

Stance Detection in Twitter Conversations Using Reply Support Classification

Parul Khandelwal¹^a, Preety Singh¹^b, Rajbir Kaur¹^c and Roshni Chakraborty²^d

¹The LNM Institute of Information Technology, Jaipur, India

²ABV IITM, Gwalior, India

{parul.khandelwal.y20pg, preety, rajbirkaur}@lnmiit.ac.in, rcrimi08@gmail.com

Keywords: Stance, Twitter, Reply Support Label, LSTM, BERT.

Abstract: During crisis, social media platforms like Twitter play a crucial role in disseminating information and offering emotional support. Understanding the conversations among people is essential for evaluating the overall impact of the crisis on the public. In this paper, we focus on classifying the replies to tweets during the “Fall of Kabul” event into three classes: *supporting*, *unbiased*, and *opposing*. To achieve this goal, we proposed two frameworks. We used LSTM layers for sentence/word-level feature extraction for classification. We also employed a BERT-based approach where the text of both the tweet and the reply is concatenated. Our evaluation on real-world crisis data showed that the BERT-based architecture outperformed the LSTM models. It produced an F1-score of 0.726 for the *opposing* class, 0.738 for the *unbiased* class, and 0.729 for the *supportive* class. These results highlight the robustness of contextualized embeddings in accurately identifying the stance of replies within Twitter conversations through tweet-reply pairs.


1 INTRODUCTION


Social media platforms, especially Twitter, can be a helpful tool for disseminating information or to share emotions in case of any crisis. Tweets often provide a straight line of communication between the affected population and concerned authorities (Bukar et al., 2022). Considerable research has been done on stance classification in events. Stance, has generally been referred to as public opinion, majorly in reference to government policies, social movements, pandemics, etc. (ALDayel and Magdy, 2021). In most research, attention has been focused on determining whether a tweet will support, oppose, or maintain neutrality toward a certain topic or entity (Küçük and Can, 2020). Stance detection has been done for political discourses (Lai et al., 2019), health misinformation (Ng and Carley, 2022), rumors identification (Zheng et al., 2022; Haouari and Elsayed, 2024), crisis scenarios (Zeng et al., 2016), or other social issues.


However, less attention has been paid to the stance expressed in individual conversation threads. Classi-


fying tweet-reply interactions presents a more complex and nuanced understanding of online conversations. A reply can either be building on, challenging, or providing a little context to the original tweet (Zubiaga et al., 2016). This can be indispensable for judgment regarding public opinion. Tweets provide the initial viewpoint or information while analyzing replies can capture the full spectrum of the discourse. This gap in research can be filled through stance classification in conversation threads to supplement the dialogue dynamics especially during crisis management. The categorization of the stance of replies can reveal if important information is being amplified or contested. This will enable the authorities to evaluate how people are perceiving updates, advisories, or policies. With this additional knowledge, emergency responders can modify their communication strategies. This will also help understand the sentimental conditions of the affected population and gauge the level of solidarity and aid being shared (Hardalov et al., 2021).

Reactions to tweets can take forms of support, opposition, or additional context and can thereby significantly affect the dynamics of the virtual discussions. The classification of responses can shed more light on the actual impact of the seed tweets during crisis situations. This study focuses on automatically

^a <https://orcid.org/0000-0003-4963-156X>

^b <https://orcid.org/0000-0002-6885-1376>

^c <https://orcid.org/0000-0003-1433-1065>

^d <https://orcid.org/0000-0003-4476-403X>

classifying tweet-reply pairs to more correctly comprehend public engagement and sentiment in conversation threads. Our contributions are summarized below:

- **Conversation Stance Classification:** We propose classifying tweet-reply conversations, focusing on both their texts, to assess the support flow of online social media conversations during a crisis.
- **Experimentation on Multiple Frameworks:** We present two distinct classification frameworks, utilizing LSTM-based models (Hochreiter, 1997) and a BERT-based architecture (Devlin, 2018). Each framework handles tweet-reply embeddings and sequence processing distinctly.
- **Explainable AI:** We apply the concept of explainable AI to interpret the predictions of our stance classification model.

The paper is organized as follows: Section 2 discusses existing research on tweet classification and reply analysis in crisis. Section 3 outlines the methodology, including the dataset and description of the frameworks. Section 4 provides details on the experimental setup. Section 5 discusses the results and analysis, comparing the performance of each framework. Section 6 concludes the paper.

2 RELATED WORK

Stance classification has evolved through various methodologies to analyze public sentiment and discourse on social media platforms. (Mohammad et al., 2016) introduced stance detection by identifying support, opposition, or neutrality toward specific targets. (Zubiaga et al., 2016) extended this work to classify stance in conversational threads using tree structures, categorizing replies as supporting, denying, querying, or commenting. (Mutlu et al., 2020) explored public perceptions using TF-IDF and CNN-based methods, while (Villa-Cox et al., 2020) focused on replies and quotes in controversial Twitter conversations. Similarly, (Hamad et al., 2022) developed the StEduCov dataset to classify online education stances during the COVID-19 pandemic.

Recent advancements include multi-task frameworks like (Abulaish et al., 2023), which employed Graph Attention Networks (GAT) for jointly predicting stance and rumor veracity, and (Bai et al., 2023), who proposed the Multi-Task Attention Tree Neural Network (MATNN) for hierarchical classification of stance and veracity. (Zhang et al., 2024) introduced DoubleH, leveraging user-tweet bipartite graphs for

stance detection related to the 2020 US presidential election. While these studies showcase diverse methodologies, they primarily focus on individual tweets or structured datasets, overlooking the stance of replies toward parent tweets in tweet-reply conversations.

3 METHODOLOGY

In this section, we present our methodology to categorize the support of replies to the parent tweet in a conversation thread. A high level of *support* indicates effective communication of the messages. *Opposing* or *unbiased* replies might suggest misinformation or confusion among the affected population. Using such insights can empower crisis managers and policymakers to make appropriate, real-time adjustments to messaging strategies.

Problem Statement: Consider a dataset \mathcal{D} , consisting of n tweet-reply pairs $\{(T_i, R_i)\}_{i=1}^n$, where T_i represents the text of the i -th tweet and R_i represents the corresponding reply text. Each tweet T_i has a **tweet label** assigned to it. Tweets providing factual data related to the crisis are labelled as *information*. Those reflecting personal reactions and appeals for support are labelled as *emotion*. Tweets stating personal opinions are termed as *neutral*. Classifying tweets offers insights into the nature of the discourse being carried on during the crisis. A tweet can have multiple replies, forming many tweet-reply pairs. The goal is to classify the **reply support label** into one of the following three categories:

- **Supporting:** The reply demonstrates alignment with or endorses the parent tweet.
- **Unbiased:** Reply neither supports nor opposes the parent tweet.
- **Opposing:** The reply expresses opposition to the parent tweet.

Considering the following example:

Tweet Example (tweet label: *information*): *The US plans to completely pull all personnel from its embassy in Kabul over the next 72 hours as Taliban forces close in on Afghanistan's capital* <https://cnn.it/3CRwPXG>

- **Supporting Reply:** *"Bring them home @POTUS @JoeBiden !!! Thank you !! You are doing good."*
- **Unbiased Reply:** *#BidensSaigon*
- **Opposing Reply:** *Sickens me beyond comprehension. Biden is not only a liar and self serving crook, heâ€™s a coward. And where is he and where has he been?? He had all the answers the*

last 4 years and now he disappears and is literally putting the US in harms way.

There were 123 replies on this tweet at the time of data collection. Of these, 48 were *opposing*, 17 were *supporting* and 58 were *unbiased*. Such an analysis through different conversation threads can provide an indication on the inclination of the people’s responses to the developments during the crisis.

Formally, our problem becomes a mapping of the function $f : \mathcal{D} \rightarrow y_i \in \{-1, 0, 1\}$ for the classification of stances in tweet-reply pairs:

$$f(T_i, R_i) = \begin{cases} 1, & \text{if } R_i \text{ supports } T_i \\ 0, & \text{if } R_i \text{ is neutral towards } T_i \\ -1, & \text{if } R_i \text{ opposes } T_i \end{cases}$$

The objective is to minimize the classification error by using the sparse categorical cross-entropy loss function, computed as:

$$\text{Loss}(f) = -\frac{1}{n} \sum_{i=1}^n y_i \cdot \log(p_{\text{true}}(f(T_i, R_i)))$$

where, y_i is the true *reply support label* for the tweet-reply pair and p_{true} represents the predicted probability for the true class y_i .

3.1 Data Collection, Annotation and Preprocessing

For the dataset, we focused on a significant geopolitical event of the *Fall of Kabul* in the year 2021. The tweets in this dataset are labelled with **tweet labels** as *Information (I)*, *Emotion (E)*, *Neutral (N)*, and *Irrelevant (X)*. We curated a selection of labelled tweets from this dataset with at least one reply, ensuring engagement within the conversation. We did not consider the *Irrelevant* tweets. We extracted only the English tweets along with their corresponding English replies. This resulted in the final dataset comprising of 495 tweets along with their replies. As a tweet can have multiple replies, our final dataset is comprised of 2864 tweet-reply pairs. The resulting dataset has 200 *emotion* class tweets with 760 replies, 163 *information* class tweets with 1180 replies, and 132 *neutral* class tweets with 924 replies.

For each tweet-reply pair, annotations were done by providing a **reply support label** as *supporting*, *unbiased*, or *opposing*. The annotation process was meticulously carried out by two independent human annotators. To maintain inter-annotator reliability, the annotated datasets were cross-checked by the annotators for consistency and discrepancies were resolved through mutual discussion. The labelling resulted in

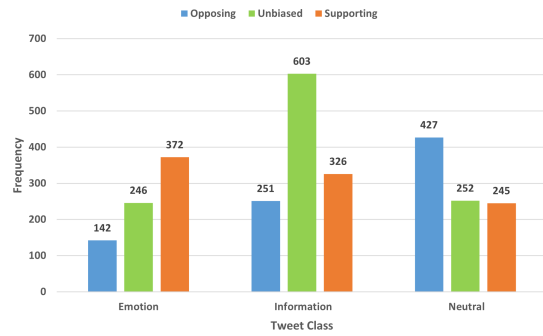


Figure 1: Distribution of *supporting*, *unbiased* and *opposing* replies across the *informative*, *emotional* and *neutral* tweets.

943 *supporting* instances, 1101 *unbiased* instances, and 820 *opposing* instances as shown in Figure 1.

To prepare the data for analysis, we focused exclusively on the textual content of the tweets and replies. All emojis were converted into their corresponding textual meanings and hashtags were split into their component words. The dataset will be made publicly available after publication of the paper.

3.2 Framework for Classification of Reply Support Labels

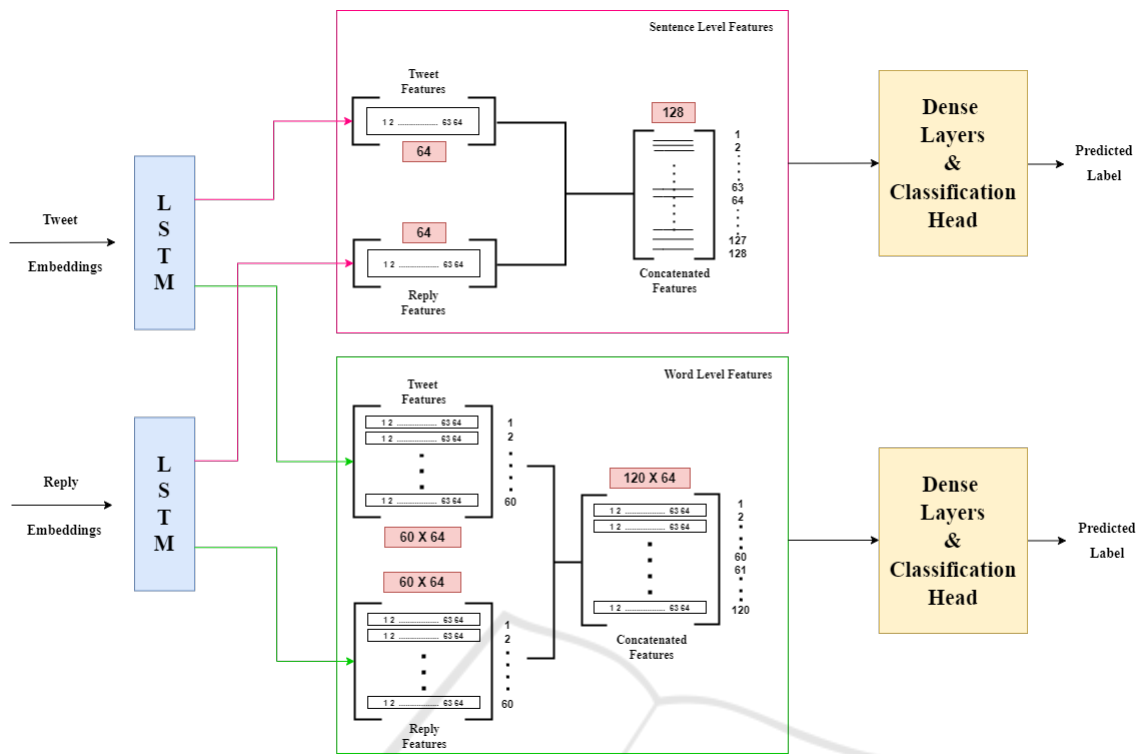
For the classification of the **reply support label**, we explored two frameworks, each designed to optimize the accuracy of reply support classification.

3.2.1 Sentence/Word-Level LSTM Features for Classification

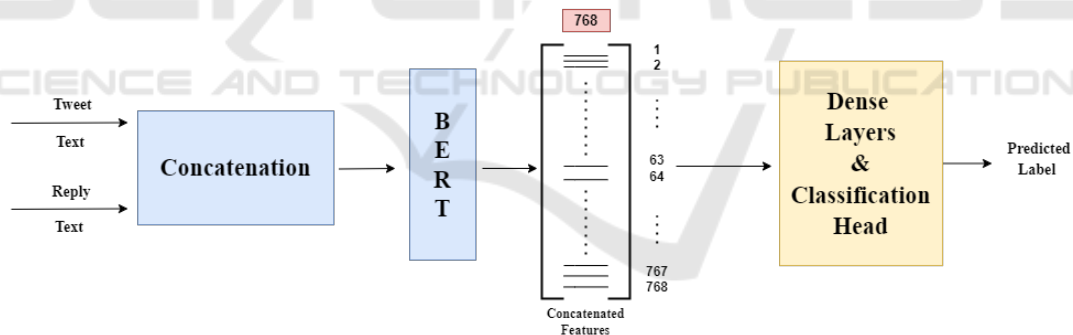
In this framework, we utilized LSTM modules for extraction of features. We employed two options: sentence-level and word-level features. For both approaches, we generated embeddings for the tweet and the reply using the GloVe technique which were passed through two identical LSTM models as shown in Figure 2a. Each model processed the entire text sequence as a whole.

For sentence-level approach, the LSTM models produced two distinct feature vectors, z_t and z_r , corresponding to the tweet and the reply. The vectors were concatenated to form a combined vector, z_{tr} , and sent as input to a network of dense layers for the final classification. This concatenation step allows the final classification model to consider both the tweet and the reply simultaneously.

For word-wise feature extraction approach, the embeddings were passed through separate LSTM models, which generated feature matrices rather than single vectors (refer Figure 2a). The resulting feature



(a) LSTM-based classification.



(b) BERT-based classification.

Figure 2: Frameworks for classification of **reply support labels**. (a) Sentence-level and word-level feature extraction from GLoVe embeddings of tweet and reply using LSTM. (b) Feature extraction from tweet and reply texts using BERT architecture.

matrices, Z_t and Z_r for tweet and reply respectively. Z_t and Z_r were concatenated to form a combined matrix Z_{tr} . Z_{tr} was flattened and sent as input through dense layers to classify the stance of the reply.

3.2.2 Attention-Based BERT for Classification

In this framework, we leverage the state-of-the-art BERT (Devlin, 2018) model to perform both em-

bedding generation and classification. This approach concatenated the tweet and reply text into a single sequence for generating embeddings and features in a unified context. The BERT model outputs a single embedding of vector size 768 which was then directly fed into a classification layer to predict the **reply support label**.

Table 1: Distribution of Classes in Training, Validation, and Test Sets across *emotional* (E), *informative* (I), and *neutral* (N) tweet classes.

	Tweet Label	Reply Support Label		
		-1	0	1
Training	E	116	205	307
	I	206	464	259
	N	334	212	188
Validation	E	13	19	33
	I	25	71	26
	N	44	20	35
Test	E	13	22	32
	I	20	68	41
	N	49	20	22

4 EXPERIMENTAL SETUP

We used 2864 tweet-reply pairs for our experiments as explained in Section 3.1. These pairs are labelled as *supporting* (+1), *unbiased* (0), or *opposing* (-1) for the **reply support label**. The dataset was split into training, validation, and test sets in the ratio of 80 : 10 : 10 using a stratified approach. Table 1 shows the distribution of the tweet-reply pairs according to their **reply support labels** and across the **tweet label** classes. We used two frameworks for the classification task as presented in Section 3.2. To evaluate how well our models classify the **reply support label**, we used Precision, Recall, F1-score, and Accuracy.

4.1 Setup for LSTM-Based Classification

Two LSTM models, each with 64 units and a dropout rate of 0.3 were used to process the tweet and the reply in both sentence-level and word-level feature extraction. For sentence-level features, the concatenated feature vector output from the LSTMs served as an input to a dense layer of 64 units activated by ReLU, with a batch size of 8. A dropout of 0.2, batch normalization and L2 regularization were applied.

For word-level features, the concatenated output was sent to three dense layers composed of 2048, 512, and 64 units respectively with ReLU activation. Batch normalization and a dropout rate of 0.2 was employed. In both cases, networks were trained using the Adam optimizer with a learning rate of $1e - 4$.

4.2 Setup for BERT-Based Classification

The concatenated text served as input to the BERT model for feature extraction. The output of the BERT model was sent through a dense layer with 256 units with ReLU activation. Batch normalization and L2 regularization and dropout were applied. A dense layer with softmax activation was used for the final classification. The model was trained for 150 epochs using the Adam optimizer with a learning rate of $1e - 5$ using sparse categorical cross-entropy loss function.

5 RESULTS AND DISCUSSION

The results for the classification of **reply support label** for the tweet-reply pairs are presented in Table 2.

5.1 Results for LSTM Models

The sentence-level features from LSTMs obtained an overall accuracy of 64.1%. It achieved the highest precision of 0.697 for the *opposing* class though the recall for this class was low. For the *unbiased* class, the model achieved a strong recall value of 0.736 and an F1-score of 0.661. The model achieved the lowest F1-score of 0.622 for the *opposing* class. The confusion matrix for the framework is shown in Figure 3a. Few *opposing* and *supporting* replies have been identified as *unbiased* replies.

For word-level feature extraction, the overall accuracy was 65.2%. Individual class metrics revealed that the model’s performance on all metrics was fairly similar for all classes. The confusion matrix for the framework is shown in Figure 3b. Few *opposing* replies were falsely labelled as *supporting* or *unbiased*. Some *supporting* replies were identified as *unbiased* replies.

5.2 Results for BERT Model

BERT classification significantly outperformed LSTM models. It achieved an overall accuracy of 73.2%. The *opposing* class showed a high recall of 0.793. There was improvement seen in the *unbiased* class also. The *supportive* class achieved high precision at 0.767 and an F1-score of 0.729, demonstrating relatively well-balanced classification for all classes. The confusion matrix for this framework is given in Figure 3c. It shows the robustness of the

Table 2: Performance Metrics for Sentence-level LSTM Features, Word-level LSTM Features and BERT Features for Classification of Reply Support Label.

Framework	Support Label	Performance Metrics			
		Precision	Recall	F1-score	Accuracy
Sentence-Level LSTM Features	-1	0.697	0.561	0.622	0.641
	0	0.600	0.736	0.661	
	1	0.663	0.600	0.630	
Word-Level LSTM Features	-1	0.675	0.634	0.654	0.652
	0	0.640	0.646	0.643	
	1	0.647	0.674	0.660	
BERT Embeddings	-1	0.670	0.793	0.726	0.732
	0	0.760	0.718	0.738	
	1	0.767	0.695	0.729	

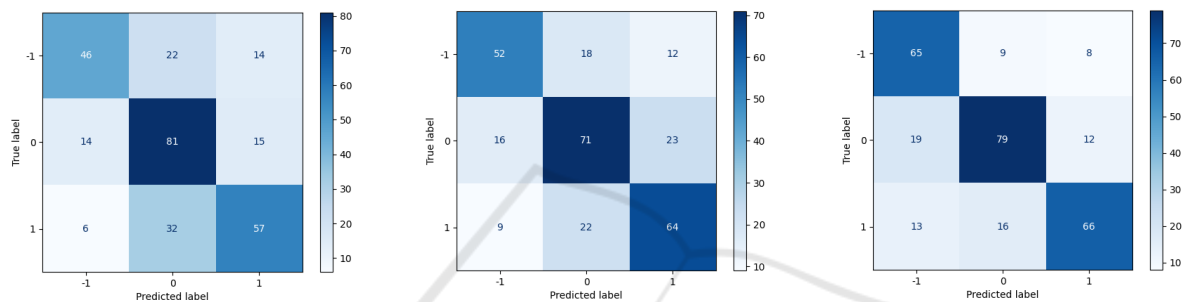
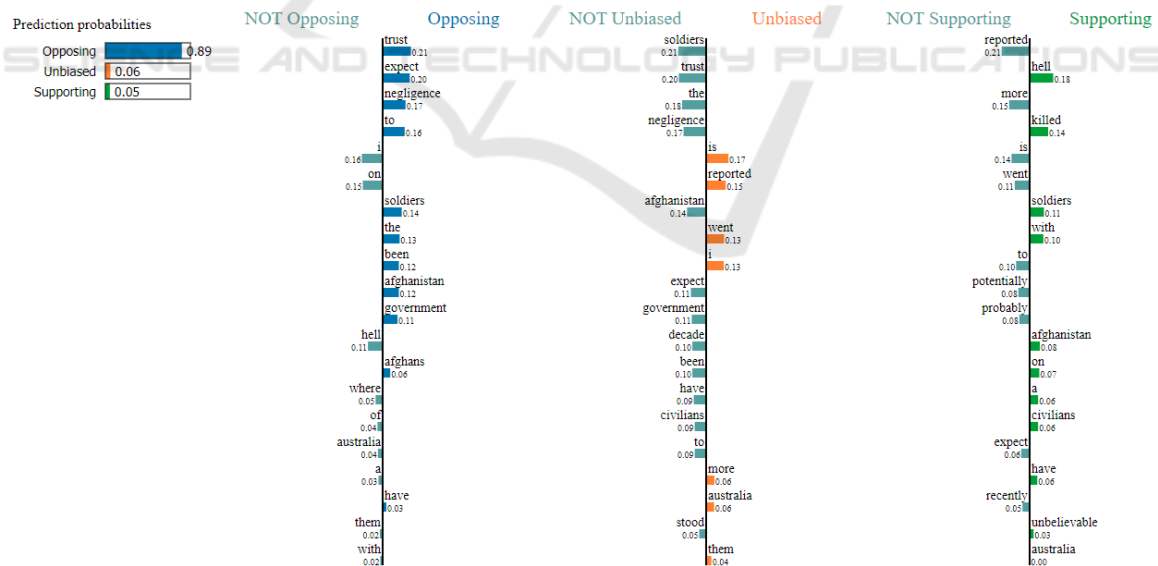


Figure 3: Confusion matrices for classification using: (a) Sentence-level features extracted from LSTM (b) Word-level features extracted from LSTM (c) Embeddings from BERT.



Text with highlighted words

this is unbelievable potentially lethal negligence on the part of the morrison government towards afghans who have stood with australia over a decade i ve been pleading with the government to bring them to australia for months where the hell is dutton.australian elite soldiers recently killed 39 unarmed afghanistan civilians probably more went un reported and you expect them to trust australia

Figure 4: Results of employing LIME on the BERT model for classification of a specific reply to a tweet. The weights assigned to the words contribute to the decision making process of the model in classifying the reply support label.

Table 3: Stance detection on varied events in current research. Our proposed model works on a more granular level by determining stance of replies in conversation threads.

Author	Objective	Stance	Dataset	Models	Results
(Mutlu et al., 2020)	Stance on treatment for Covid-19	neutral, against, favor	COVID-CQ dataset of 14374 tweets	Logistic regression	Accuracy: 0.76
(Hamad et al., 2022)	Stance on online education during Covid-19	agree, disagree, neutral	StEduCov dataset of 16,572 tweets	BERT	Accuracy: 68%
(Zhang et al., 2024)	Stance on US presidential elections	pro-Biden, pro-Trump	Self curated dataset of 1,123,749 tweets	DoubleH	Accuracy: 0.8579 ± 0.04
Proposed Reply support label	Stance of reply in tweet conversations	supporting unbiased opposing	Twitter dataset of event "Fall of Kabul", 2864 tweet-reply pairs	Sentence-Level LSTM Word-Level LSTM BERT	F1-score: 0.637 F1-score: 0.652 F1-score: 0.731

model in correctly classifying *opposing* and *unbiased* responses with moderate misclassification between *unbiased* and *supportive* replies. To summarize, the results indicate that BERT provides the most effective classification of the **reply support labels**, which can play a crucial role in understanding social interactions in crisis contexts.

5.3 Results for Explainable BERT Using LIME

To provide insights into the model’s decision-making process, we focus on enhancing the interpretability of our BERT-based classification model by employing LIME (Local Interpretable Model-agnostic Explanations). Figure 4 demonstrates the results generated by LIME for the following tweet-reply pair.

Tweet: *This is unbelievable, potentially lethal negligence on the part of the Morrison Government towards Afghans who have stood with Australia over a decade. Iâ€™ve been pleading with the government to bring them to Australia for months. Where the hell is Dutton.*<https://t.co/2OrI4HcfTU>
Reply: *https://theguardian.com/australia-news/2020/nov/19/australian-special-forces-involved-in-of-39-afghan-civilians-war-crimes-report-alleges Australian elite soldiers recently killed 39 unarmed Afghanistan civilians. Probably more went unreported. And you expect them to trust Australia?!?*

The visualization demonstrates that the model has classified the reply as *opposing* with a high probability of 0.89. The probabilities for *unbiased* and *supportive* classes are very low, being 0.06 and 0.05 respectively. This suggests that the model is quite con-

fidant about its decision. LIME also highlighted that words such as *{trust, expect, negligence, soldiers}* highly influenced the classification of the reply stance as *opposing*, being assigned weights of 0.21, 0.20, 0.17 and 0.14 respectively. Words such as *{killed, soldiers}* contributed towards the *supporting* class also and *{reported}* was responsible for aiding the *unbiased* class, but the overall influence is minimal.

5.4 Discussion of State-of-Art

Usually stance detection has been done in totality for specific events/topics. Few of these works are shown in Table 3. Our research differs from them as the replies were not considered in those studies. Our framework extends the work by analyzing replies of tweets, deepening the granularity of online discourses. (Mutlu et al., 2020) and (Hamad et al., 2022) reported results for stance on singular tweet classifications like COVID-19 treatment opinions or online education. (Zhang et al., 2024) predicted the binary stance context on political tweets. Our contribution stands out by the fact that it focuses on tweet-reply pairs, a rather unexplored field in stance detection. This conversation analysis can deliver even more subtle insights during crisis where engagement in replies significantly shapes the discourse. Our best framework makes use of a BERT-based framework optimized for the **reply support label** classification achieves an F1-score of 0.73, showcasing reasonable performance in the multi-class setting.

6 CONCLUSIONS

This paper proposed a novel idea of determining the **reply support labels** of tweets during crises to understand the conversation dynamics. By evaluating LSTM and BERT-based models, we demonstrated the reply stance classification in social media conversations, achieving an F1-score of 0.731, recall of 0.735, and accuracy of 0.732. The interaction analysis of tweets and replies is informative in terms of public engagement compared to a sole focus only on the stance of tweets. This knowledge can contribute towards improved information management and offer actionable insight for authorities to tailor their strategies in real-time.

REFERENCES

- Abulaish, M., Saraswat, A., and Fazil, M. (2023). A multi-task learning framework using graph attention network for user stance and rumor veracity prediction. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 149–153.
- ALDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Inf. Process. Manage.*, 58(4).
- Bai, N., Meng, F., Rui, X., and Wang, Z. (2023). A multi-task attention tree neural net for stance classification and rumor veracity detection. *Applied Intelligence*, 53(9):10715–10725.
- Bukar, U. A., Jabar, M. A., Sidi, F., Nor, R. B., Abdullah, S., and Ishak, I. (2022). How social media crisis response and social interaction is helping people recover from covid-19: an empirical investigation. *Journal of computational social science*, pages 1–29.
- Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Hamad, O., Hamdi, A., Hamdi, S., and Shaban, K. (2022). Steducov: an explored and benchmarked dataset on stance detection in tweets towards online education during covid-19 pandemic. *Big Data and Cognitive Computing*, 6(3):88.
- Haouari, F. and Elsayed, T. (2024). Are authorities denying or supporting? detecting stance of authorities towards rumors in twitter. *Social Network Analysis and Mining*, 14(1):34.
- Hardalov, M., Arora, A., Nakov, P., and Augenstein, I. (2021). A survey on stance detection for mis- and disinformation identification. *arXiv preprint arXiv:2103.00242*.
- Hochreiter, S. (1997). Long short-term memory. *Neural Computation MIT-Press*.
- Küçük, D. and Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.
- Lai, M., Tambuscio, M., Patti, V., Ruffo, G., and Rosso, P. (2019). Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124:101738.
- Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.
- Mutlu, E. C., Oghaz, T., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., Ozmen, O., and Garibay, I. (2020). A stance data set on polarized conversations on Twitter about the efficacy of hydroxychloroquine as a treatment for COVID-19. *Data in brief*, 33:106401.
- Ng, L. H. X. and Carley, K. M. (2022). Pro or anti? a social influence model of online stance flipping. *IEEE Transactions on Network Science and Engineering*, 10(1):3–19.
- Villa-Cox, R., Kumar, S., Babcock, M., and Carley, K. M. (2020). Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*.
- Zeng, L., Starbird, K., and Spiro, E. (2016). #unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Proceedings of the international aaai conference on web and social media*, volume 10, pages 747–750.
- Zhang, C., Zhou, Z., Peng, X., and Xu, K. (2024). Double: Twitter user stance detection via bipartite graph neural networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1766–1778.
- Zheng, J., Baheti, A., Naous, T., Xu, W., and Ritter, A. (2022). Stanceosaurus: Classifying stance towards multicultural misinformation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 2132–2151.
- Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., and Lukasik, M. (2016). Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028*.