Weak Supervision and Transformed-based Sentiment Analysis on Multi-lingual Data

Shubhangi Rastogi
School of Computer Science Engineering & Technology
Bennett University-The Times Group
Greater Noida, India
shubhangi.rastogi@bennett.edu.in

2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS) | 978-1-6654-7706-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/COMSNETS56262.2023.10041286

***Abstract*—Sentiment analysis on social media is an important research field with applications in many domains. This can be leveraged to predict unrest situations, mob violence, etc., on the occurrence of an event concerning national security. The paper aims to develop an automated intelligent framework to detect the sentiments of users in a multilingual and geospatial manner. This paper covers an important case study to obtain a comprehensive dataset in three languages, i.e., English, Hindi, and Urdu, from three regions. An automated approach for data annotation using weak supervision is presented to highlight the challenge of obtaining training data. We propose a BERT-based transformer model to effectively identify the sentiment considering the context. The study also presents an interactive visualization of results to obtain valuable insights.**

***Index Terms*—Sentiment, Weak supervised, transformer, multi-lingual data, Kashmir case study**

I. Introduction

Social media platforms, especially Twitter, have become a common tool where the users actively display their opinions and sentiments on any topic [1]. Sentiment analysis is a popular research area in social media analysis and has applications in different domains. Business organizations leverage this area to evaluate online customer reviews. Also, political parties use this field of research to draft their political campaigns. However, limited work has been done in the national security domain covering sensitive information about any crisis or conflicts [2]. Such application of sentiment analysis allows analysts to generate early warnings regarding illicit acts and anticipate sabotage campaigns. This paper covers an important dispute, i.e., the Kashmir territorial conflict that has been the reason for two wars between India and Pakistan since the two countries acquired independence in 1947. The case study covered in this paper has been explained in Section IV. The data has been extracted from Twitter in multiple languages covering different events related to the Kashmir conflict. However, the labeling of data is a challenge due to the voluminous data extracted. To this aim, the paper proposes an automated approach to annotate data and develop a training labeled dataset.

Even though sentiment analysis is an extensively researched area, techniques that completely exploit the context, semantics, exact emotion, and sentiment present in a text are limited [3]. In the case of short texts like tweets, the problem is more complex. Therefore, we have proposed an intelligent framework based on transformers to efficiently analyze the sentiments of multi-lingual data in order to predict unrest situations in different regions of the nation. The key contributions of this work focusing on the aforementioned research challenges are summarised below:

- C1: Develop a dataset that is multi-lingual, geo-spatial, extensive metadata, and covers major events related to the Kashmir conflict in order to develop a comprehensive corpus.
- C2: Propose an automated approach for data annotation using a weak supervision technique.
- C3: Propose an intelligent transformer-based system to identify the sentiments behind the text with high efficacy.
- C4: Extensively visualize the sentiments obtained using interactive visualization techniques in order to obtain valuable insights.

The remaining sections are structured as follows. Section 2 presents a review of related works. Section 3 introduces the methods covering data collection, annotation, experimental results, and visualization. Section 4 discusses the case study and limitations of the work. Section 5 makes concluding remarks.

II. Related Work

Literature highlights several techniques to analyze sentiments of textual data. Broadly, two popular approaches to sentiment analysis are Lexicon-based, and Machine learning-based [4]. Lexicon-based are the conventional techniques for sentiment analysis. A number of lexicons have been presented in the literature, such as Textblob, NRC, VADER, AFFIN, Flair, etc. These are unsupervised approaches that do not require a labeled training dataset [5]. Regardless of the time-efficient and speed of this approach, researchers have always challenged their performance. Whereas several researchers have conducted sentiment analysis on Twitter using machine learning algorithms and obtained substantial results [6]. Rakshitha et al. performed sentiment analysis using eight machine learning algorithms and achieved the highest accuracy between 85-90% [7]. They also observed that accuracy decreases by increasing the number of classification classes. Whereas, Jacon et al. obtained an accuracy of 97% with a machine learning algorithm by introducing a novel feature [8]. Clearly, the performance of both the above-mentioned approaches vary with the data and experiment environment. To this aim, Rastogi et al. proposed an adaptive approach for sentiment analysis using both lexicon and machine learning-based algorithms [9]. They observed that the lexicon-based approach outperformed the machine learning approach with their dataset. Hence, various research has been done to compare the performance of unsupervised and supervised techniques [10]. Moreover, rich literature exists in sentiment analysis considering different topics or entities (i.e., aspects). Analyzing sentiments corresponding to a particular aspect is called aspect-based sentiment analysis [11]. Social media provides data on numerous aspects; thus, the aspect-based approach provides a fine-grained analysis. Social media platforms also provide users with a feature to post in different languages other than English. Limited work has been done in this field considering multi-lingual data. Recent developments in this domain have significantly improved sentiment extraction using natural language processing and transfer learning. In 2018, Google introduced open-sourced transformer models in the hugging face library. The high performance of BERT transformers has been observed in sentiment analysis, especially for multi-lingual data. Essentially, the literature shows an evaluation of sentiment analysis from lexicons to encoders and up to the latest transformers [12]. Sun et al. obtained state-of-the-art results for aspect-based sentiment analysis on two publicly available datasets by fine-tuning the pre-trained BERT model [13]. This work has been further improved by Liao et al. by applying RoBerta, which was introduced by Facebook in 2019 and offered an alternative optimized version of BERT [14]. The researchers focused on feature extraction of both text and aspect tokens. Several versions of BERT have been proposed, such as XLNet, FinBERT, XLM, Albert, BART, DistilBERT, etc. Each version has some advantages or disadvantages over the other. For instance, Tao et al. employed XLNet in multi-label sentiment analysis on existing data as an improvement over BERT by leveraging a generalized AR model [15]. Therefore, the literature highlights the importance of transfer-learning using transformers in sentiment analysis. However, to the best of our knowledge, no work has employed such techniques in a sensitive issue and further extended the work to obtain exciting visualizations. In this paper, we have applied state-of-the-art transformer techniques to analyze the sentiments of multi-lingual and geospatial data covering a sensitive issue. The study also presents extensive visualization, which can be used by security agencies or government authorities to predict unrest situations, mob violence, etc., in various regions. The following section covers the methodology followed to achieve this goal.

III. Materials & Methods

Figure 1 demonstrates the steps involved in the proposed methodology for sentiment analysis. The following sections explain these steps in detail.

A. Data Collection

A data crawler has been developed to extract data from Twitter using python library, *Snscrape* to overcome the Twitter data access constraints. The goal of this step is to build a geospatial and multi-linguistic corpus covering different events related to Kashmir territorial conflict. Finally, the purpose is to use this corpus to obtain the sentiments of users from the two countries, India and Pakistan, and the disputed state of Kashmir in two different languages, i.e., Hindi and English. It consists of tweets, location, time, language, user-oriented details, etc. Table I shows the events and corresponding hashtags and keywords used to extract data. Furthermore, the duration to collect data for each event has been decided by studying the timeline of events and their trending pattern on Twitter. To the best of our knowledge, a geospatial dataset in two languages considering different events related to a territorial conflict has not been given in the literature. Moreover, this dataset has been constructed to be used for fake news detection in this conflict in the future. Sentiment analysis is an important module for fake news detection. However, obtaining a balanced dataset for fake news detection is a challenging task. Even though the spread of fake news is vast, the number of fake news is always less than the number of real news on an event. Thus, this study has mindfully selected the hashtags and keywords for data collection which have been used in fake news. Table II highlights the number of tweets collected from three regions in three languages, English, Hindi, and Urdu, depending on the regional language. The tweets collected from India are in Hindi and English languages, while comparatively fewer tweets have been collected from Pakistan in Urdu and English language. The reason for less number of tweets collected from J&K state could be directly attributed to the population of social media users.

B. Data Pre-processing & Translation

Twitter data is unstructured data with syntactical errors, typos, etc.; thus, it requires to be cleaned. The collected data has been passed through the generic data cleaning pipeline to remove special characters, numbers, symbols, RTs, stopwords, white spaces, etc. [16]. Natural language processing techniques like lemmatization have also been applied. However, it has been observed using Table II that data is unbalanced in the three regions. This Unbalanced data has been dealt with the help of random oversampling, which provided us with an equal number of training labels in each dataset. However, slight improvements have been observed by using balanced data over unbalanced raw data.

C. Automated Annotation of Data

Obtaining labeled training data for supervised machine learning techniques is a challenging task that directly affects the performance of the model. Data annotation can be performed in two ways: Expert-based manual annotation and automated annotation. Expert-based manual annotation gives better efficacy with supervised machine learning, but it is time-consuming, resource-intensive, and subject to human bias.

Event	Popular hashtags/keywords	Queries to collect related fake news
Balakot Air strike (February 2019 to August 2019)	#SurgicalStrike2, #India Strikes back, #Joshishigh, #ModipunishesPak, #BalakotAirStrike, #Balakot	292(#Balakot OR #surgicalstrike), 200(#Balakot OR #surgicalstrike), IAF troll(#Balakot OR #surgicalstrike), zafar(#Balakot OR #surgicalstrike), twoindianjets(#Balakot OR #surgicalstrike)
Article 370 (August 2019 to December 2020)	#KashmirBleeds, #SaveKashmir, #KashmirwithModi, #kashmirWelcomesChange, #Article370, #RedForKashmir, #KashmirWantsFreedom, #KashmiriLivesMatter, #5AugustBlackDay	Ambedkar(#Kashmir OR #article370), bombs army(#KashmirBleeds AND #article370), Army India(#KashmirBleeds AND #article370), Army India beat(#KashmirBleeds AND #article370), army(#saveKashmir AND #article370), Yasin Malik dead(#KashmirBleeds OR #article370)
Galwan Valley (June2020 to December 2020)	#BoycottChina, #GalwanValley, #China, #India, #GalwanValleyClash, #WeStandWithIndianArmy	coffins(#galwan), 50 coffins (#galwan), 56 chinese (#galwan), list chinese (#galwan), #chinese #territory #Galwan #india, #pakistan
Pulwama Attack(February 2019 to August 2019)	#Pulwama, #PulwamaTerrorAttack, #CRPF, #KashmirTerrorAttack, #StandWithForces, #PakistanNahiSudhrega, #RemoveArticle370, #RemoveArticle370and35A, #ExposeDeshdrohis, #KashmiriMuslims	40(#Pulwama), CRPF(#Pulwama), Video(#Pulwama), Priyanka Gandhi Vadra(#Pulwama), Rahul Gandhi(#Pulwama), video(#Pulwama), 2 militants(#Pulwama)

TABLE I

EVENTS AND CORRESPONDING HASHTAGS/KEYWORDS/ QUERIES FOR DATA COLLECTION

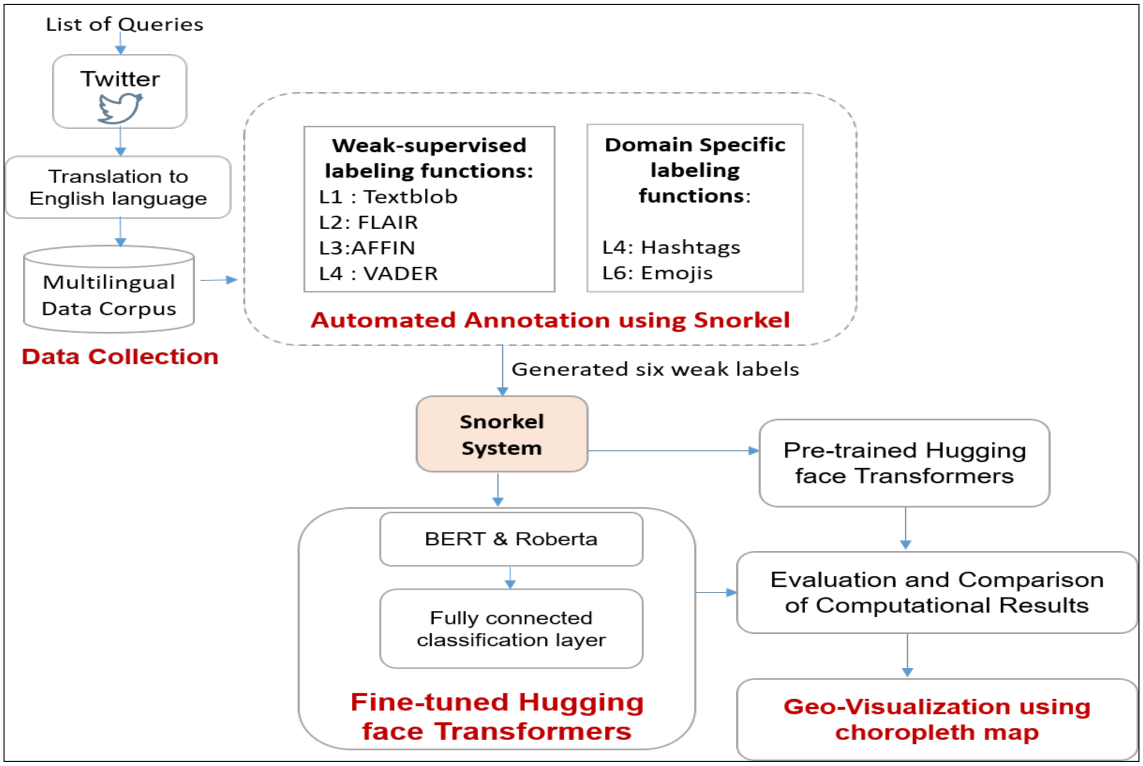
List of Queries
Twitter
Translation to English language
Multilingual Data Corpus
Data Collection
Weak-supervised labeling functions:
L1 : Textblob
L2: FLAIR
L3:AFFIN
L4 : VADER
Domain Specific labeling functions:
L4: Hashtags
L6: Emojis
Automated Annotation using Snorkel
Generated six weak labels
Snorkel System
Pre-trained Hugging face Transformers
BERT & Roberta
Fully connected classification layer
Fine-tuned Hugging face Transformers
Evaluation and Comparison of Computational Results
Geo-Visualization using choropleth map

Fig. 1. Proposed methodology for sentiment analysis.

Corpus per Region	Total tweets
India	60000 in Hindi and English
Pakistan	40000 in Urdu and English
J& K state	10000 in English, Hindi, Urdu

TABLE II

TOTAL TWEETS IN THE DEVELOPED CORPUS FROM THREE REGIONS

Automated data annotation is required to overcome this limitation. In this paper, we have proposed an automated approach for data annotation using weak supervised techniques [17]. Snorkel is a system that presents a method of creating training datasets without manual labeling [18]. The first element of a Snorkel pipeline covers labeling functions, intended to be weak heuristic functions to predict the label for the given data. We have defined five labeling functions, including three weak classifiers and two domain-specific classifiers, by creating lexicons for positive and negative hashtags and emojis from the dataset. Labeling functions for Snorkel include three weak classifiers and two domain-specific lexicon-based classifiers, which are described as follows:

1) Three weak classifiers :
   - Textblob: Textblob is a Python library to process text data and perform different tasks in natural language processing (NLP) [19]. It performs sentiment analysis using a lexicon-based approach to depict the polarity and subjectivity of a text by applying statistical rules based on certain words present in the given text.
   - VADER: VADER (Valence Aware Dictionary for sentiment Reasoning) is a simple rule-based approach to sentiment analysis [20]. It leverages both quantitative and qualitative methods to create and validate lexical features with associate sentiment scores. Then, the generated lexical features encompass grammatical and syntactical conventions in order to convey the intensity of sentiments.
   - Flair: Flair is the state-of-the-art NLP technique [21] with several functionalities, such as pre-trained sentiment analysis models, text embeddings, and name entity recognition (NER), etc. It is trained on IMDB data but can be custom trained by generating embeddings from the dataset under investigation. Therefore, Flair comprehends the context and provides better performance.
2) Two domain specific lexicon-based classifiers:
   - Leveraging hashtags: Hashtags used in a sentence have been used to predict the sentiment of the sentence [22]. To this aim, two lexicons with positive and negative words present in the hashtags have been constructed and leveraged to generate a weak label for each tweet in the dataset.
   - Leveraging Emojis: Similarly, two lexicons for emojis used in the dataset have been built representing positive and negative emojis. The description of emojis has been used to depict the weak label for the sentiment of tweets.

After applying these defined labeling functions, we obtain six weak labels, which in turn, are fed into the snorkel model to create one label for each data record. The snorkel system outputs two labels based upon sentiment polarity, i.e., positive if $polarity > 0$ or negative if $polarity < 0$. The overlapping conflicts are negligible and, therefore, assumed as no conflict is observed in this process. To end, this labeled data is used as a training dataset in further steps of this work. Furthermore, the ground truth of these annotations has been done by a few human experts in this domain, who randomly validated the generated automated annotations.

D. Hugging Face BERT Transformer Modeling

Two transformer models from the hugging face library have been employed in our pilot study, namely Bert-base-uncased and Twitter-roberta-base-sentiment.

BERT-base-uncased: BERT (Bidirectional Encoder Representations) has transformer-based architecture and is pre-trained on a large corpus in English language [23]. The training approach used is self-supervision with unlabeled data. It processes information from both sides (left and right) of a token's context during the training phase. As a result, it can be fine-tuned to a wide range of NLP tasks such as question-answering, text classification, etc.

Twitter-roberta-base-sentiment: It is a Robustly Optimized BERT, Pre-trained on Google's BERT model [24]. It is trained on 58M tweets and fine-tuned for applications like sentiment analysis.

We have selected these two baselines because RoBerta is extensively trained on Twitter data, whereas BERT-base-uncased has been employed to compare the results of Roberta with the base model. The models have been experimented using both translated data, i.e., single language as well as multi-linguistic data. Essentially, Roberta is 16 times more computationally expensive than Bert; it took more time and resources. Roberta Tokeniser has been used in the process.

E. Experiments & Results

Experimental training has been conducted on a Tesla P100-PCIE with GPU memory of 16GB. Adam optimizer has been utilized with a learning rate of 1e-05 and decay rate 1e-07 for 3-4 epochs and varying batch sizes of 8 and 16. Bert model has been pre-trained on a large corpus in the English language using self-supervised learning, whereas the Roberta model has been pre-trained on 58M tweets and fine-tuned for sentiment analysis. Table III shows that models used in this work have given substantial results for positive sentiments, while negative sentiment classification has not performed that well. Also, models have not performed well on non-translated data. Therefore, in our pilot study, models have shown better results when applied to single-language data. Overall, for the translated dataset, there is less difference in the performance of Bert and Roberta. However, BERT takes less memory and time

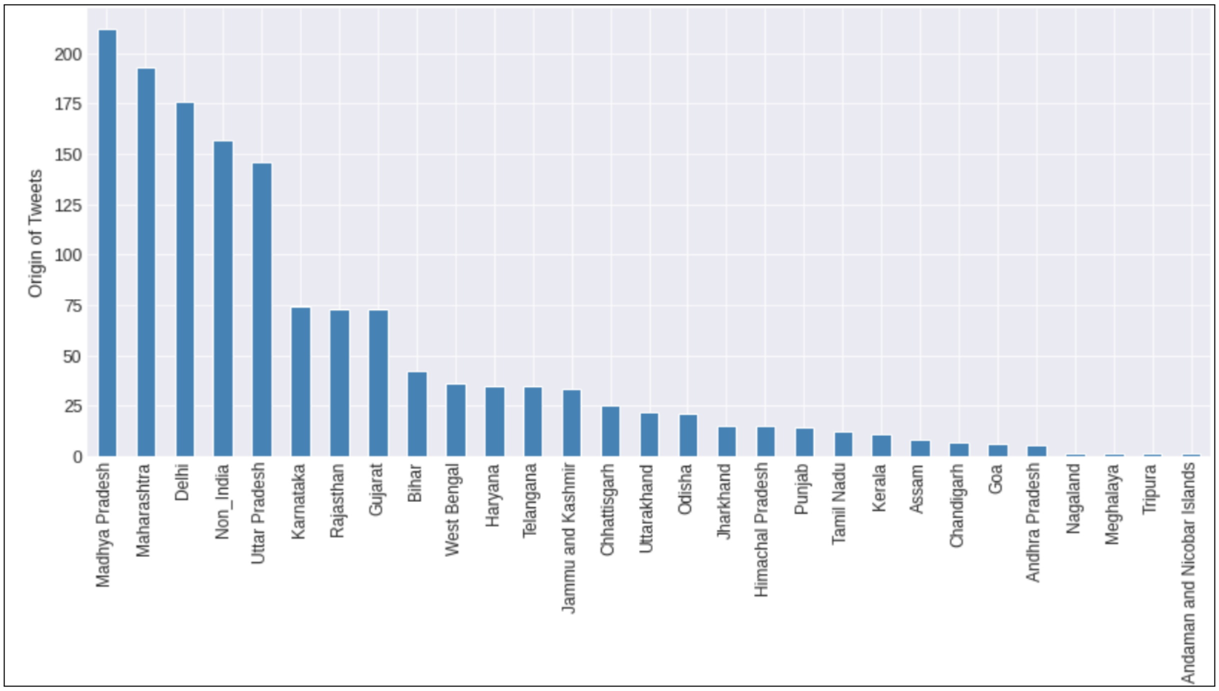

Fig. 2. State-wise ranking based on the collected Twitter data.

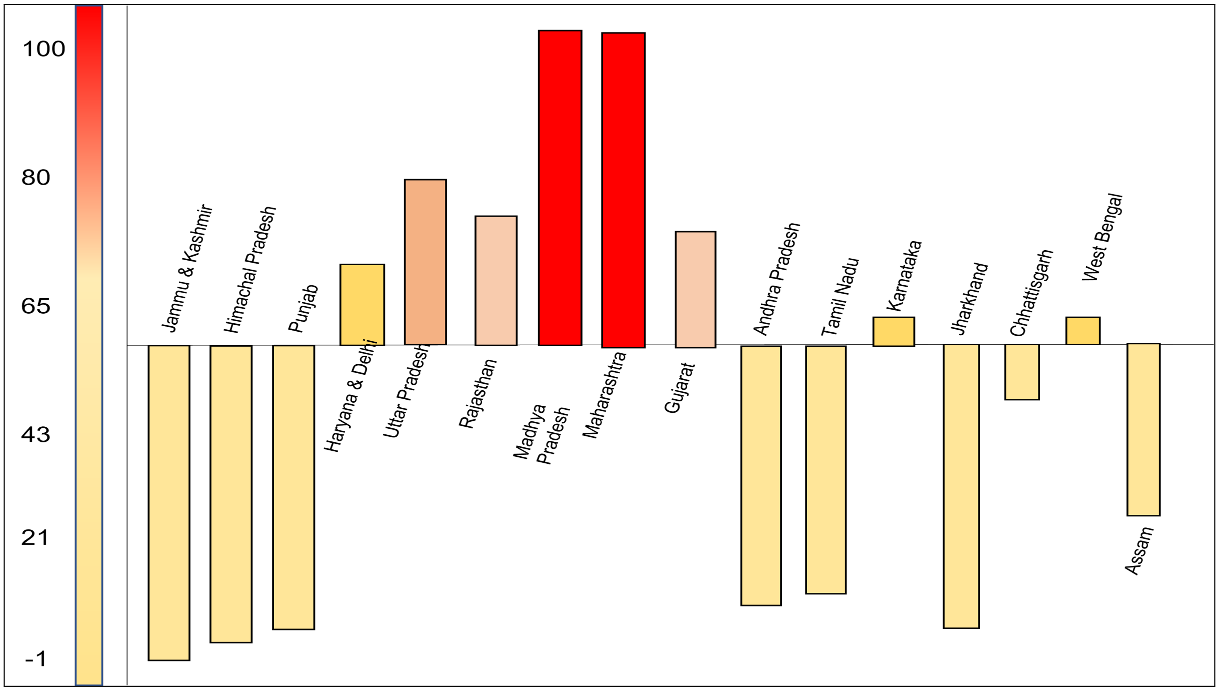

Fig. 3. Visualization of geospatial sentiments

Batch Size	Label	Preci.	Recall	F1
8 (BERT)	-1	0.77	0.81	0.79
Translated	1	0.92	0.90	0.91
16 (Bert)	-1	0.79	0.77	0.78
Translated	1	0.91	0.92	0.91
8 (RoBerta)	-1	0.80	0.77	0.79
Translated	1	0.91	0.92	0.92
16 (RoBerta)	-1	0.79	0.76	0.78
Translated	1	0.91	0.92	0.91
8 (Roberta)	-1	0.29	0.36	0.32
Non-Translated	1	0.72	0.66	0.69
16 RoBerta	-1	0.29	0.32	0.31
Non-Translated	1	0.72	0.69	0.70

TABLE III
COMPARISON OF EMPIRICAL RESULTS OF BERT AND ROBERTA MODELS.

to train; thus, it is considered a more efficient and preferable model.

F. Visualization & Observations

Visualizing the analytical data enriches the insights provided by the data. In order to visualize the obtained insights from data, we have leveraged different libraries from the python framework. Firstly, we have used *geopy* library in the python framework to extract the respective states from the given geo-coordinates (i.e., longitude, latitude, and radius). Secondly, libraries and packages such as *shapely, folium, and plotly* have been utilized to represent sentiments of each state with a warm color schema. As shown in Figure 2, the maximum number of tweets were extracted from Madhya Pradesh, followed by Maharashtra and Delhi. The study aims to visualize the sentiments of users from different states of India on the topic of the Kashmir conflict. Therefore, the non_Indian tweets, i.e., tweets originating from outside of India, have not been considered for the final Visualisations. Furthermore, we have attempted to visualize the results obtained from our proposed model using the *choropleth maps* of Indian locations as shown in Figure 3. A choropleth map is a kind of statistical thematic map where the intensity of color corresponds to the count of a spatial enumeration unit. Since the boundary of Kashmir is a politically contentious issue, we have shown it using a bar chart in this paper. The state-level sentiment scores can be seen with the help of this visualization, where the emotions vary from negative to positive, with colors from light to dark shades. This geographical representation is a useful tool to discover the sentiments of users in a geospatial manner. The bar chart shows the sentiment scale for major populated states of the country.

IV. BACKGROUND OF CASE STUDY

Kashmir territorial conflict is an important unresolved conflict between India and Pakistan since the two countries acquired independence in 1947 [25]. Since then, it has resulted in wars and regular clashes between the involved nations. The state is still under dispute, leading to mob violence and civil unrest and is being used as info warfare by foreign advisories. In the past few years, several events happened due to military and government actions that have affected this issue. The study covers such events to comprehend the sentiments of people from different regions:

- Article 370 by the Indian constitution gives special privileges to Jammu and Kashmir state to exert autonomy and formulate its own laws. However,
- Balakot Airstrike: On 26 February 2019, Indian warplanes conducted an airstrike on the terrorist camps in Balakot, Pakistan. Indian authentic sources claimed 200–350 JeM militants killing and the distortion of terrorist camp.
- Galwan Valley: Beginning from 5 May 2020 onwards, tensions across the Indo-China border escalated into a series of skirmishes between Indian and Chinese forces. It led to losses on both sides.
- Pulwama attack: On 14 February 2019, the convoy of CRPF vehicles was attacked by a suicide bomber in the Pulwama district of J&K, which led to a loss of 40 CRPF personnel.

V. DISCUSSION & LIMITATIONS

We have performed fine-tuning by training the entire architecture, which takes a large amount of time and memory. Hence, we have performed experiments using mini-batch sizes of 8 and 16 sizes due to the resource constraint. The Snorkel framework does not support multilingual data, and the data needs to be translated before being fed into the transformer model. Therefore, in our pilot study, we have not leveraged the multilingual feature of transformer models. Furthermore, we have collected data from Twitter using four popular events related to the Kashmir conflict. However, this data does not claim to demonstrate the whole population of users.

VI. CONCLUSION

The paper presents comprehensive multilingual data covering an important case study, i.e., Kashmir territorial conflict. We highlight an important challenge of creating labeled trained data by proposing an automated approach using weak supervision. First, we translated the multilingual data into a single language, i.e., English. Secondly, we have defined six labeling functions to generate six weak labels, then fed them into the Snorkel system to obtain one label for training the model. The generated annotations have been validated by human domain experts in order to provide ground truth. It has been observed that the annotations obtained by the automated method were not completely correct; however, this automated annotation proposed method is useful to train machines with a huge amount of data. Once the labels had been obtained, we applied two baseline models, BERT-base-uncased and RoBerta, on translated as well as initial data in three languages. In our pilot study, both the models have performed well, whereas considering the resource constraints involved in RoBerta, the BERT base model is preferred. In this paper, the sentiments have been visualized geographically in order to predict regions where the probability of an unrest situation is high if sentiments circulated from the region are negative.

References

[1] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, no. 2, pp. 617–663, 2019.

[2] S. Sharma and A. Jain, "Role of sentiment analysis in social media security and analytics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1366, 2020.

[3] S. T. Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, p. 100157, 2022.

[4] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, "Sentiment analysis: A literature review," in *2012 International Symposium on Management of Technology (ISMOT)*. IEEE, 2012, pp. 572–576.

[5] A. Sadia, F. Khan, and F. Bashir, "An overview of lexicon-based approach for sentiment analysis," in *2018 3rd International Electrical Engineering Conference (IEEC 2018)*, 2018, pp. 1–6.

[6] B. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," in *2017 International conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2017, pp. 216–221.

[7] K. Rakshitha, H. M. Ramalingam, M. Pavithra, H. D. Advi, and M. Hegde, "Sentimental analysis of indian regional languages on social media," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 414–420, 2021.

[8] S. S. Jacob and R. Vijayakumar, "Sentimental analysis over twitter data using clustering based machine learning algorithm," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2021.

[9] S. Rastogi and D. Bansal, "Visualization of twitter sentiments on kashmir territorial conflict," *Cybernetics and Systems*, vol. 52, no. 8, pp. 642–669, 2021.

[10] H. Zhang, W. Gan, and B. Jiang, "Machine learning and lexicon based methods for sentiment classification: A survey," in *2014 11th web information system and application conference*. IEEE, 2014, pp. 262–265.

[11] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," *Applied Intelligence*, vol. 48, no. 5, pp. 1218–1232, 2018.

[12] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: from lexicons to transformers," *IEEE access*, vol. 8, pp. 131 662–131 682, 2020.

[13] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," *arXiv preprint arXiv:1903.09588*, 2019.

[14] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta," *Applied Intelligence*, vol. 51, no. 6, pp. 3522–3533, 2021.

[15] J. Tao and X. Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," *Journal of Big Data*, vol. 7, no. 1, pp. 1–26, 2020.

[16] S. Pradha, M. N. Halgamuge, and N. T. Q. Vinh, "Effective text data preprocessing technique for sentiment analysis in social media data," in *2019 11th international conference on knowledge and systems engineering (KSE)*. IEEE, 2019, pp. 1–8.

[17] N. Jain, "Customer sentiment analysis using weak supervision for customer-agent chat," *arXiv preprint arXiv:2111.14282*, 2021.

[18] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," in *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, vol. 11, no. 3. NIH Public Access, 2017, p. 269.

[19] S. Loria *et al.*, "textblob documentation," *Release 0.15*, vol. 2, p. 269, 2018.

[20] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.

[21] S. M. Yimam, H. M. Alemayehu, A. Ayele, and C. Biemann, "Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1048–1060.

[22] K. W. Lim and W. Buntine, "Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon," in *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 2014, pp. 1319–1328.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[25] D. A. Majid and D. M. Hussin, "Kashmir: A conflict between india and pakistan," *South Asian Studies*, vol. 31, no. 1, 2020.