ISSN: 0974-8571 Vol.14 No. 1 June, 2022

International Journal of Computational Intelligence in Control

Multilingual Sentiment Analysis of Low Resource Language Reviews

(Urdu, Roman Urdu, Mixed with English)

1. **Mir Ahmad Khan 2. Dr. Aurang Zeb Khan 3. Muhammad Bilal 4. Ayaz Ali Khan 5. Irfan Ullah Khan**

1. PhD Scholar, Department of Computer Science, University of Science & Technology Bannu
2. Professor, Department of Computer Science, University of Science & Technology Bannu
3. Assistant Professor, Department of Computer Science & IT, University of Lakki Marwat
4. Assistant Professor, Department of Computer Science & IT, University of Lakki Marwat
5. Assistant Professor, Department of Education & Research, University of Lakki Marwat

ABSTRACT

Due to daily interaction with social media and the rapid growth of the internet, people are increasingly motivated to engage in online conversations. Due to which, big data/large volume of useful data has been generated on the web. This data contains valuable information that is highly beneficial for organizations. Many people from rural areas use their native or local languages to express reviews on social media. Extracting information from these reviews is challenging because these languages often lack a standard format. Standard languages have been thoroughly researched; however, these local languages receive less attention although having a significant number of speakers. Some research has been conducted on sentiment analysis for these languages using alternative methods, as basic resources need for information extraction are unavailable Therefore, it is proposed to develop a direct approach where resources are created from scratch, instead of using alternative methods to extract information from reviews. This approach involves creating resources for sentiment analysis in low-resource languages and has achieved a commendable accuracy of 89.5%.

1. INTRODUCTION

In this age of internet, a large volume of data has been generated on the web due to the increasing use of internet devices. This data contains hidden information that is very useful for organizations to understand their potential users [1]. Sentiment analysis, a subfield of NLP (Natural Language Processing) [2], is used to classify and identify the emotions or opinions expressed by users in the data [3]. Sentiment analysis determines whether the writer's attitude is positive, negative, or neutral [4]. Developed languages in the field of opinion mining have been thoroughly researched, and many systems have been developed to determine opinions. However, local languages such as Urdu and Roman Urdu mixed with English are not as well supported despite having a large number of speakers. A significant number of social media users prefer to use their native or local mixed language to express their thoughts/reviews about a particular service or topic i.e. (i) **Nice Bahoot کا ہے پکچرCameraنیکونمیرے نیو**(ii) **IPhone mobile kaa camera bahoot achaa ha.** (iii) **لیپ ٹاپ کا ڈسپلے بہت خوبصورت ہے۔ایچ پی**[5]. Local languages (Urdu and Roman Urdu) are not linguistically rich, leading to basic issues such as the lack of a standard format or morphological structure. These issues complicate the sentiment analysis process and the extraction of opinions from these reviews [6]. Similarly, due to

limited or non-availability of basic resources such as a dataset, tokenizer, part-of-speech (POS) tagger, and lexicon, extracting opinions from these reviews is difficult and challenging [7]. Some researchers have proposed indirect or alternative methods for extracting opinions from these reviews. To extract information from reviews posted in local languages, they are first translated into a more widely-used language, and then traditional sentiment analysis methods are applied [8]. However, this translation or indirect method is time-consuming, costly, and less accurate. Developing resources specifically for these local or low-resource languages is essential to improving the sentiment analysis process for these languages [9].

2. LITERATURE REVIEWS

An overview of the related work is presented in this section. in the field of opinion mining less attention has been given to low-resource languages (Urdu, Roman Urdu), While a substantial amount of research has been conducted in developed languages. there has been comparatively limited exploration in languages with limited resources. Some of the related studies concerning for sentiment analysis in low-resource languages.

Gupta et al. [11] used deep learning methods to extract opinions from mixed-language reviews. To tackle various challenges, they developed resources for their model. Their approach effectively addressed three key issues including opinion mining, language identification, and machine translation, with achieving good accuracy. In [12], Sharf and Rahman proposed a model for lexical normalization to handle differences in the writing of Roman Urdu text on social media platforms. Across various languages, including Chinese, Japanese, Roman Urdu, Arabic, Bangla, and Dutch, they conducted a comparative analysis of normalization processes with the goal of achieving textual regularity. Sharf et al. in [13] conducted research for sentiment analysis of Roman Urdu using discourse based approach. A large amount of related data has been gathered from various social media platforms. preprocessed the data for standard representation. Subsequently, they conducted sentiment analysis by using neural network models, on the datasets of Roman Urdu differentiating between the absence and presence of discourse elements. In [14] the author's presents a (DFST), discriminative feature spamming technique which categorizes unique terms. They used 1100 own created reviews datasets for experiment. A customized tokenizer was developed to increase accuracy and enhanced outcomes. Rafique et al. [15] used (ML) algorithms including Logistic Regression, Naive Bayes (NB) and Support Vector Machine (SVM)for sentiment analysis of Roman Urdu. Their dataset contains 806 comments, where 400 were marked as positive and 406 as negative reviews. SVM performed well as compared to other algorithms with an achieved accuracy of 87.22%.

In [16] the authors proposed "Self attention Bidirectional LSTM (SA-BiLSTM) a novel network specifically aimed to address the problem of spelling variations and inconsistence sentence structure of Roman Urdu. Through this approach they achieved good accuracy on preprocessed and normalized datasets. Mehmood et al. in [17] the authors worked on sentiment analysis of Roman Urdu. They collected about 11,000 reviews from various online forums and using a multi annotator methodology to annotated the datasets. state of the art algorithms at both word level and character level features were used for sentiment analysis of Roman Urdu, resulting in achieving good accuracy of 80.07%.

3. METHODOLGOY

Due to the unavailability of resources, local languages are not extensively explored in research [18]. To some extent, alternative methods, such as conversion to more developed languages, are used instead of direct sentiment analysis methods. Local or low-resource languages need to develop their own resources

to achieve better accuracy [19]. This section describes the detailed methodology applied to the opinion mining of low resource languages (Roman Urdu, Urdu, and mixed with English). The purpose of this method is to provide a solution for sentiment analysis of low-resource languages by creating new customized resources and enhancing existing ones with the help of field experts. In first step related data were collection from different online related forums. The data contained unwanted or noisy elements, which were removed. For each language a customized tokenizer, capable of handling all three languages, was applied to split the data into tokens. After that, a customized part of speech tagger was used to tag the low resource languages words with part speech tags. Our own datasets, which were developed in Urdu, Roman Urdu, and combined with English language, were used to train the algorithms, for each language, separate validation and test datasets were used to check the model's performance.

I. Tokenizer

In the lexical analysis phase the tokenizer receives the input text, which is Urdu, Roman Urdu or mixed English text. The tokenizer processes it and creates output tokens [20] the resulting tokens are then forwarded to the subsequent stage known syntax analysis for further processing.

Figure3.1: Tokenization Process of Urdu and Roman Urdu review

II. Part of Speech Tagger

Totag words in a sentence, a part-of-speech (POS) tagger is an essential tool [21]. With the assistance of a field expert, a customized part-of-speech tagger was constructed to tag words in the target language.

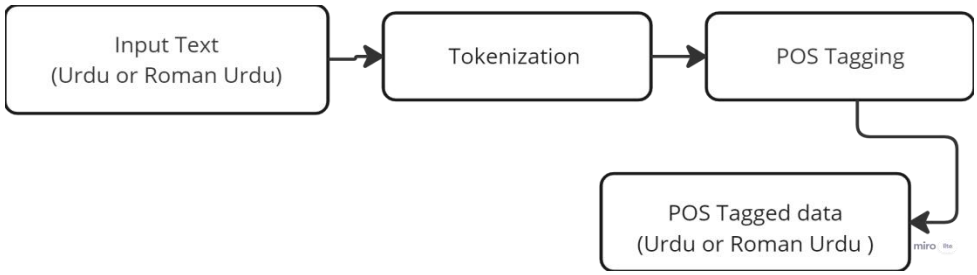

Figure 3.2: POS Tagging process of low resource languages

III. Datasets

We created our won customized datasets for Roman Urdu, Urdu, and for English mixed with this language. total 10 (ten) thousand reviews were collected in which 5,000 Positive, 3,000 negative, and 2,000 neutrals. These datasets were used to train the model. 80% of the data were used for training and 20% for testing purpose.

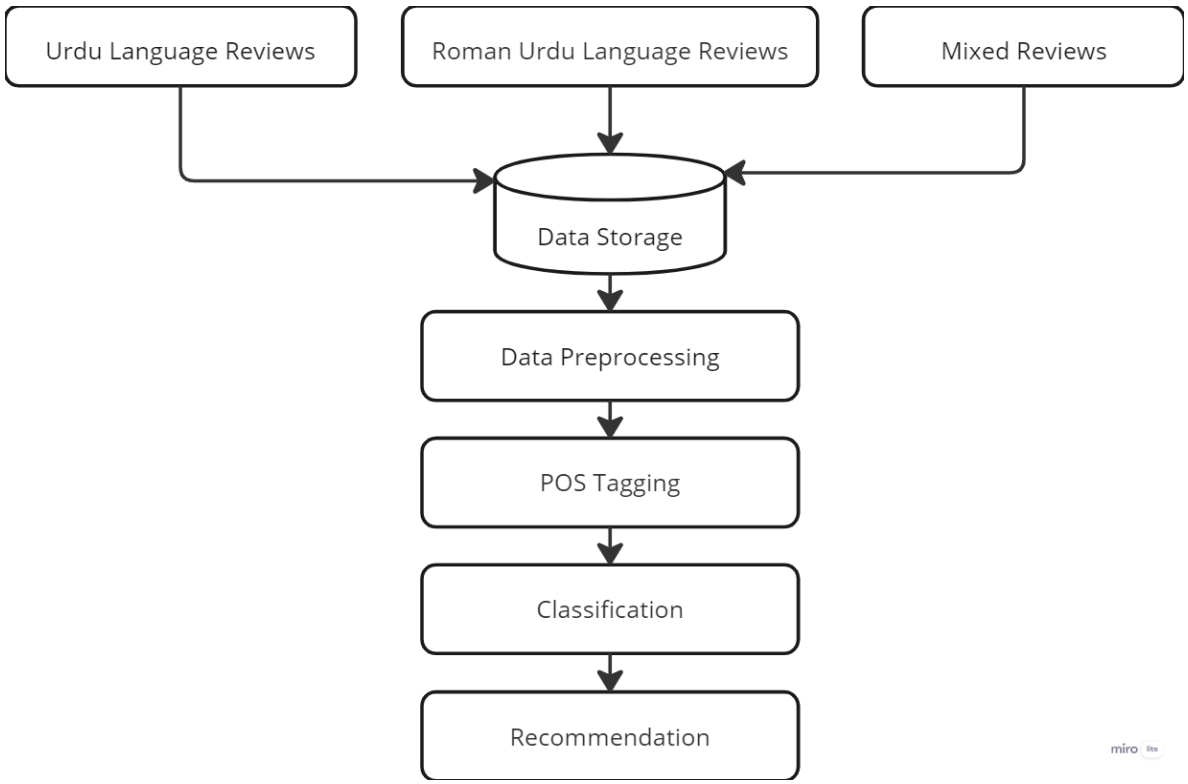

Figure 3.3: Sentiment analysis process of the target low resource languages

4. RESULT AND DISCUSSION

The results of the proposed approach for the target low resource languages are presented in this section. To classify the reviews, experiments were performed using Naive Bayes, Decision Trees, Logistic Regression and Support Vector Machines algorithms. our own customized datasets which contain data of Roman Urdu, Urdu or mixed with English were used to train the algorithms/models. the reviews were classified as positive, negative and neutral with 2000 neutral, 3000 negative and 5000 positives. sentiment labels were assigned to this reviews. 80% labeled data were used for training purpose while 20% data were used for testing purpose. Examples of Urdu language reviews with sentiment labels are presented in Table 4.1. Examples of Roman Urdu reviews are shown in Table 4.2. Examples of mixed language reviews (i.e., Urdu, Roman Urdu mixed with English) are shown in Table 4.3.

Table 4.1: Example of Urdu reviews with sentiment labels

Review Text (Urdu/Roman Urdu)	English Translation	Sentiment Label
میری کیمرہ کی کوالٹی بہت بہتر ہے۔	The quality of my camera is much better.	Positive
میرے لیپ ٹاپ کی کارکردگی اچھی ہے۔	The performance of my laptop is good.	Positive
تعلیم کی کوئی قیمت نہیں ہوتی۔	Education has no price.	Positive
میں ان سیاستدانوں کا حمایت نہیں کرتا۔	I do not support those politicians.	Negative
یہ کتاب میرے وقت کا ضائع کرتی ہے۔	This book wastes my time.	Negative
یہ کتاب دلچسپ لگ رہی ہے۔	This book seems interesting.	Positive
میں سیاست کو بہت زیادہ پسند کرتا ہوں۔	I like politics very much.	Positive
ہماری ٹیم نے شاندار کارکردگی کی۔	Our team performed brilliantly.	Positive
کتاب کی کہانی بہت بیکار ہے۔اس	The story of this book is very worthless.	Negative

Table 4.2: Example of Roman Urdu reviews with sentiment labels

Review Text Roman Urdu)	English Translation	Sentiment
IPhone kaa camera bahoot humda ha.	IPhone camera is very nice	Positive
Nikon camera kaa pic saafni ha	The picture of Nikon camera is not clear.	Negative
Mujheyehkitaabbohotpasandaayi.	I really liked this book.	Positive
Woh film bohotachithi.	That movie was very good.	Positive
Woh camera ki quality bohotacchihai.	The quality of the camera is very good.	Positive
Meri education kaafibekaarrahihai.	My education has been quite poor.	Negative
Pakistan ma education ka system ach ni ha .	Education system in Pakistan is not good.	Positive

Table 4.3: Example of Urdu, Roman Urdu and English mixed reviews with sentiment labels

Review (Urdu, Roman Urdu and English)	English Translation	Label
ہے fast بہت laptop میرا	My laptop is very fast.	Positive
-تھی helpful بہت کتاب وہ	That book was very helpful.	Positive
-تھیfilm boring بہتوہ	That film was very boring.	Negative
-ہے achi بہت quality ki camera میرا	The quality of my camera is very good.	Positive
میں politics میں interested -ہوں	I am interested in politics.	Positive
-ہے bekaar کافیeducation میری	My education is quite poor.	Negative

Table 4.4: Average Precision, Recall, and F1 score for each algorithm

Algorithm/Model	Average Precision%	Average Recall%	Average F1 Score%
Decision Trees	76.67	75.33	75.67
Naive Bayes	77.67	80.00	78.67
Support Vector Machines (SVM)	77.00	75.33	76.00
Logistic Regression	75.67	78.67	77.00

Table 4.4 shows that four different algorithms Naive Bayes, Decision Tree, Logistic Regression, and Support Vector Machine (SVM) were applied with average performance metrics: precision, recall, and F1 score. The models were trained on a dataset of reviews containing Urdu, Roman Urdu, or mixed with English language. Naive Bayes performed the best, with the highest average F1 score of 78.67% and an average recall of 80%.

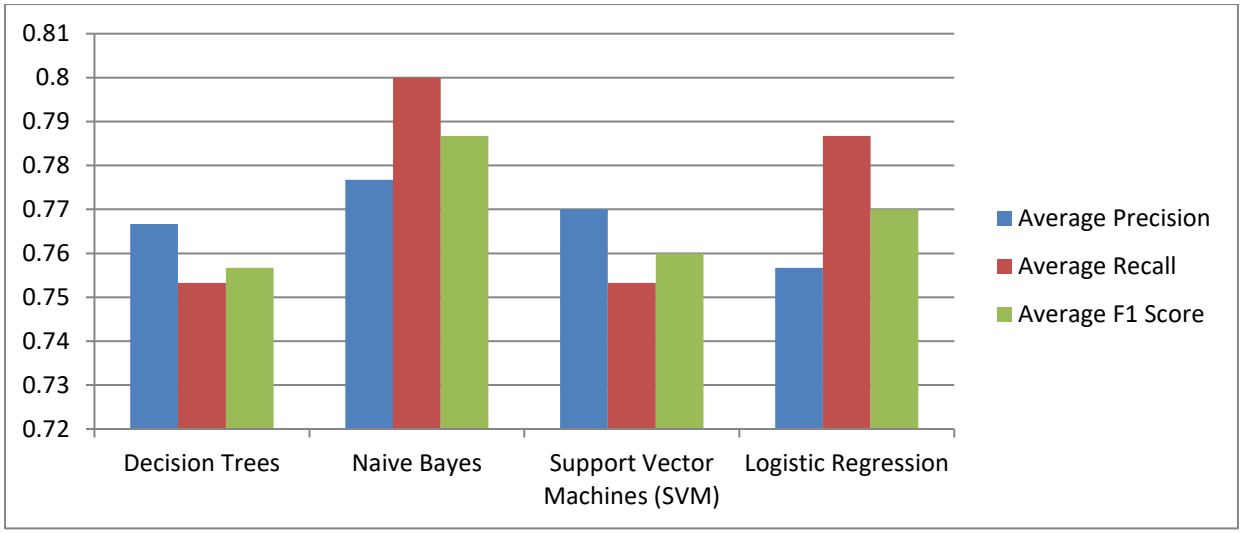

Figure 4.1:Average Precision, Recall and F1 Score of all algorithms

Graph 4.1 shows that Naïve Bayes performed well compared to other algorithms, with an average precision of 77.67%, a recall of 80%, and an F1 score of 78.67%. The average precision, recall, and F1 score for the Decision Tree are 76.67%, 75.33%, and 75.67%, respectively. SVM achieved an average precision of 77.00%, a recall of 75.33%, and an F1 score of 76%. Logistic Regression achieved an average precision of 75.67%, a recall of 78.67%, and an F1 score of 77%.

5. CONCLUSION AND FUTURE DIRECTION

The rapid growth of data generated on the web, particularly on social media, contains valuable information that can be used for various analytical purposes, such as sentiment analysis. A significant portion of social media users prefer to express their opinions and post reviews in their native or local languages. However, due to the lack of essential resources like tokenizers, part-of-speech (POS) taggers, and annotated datasets for many of these languages, researchers often resort to indirect methods, such as translating text into more resource rich languages for analysis. These conversions based methods are not only expensive and time consuming but also tend to produce suboptimal accuracy.To address this challenge, it is proposed to develop and provide resources such as tokenizers, POS taggers, and relevant datasets for these underrepresented languages, enabling direct sentiment analysis on native language reviews. This approach would establish a foundational platform for conducting sentiment analysis in these languages. However, to ensure high accuracy and reliability, there is a continuous need to improve and refine these resources. Investing in the development and enhancement of these tools will significantly advance sentiment analysis capabilities in local languages, making the process more efficient and accurate.

6. REFRENCES

1. Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, *15*(2), 94-101.
2. Cambria, E., Poria, S., Gelbukh, A., &Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, *32*(6), 74-80.

3. Yaakub, M. R., Latiffi, M. I. A., &Zaabar, L. S. (2019, August). A review on sentiment analysis techniques and applications. In *IOP conference series: materials science and engineering* (Vol. 551, No. 1, p. 012070). IOP Publishing.
4. Astya, P. (2017, May). Sentiment analysis: approaches and open issues. In *2017 International Conference on computing, Communication and automation (ICCCA)* (pp. 154-158). IEEE.
5. Kizgin, H., Dey, B. L., Dwivedi, Y. K., Hughes, L., Jamal, A., Jones, P., ... & Williams, M. D. (2020). The impact of social media on consumer acculturation: Current challenges, opportunities, and an agenda for research and practice. *International Journal of Information Management*, *51*, 102026.
6. Mehmood, F., Ghani, M. U., Ibrahim, M. A., Shahzadi, R., Mahmood, W., &Asim, M. N. (2020). A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis. *IEEE Access*, *8*, 192740-192759.
7. Khattak, A., Asghar, M. Z., Saeed, A., Hameed, I. A., Hassan, S. A., & Ahmad, S. (2021). A survey on sentiment analysis in Urdu: A resource-poor language. *Egyptian Informatics Journal*, *22*(1), 53-74.
8. Masroor, H., Saeed, M., Feroz, M., Ahsan, K., & Islam, K. (2019). Transtech: development of a novel translator for Roman Urdu to English. *Heliyon*, *5*(5).
9. Rimell, L., Lippincott, T., Verspoor, K., Johnson, H. L., &Korhonen, A. (2013). Acquisition and evaluation of verb subcategorization resources for biomedicine. *Journal of biomedical informatics*, *46*(2), 228-237.
10. Daud, M., Khan, R., & Daud, A. (2015). Roman Urdu opinion mining system (RUOMiS). *arXiv preprint arXiv:1501.01386*.
11. Gupta, A., Agrawal, A., &Rajendra, K. R. (2016). Deep learning for opinion mining of mixed-language reviews. Procedia Computer Science, 78 https://doi.org/10.1016/j.procs.2016.02.032.
12. Sharf, Z., & Rahman, S. U. (2017). Lexical normalization of roman Urdu text. *International Journal of Computer Science and Network Security*, *17*(12), 213-221.
13. Sharf, Z., & Ali, H. M. (2019). DISCOURSE BASED OPINION MINING ON ROMAN URDU DATA. *Journal of Independent Studies and Research Computing*, *17*(1).
14. Mehmood, K., Essam, D., Shafi, K., & Malik, M. K. (2019). Discriminative feature spamming technique for roman urdu sentiment analysis. *IEEE Access*, 7, 47991-48002.
15. Rafique, A., Malik, M. K., Nawaz, Z., Bukhari, F., &Jalbani, A. H. (2019). Sentiment analysis for roman urdu. *Mehran University Research Journal of Engineering & Technology*, *38*(2), 463-470.
16. Manzoor, M. A., Mamoon, S., Tao, S. K., Ali, Z., Adil, M., & Lu, J. (2020). Lexical Variation and Sentiment Analysis of Roman Urdu Sentences with Deep Neural Networks. *International Journal of Advanced Computer Science and Applications*, *11*(2).
17. Mehmood, F., Ghani, M. U., Ibrahim, M. A., Shahzadi, R., Mahmood, W., &Asim, M. N. (2020). A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis. *IEEE Access*, *8*, 192740-192759.
18. Cambria, E., Poria, S., Gelbukh, A., &Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, *32*(6), 74-80.
19. King, B. P. (2015). *Practical Natural Language Processing for Low-Resource Languages* (Doctoral dissertation).
20. Vijayarani, S., & Janani, R. (2016). Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, *3*(1), 37-47.
21. Naseem, A., Anwar, M., Ahmed, S., Satti, Q. A., Hashmi, F. R., & Malik, T. (2017). Tagging Urdu Sentences from English POS Taggers. *International Journal of Advanced Computer Science And Applications*, *8*(10), 231-238.