaning that differs from the literal meaning of the words used is referred to as an idiomatic expression. For instance,

the idiomatic expression "beating around the bush" refers to avoiding a topic's direct discussion rather than really hitting a bush.

◦ Domain-specific language: There is specialized vocabulary and terminology that is used in numerous domains, such as healthcare, politics, and finance. It may be difficult for sentiment analysis algorithms to precisely discern the sentiment represented in text published in that domain due to the specialized language and terminology used.

III. Related Works

Substantial research on low-resource semantic analysis is done, for example in Urdu and Bengali. In the Khasi language, there is almost no research is completed on any text-processing task [1].SentNoB, a sentiment analysis dataset for noisy Bangla text, with three labels positive, negative, and neutral, with categories ranging from politics to agriculture. BEmoC, Bengali Emotion Corpus, is another dataset created for emotion analysis on Bengali text, with six classes of sentiment. These are a few examples of text sentiment classification that have been published and also serve as an inspiration for sentiment analysis on low resource analysis. Table 1 shows a comparison between a few sentiment analysis research and study for low resource language. Machine Translation using the concept of frequent item set or rule mining has already been established and has been completed for the language pair of English-Arabic [7] [8]. Translation triggers association rules joining phrases present in the source language, and phrases present in the target language, these rule can be adapted to perform translation for word item set, to obtain association rules for word-to-word translation, this adaptation can filter used to filter sentiment lexicons and form association rules for sentiment lexicon [6].

IV. Issues on Low-Resource Sentiment Analysis

Data scarcity: The lack of annotated data, which is necessary to train models, is one of the main obstacles in sentiment analysis for low resource languages. Text that has been manually annotated with emotions, such as positive, negative, or neutral, is known as annotated data.

TABLE I: Existing Methods Comparison

Method	Dynamic Data	Grammar Error Check	Low Resource	Emoji and Symbols	Pre-processing	Polar Lexicons	Translated Data	Labeled Data
Assamese POS LEX Analysis [2]	No	✓	✓	×	✓	×	×	✓
Translated VADER [3]	No	✓	✓	✓	✓	✓	×	×
Bengali Cross-Lingual [4]	Yes	×	✓	×	✓	×	✓	✓
Machine Learning Classifier [5]	No	✓	✓	×	✓	×	×	✓
Deep Learn- ing Classifier [5]	No	✓	✓	×	✓	×	×	✓
Verb Based Manipuri Text [6]	No	×	✓	×	✓	×	×	✓

Lack of NLP tools: Another problem in sentiment analysis for low resource languages is the lack of available NLP tools and resources. This comprises tools for text preparation such as tokenization, POS-tagging, and named entity identification. It might be difficult to extract useful information from text data and prepare it for sentiment analysis without these tools.

Cultural and linguistic differences: When compared to high resource languages, low resource languages may have distinct cultural and linguistic norms, which might alter how feeling is represented. Due to cultural and linguistic variations, sentiment analysis algorithms trained on high resource languages may be unable to reliably predict sentiment in low resource languages.

Domain-specific challenges: Sentiment analysis methods may struggle with low resource language domain-specific phrases and terminology. This can lead to inaccurate sentiment projections.

V. Issues on Sentiment Analysis with Khasi Language

Some research issues for Khasi language in the field of sentiment analysis include:

Lack of annotated data: There is a need for the creation of large, high- quality sentiment-annotated corpora in Khasi to train and evaluate sentiment analysis models.

Inadequate lexical resources: The lack of sentiment dictionaries, word embedding, and other lexical resources specific to Khasi can limit the accuracy of sentiment analysis models.

Lack of computational tools: There is a need for the development of NLP tools and pre-trained models specifically designed for Khasi to make sentiment analysis more accessible and efficient.

Cultural and linguistic differences: Sentiment analysis models trained on data from other languages may not accurately predict sentiment in Khasi due to cultural and linguistic differences between languages.

Sentiment shifting and negation: Sentiment in Khasi can shift in a single sentence, making it challenging to accurately predict sentiment. Addition- ally, negation words in Khasi can invert the sentiment expressed, which can also pose difficulties for sentiment analysis models.

VI. Conclusion

This paper explores the concept of low-resource sentiment analysis and the difficulties that come with it across multiple languages also present an experiment that aims to analyze and understand the sentiment of the text in the Khasi language, which is a low-resource language. The paper also identifies and discusses the specific challenges faced when conducting NLP tasks in the Khasi language, such as the lack of annotated data and the complexity of the language itself. This information is intended to shed light on the current state of low-resource sentiment analysis and help researchers and practitioners to better understand the challenges and limitations of NLP in low-resource languages like Khasi.

References

[1] E. Haddi, X. Liu, Y. Shi, The role of text pre-processing in sentiment anal- ysis, Procedia Computer Science 17 (2013) 26–32. doi:10.1016/j.procs. 2013.05.005.

[2] R. Das, T. D. Singh, A Step Towards Sentiment Analysis of Assamese News Articles Using Lexical Features, 2021.

[3] I. Hossain, A. Amin, Bengali vader: A sentiment analysis approach using modified vader, Ph.D. thesis (06 2018). doi:10.1109/ECACE.2019.8679144.

[4] S. Sazzed, Cross-lingual sentiment analysis in bengali utilizing a new bench- mark corpus, 2020.

[5] L. Meetei, T. D. Singh, S. Borgohain, S. Bandyopadhyay, Low resource lan- guage specific pre-processing and features for sentiment analysis task, Language Resources and Evaluation 55. doi:10.1007/s10579-021-09541-9.

[6] K. Nongmeikapam, D. Khangembam, W. Hemkumar, S. Khuraijam, S. Bandyopadhyay, Verb based manipuri sentiment analysis, Special Issue on NLPACC, International Journal on Natural Language Computing (IJNLC) 3. doi:10.5121/ijnlc.2014.3311.

[7] H. A. H. Mahmoud, H. A. Mengash, Machine translation utilizing the frequent-item set concept, Sensors 21 (4). doi:10.3390/s21041493. URL https://www.mdpi.com/1424-8220/21/4/1493

[8] P. B. S, R. K. G. K, Efficient incremental itemset tree for approximate frequent itemset mining on data stream, in: 2016 2nd International Confer- ence on Applied and Theoretical Computing and Communication Technol- ogy (iCATccT), 2016, pp. 239–242. doi:10.1109/ICATCCT.2016.7912000.