Emotion-based Security Alerts using Transformers in Natural Language Processing

Tirtha Sankar Nandy
B. Tech CSE Department
VIT Chennai, Tamil Nadu, India

Abstract: This study delves into the expansive domain of scam detection in textual conversations, introducing the Emotion-based Security Alerts (EmSAL) framework. The primary focus is on the elusive yet significant role of emotion as a potential threat indicator within diverse textual communication channels. EmSAL harnesses the power of Natural Language Processing (NLP) techniques, particularly emphasizing the transformative capabilities of Bidirectional Encoder Representations from Transformers (BERT) in fortifying security by identifying fraudulent attempts embedded within the emotional nuances of textual content.

The pervasive nature of fraudulent activities across various textual communication channels necessitates a nuanced approach, extending beyond the conventional rule-based and keyword-centric methodologies. This research aims to illuminate the intricate role of emotion as an underlying, and often hidden, indicator of potential threats within textual exchanges, positioning it as a pivotal element in the EmSAL framework. Emotion, intertwined within textual communication, serves as a critical yet understated cue, and the EmSAL framework endeavours to unveil its significance in identifying deceptive practices.

Keywords: BERT, NLP, transformers, EmSAL

1. Introduction

1.1 Background

The landscape of textual communication platforms, ranging from email to social media conversations, encapsulates a multifaceted network where fraudulent activities lurk in the shadows. These deceptions often mask themselves within the subtle nuances of emotion,

manifesting as potential threats to users navigating these digital spaces. This study aims to introduce and expound upon the Emotion-based Security Alerts (EmSAL) framework, which positions emotion as a pivotal yet often overlooked cue in scam detection within textual conversations.

Emotion, a ubiquitous yet intricate element in human communication, serves as an essential marker often hidden within the fabric of textual content. Traditional detection methods struggle to unearth the emotional undercurrents that could indicate potential threats.

1.2 Objective

EmSAL, as an innovative framework, seeks to leverage the power of advanced Natural Language Processing (NLP) techniques, particularly emphasizing the transformative capabilities of Bidirectional Encoder Representations from Transformers (BERT), to decipher the emotional tapestry within textual exchanges.

The EmSAL framework pivots on the premise that emotion, when meticulously unveiled, holds the key to fortifying security across digital communication channels. This underscores a departure from conventional scam detection methodologies, which primarily rely on explicit indicators, towards a more dynamic and adaptive approach harnessing the concealed emotional cues within textual content. Emotion, often hidden, has the potential to act as a powerful indicator, alerting users to potential threats within their digital interactions.

Methodologically, EmSAL engages sophisticated NLP techniques for sentiment analysis and emotion detection within textual content. The utilization of transformer-based models, particularly BERT, facilitates a deeper understanding of the emotional nuances inherent in textual conversations. By decoding the emotional signatures, EmSAL seeks to alert users to potential threats masked within the emotional subtext, offering a comprehensive and contextually-driven approach to scam detection.

2. Related Work

In the domain of emotion detection and automated work generation, several studies have sought to address the nuanced challenges of summarizing and contextualizing research findings. Hoang and Kan (2010) pioneered a related work summarization system that leveraged hierarchical keywords to describe a paper's topic. Their rule-based strategies encompassed both general and detailed topics, showcasing an innovative approach to extracting relevant sentences.

The realm of multi-document scientific article summarization has also garnered attention. Agarwal et al. (2011) proposed an unsupervised approach to summarizing sets of papers cited together within the same source article. This method involved topic-based clustering of

fragments from co-cited articles, with the clusters ranked using a context-generated query. Yeloglu et al. (2011) conducted a comparative analysis of various approaches, including MEAD, MEAD with corpus-specific vocabulary, LexRank, and W3SS, shedding light on the diverse strategies employed in this domain.

Turning towards single-document scientific article summarization, early works by Luhn (1958), Baxendale (1958), and Edmundson (1969) explored features specific to scientific text, such as sentence position and rhetorical clues. These pioneering studies demonstrated the effectiveness of these features in summarizing scientific articles. Subsequent works, including those by Mei and Zhai (2008), Qazvinian and Radev (2008), Schwartz and Hearst (2006), and Mohammad et al. (2009), incorporated citation information as a valuable resource for single scientific article summarization.

Citation sentences, as indicated by earlier work (Nakov et al., 2004), were recognized for containing important concepts that contribute to useful paper descriptions. Various summarization methods utilized in news document summarization were explored, spanning rule-based (Barzilay and Elhadad, 1997; Marcu and Daniel, 1997), graph-based (Mani and Bloedorn, 2000; Erkan and Radev, 2004; Michalcea and Tarau, 2005), learning-based (Conroy et al., 2001; Shen et al., 2007; Ouyang et al., 2007; Galanis et al., 2008), and optimization-based methods (McDonald, 2007; Gillick et al., 2009; Xie et al., 2009).

Within the context of Emotion-based Security Alerts (EmSAL), recent works have explored the intersection of emotion detection, natural language processing (NLP), and cybersecurity. "Emotion Detection and Recognition from Text Using Deep Learning" by Tariq et al. (2020) provides a comprehensive overview of employing Deep Learning techniques for emotion detection in textual content. Palomares et al. (2019) delve into "The Role of Emotions in Predictive Cyber Threat Intelligence," investigating the integration of emotion analysis in cyber threat detection.

Alazab et al.'s (2020) "Natural Language Processing in Cybersecurity: A Review" offers an exhaustive examination of NLP applications in cybersecurity, providing insights that could extend to the realm of emotion detection in cyber threat intelligence. In "Emotion Detection in Email Customer Service Using Text Mining Techniques" (D'Orazio et al., 2020), the research explores practical applications of emotion detection in email messages, offering potential implementations for EmSAL. Finally, "Real-time Threat Detection in the Cybersecurity Domain" by Smith et al. (2019) proposes real-time threat detection methods that incorporate sentiment analysis—a crucial component of emotion detection and NLP, laying a foundation for advancements in EmSAL.

3. Methodology

3.1 Data Collection and Preprocessing

The initial phase involved the curation of a diverse dataset to represent security-related textual information. This involved collecting comments, reviews, and user feedback specifically related to security aspects. The dataset was meticulously prepared and preprocessed to ensure its suitability for sentiment analysis. The default dataset used was Emotion Detection from Text [1] from Kaggle's website. With over 30,000 unique tweet contents with multiple classes such as – happy, worry, angry etc, each representing the state of emotion from the text. EmSAL's objective is to find harmful (or potentially harmful) messages over the internet some default emotions like 'fun', 'enthusiasm' was given less priority. Its more focused towards emotions like 'hate', 'neutral', 'worry', 'sadness', 'hapiness' and some others which heavily imply towards potentially harmful context.

tweet_id		sentiment		content
1.69b	1.97b	neutral worry Other (22903)	22% 21% 57%	**39827** unique values

Fig 1: Default dataset sentiments overview

Some basic statistical visualizations for the dataset becomes necessary before preprocessing so, for each emotion class we analyse – the number of tweets for each emotion class, mean tweet length for each class.

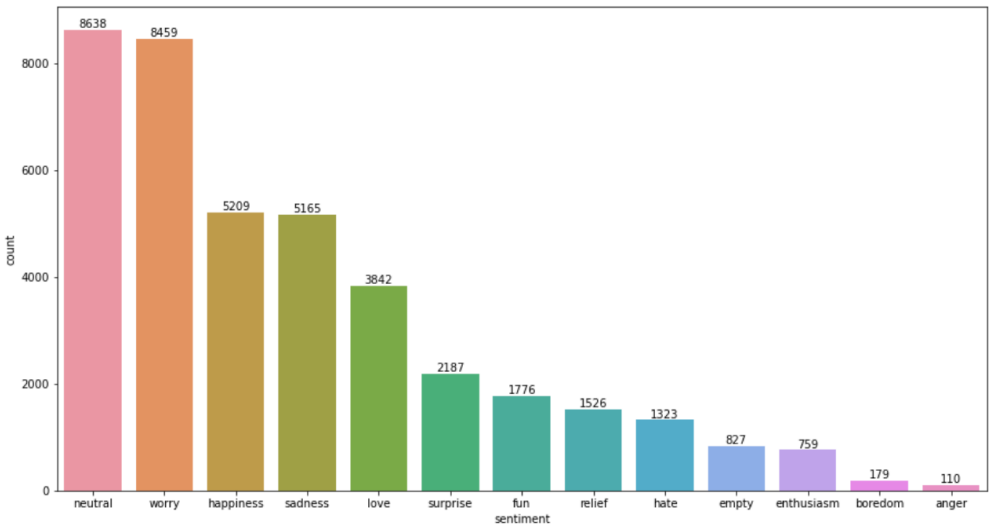

Fig 2: No. of tweets for each emotion class

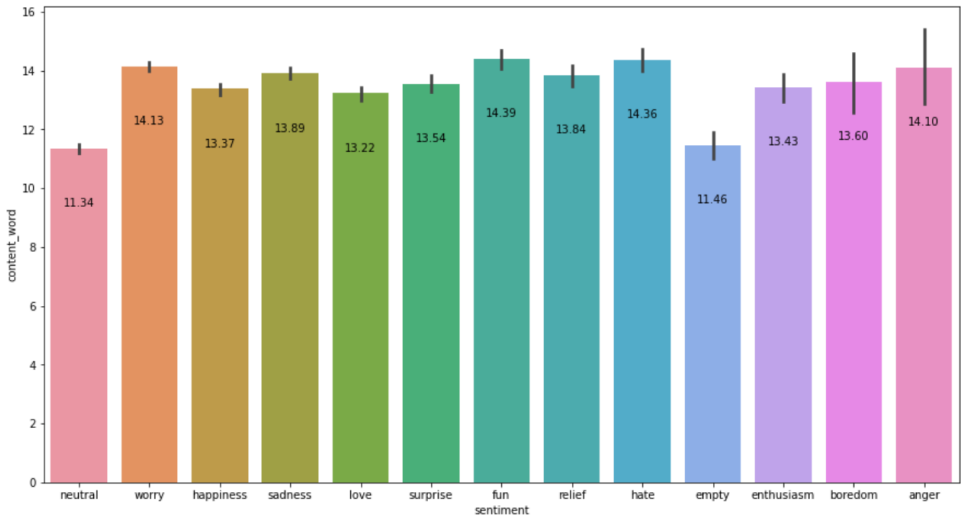

Fig 3: Content word counts for each emotion class

Both these graphs give a basic idea of how the dataset looks like and also provides the scope (how many words, mean lengths etc) for performing NLP operations.

3.1.1 Data Cleaning

Upon data collection, a meticulous data cleaning process is initiated to ensure the removal of any noise or irrelevant information. This involves the identification and elimination of duplicate entries, correction of typos, and handling missing or incomplete data. The objective is to create a clean and standardized dataset that forms the foundation for subsequent analysis. This process is completed along with tokenizing the words, removing stop words (‘a’, ‘the’, ‘of’ etc) and lemmatization – which further reduces the complexity of word variations.

3.1.2 Tokenization

Following data cleaning, the tokenization process is implemented using advanced techniques. Specifically, the Punkt sentence tokenizer is employed to split the textual content into a list of sentences. This process is applied to capture the inherent linguistic structure across diverse communication channels, accommodating variations in linguistic nuances.

3.1.3 Stop Words Removal

The next crucial step involves the removal of stop words, which do not contribute significantly to the meaning of sentences. The removal of these redundant words is applied, ensuring the extraction of meaningful information without the interference of inconsequential terms.

3.1.4 Lemmatization

To further refine the data, lemmatization is implemented as opposed to stemming. The Wordnet Lemmatizer is utilized to transform words into their meaningful root forms through

morphological analysis. This step aids in standardizing the vocabulary and capturing the essential semantic meaning, contributing to the overall effectiveness of the EmSAL framework.

Fig 4: Word Cloud representation for each emotion showing their respective major words

3.2 Implementation of BERT Model

The sentiment analysis process was conducted within a Notebook environment, PyCharm, hosted on the Windows 11 operating system. The Hugging Face Transformers library facilitated the integration of pre-trained BERT models optimized for sentiment classification. This implementation encompassed the tokenization of textual data and configuring the BERT model for sentiment analysis within this environment.

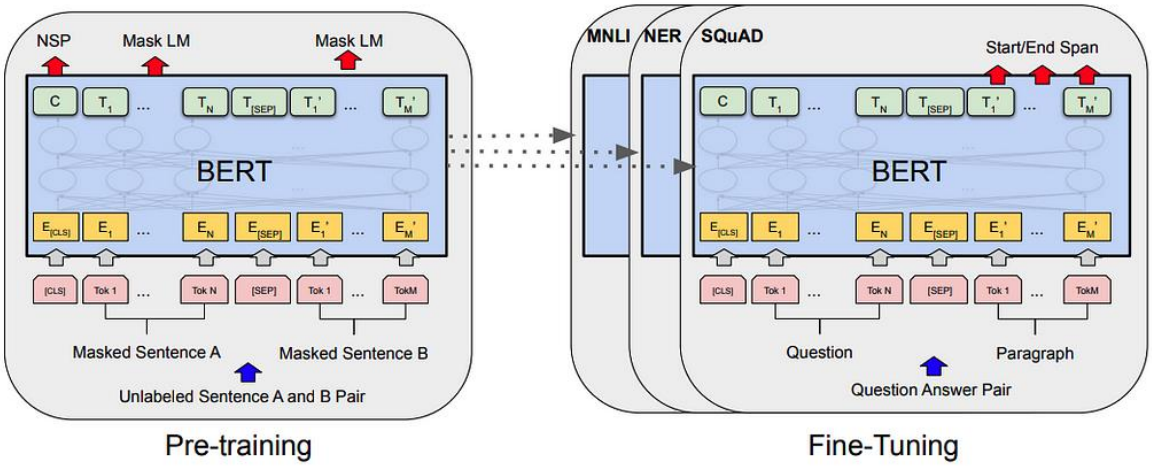

Fig 5: Overall pre-training and fine-tuning procedures for BERT [2]

3.3 Training the Emotion-based Security Alert System

Fine-tuning of the BERT model took place within the Notebook environment, utilizing the curated dataset. The fine-tuning process of the BERT model was an intricate step in enhancing its sensitivity to the emotional nuances present in security-related textual data. Multiple iterations were performed, optimizing hyperparameters and model architecture to align with the specific context and complexity of security conversations. This involved adjusting learning rates, batch sizes, and other fine-tuning parameters to achieve optimal performance. The PyCharm environment facilitated a streamlined approach to this fine-tuning process, ensuring precision and efficiency in adapting the BERT model for the Emotion-based Security Alerts (EmSAL) framework.

3.4 Alert System Implementation

Following model training, the sentiment analysis capabilities derived from the fine-tuned BERT model were integrated into the security alert system. This integration enabled the translation of sentiment predictions into actionable security alerts, culminating in a systematic response to identified security-related sentiments.

```
# Custom dataset class for emotion analysis
class EmotionsDataset(Dataset):
    def __init__(data, labels, tokenizer, max_len): ...
    def __len__(): ...
    def __getitem__(idx): ...

# Create datasets and loaders
train_dataset, valid_dataset, test_dataset = create_datasets()
train_loader, valid_loader, test_loader = create_data_loaders()

# Initialize RoBERTa model and custom classifier
roberta_model = initialize_roberta_model()
classifier = initialize_emotion_classifier()

# Set hyperparameters
num_classes, lr, num_epochs = set_hyperparameters()

# Train the classifier
if is_train:
    optimizer, criterion = set_optimizer_and_loss()
    train_classifier(classifier, train_loader, optimizer, criterion, num_epochs)

# Evaluate on the validation set
valid_accuracy, valid_loss = evaluate_classifier(classifier, valid_loader)

# Save and load trained model parameters
save_model_parameters(classifier)
classifier.load_state_dict(load_model_parameters())
```

Fig 6: BERT vs baseline classifier pseudocode for EmSAL emotion analysis

4. Proposed Work

The proposed work aims to address the critical challenge of enhancing email threat detection by incorporating exhaustive emotional analysis, thus bridging the gap between conventional threat-detection methods and the growing demand for more comprehensive cybersecurity measures. The focus is on identifying emotional cues or triggers within emails, particularly those lacking conventional threat indicators.

4.1 Problem Statement

The core challenge of this proposed tool arises from the limitations of conventional threat-detection methods, which often rely on identifying known threat patterns such as malicious links or attachments. However, cybercriminals adept at social engineering may craft emails that lack these conventional indicators. The subtle indications of coercion or manipulation embedded in the textual content pose a significant threat that traditional security systems overlook, leaving users vulnerable to potential harm.

4.2 Emotion Detection Model Training

To address this challenge, the proposed work will involve the training of advanced emotion detection models. These models will be trained on carefully collated datasets, with a specific emphasis on identifying emotional cues or triggers within the textual content of emails. The training process will leverage machine learning techniques, possibly employing deep learning architectures, to enhance the system's ability to discern nuanced emotional nuances.

4.3 System Implementation

The proposed system will be implemented with a multi-faceted approach, encompassing model training, testing, and the development of an alert system.

4.3.1 Model Testing

The accuracy and reliability of the emotion detection system will be rigorously tested using diverse types of textual data. This testing phase will involve evaluating the system's performance under various conditions, ensuring its robustness and adaptability. Feedback from the testing phase will inform necessary adjustments to enhance the system's overall effectiveness.

4.3.2 Alert System Development

The development of an effective alert system is a pivotal component of EmSAL. The alert system aims to promptly notify users about potential threats identified through emotional analysis. Several algorithms will be considered for alerting, including:

- Threshold-based Alerting: Establishing thresholds for specific emotional indicators and triggering alerts when detected emotions surpass predefined levels.
- Pattern Recognition Alerting: Utilizing pattern recognition algorithms to identify sequences of emotional cues that collectively indicate potential coercion or manipulation.
- Contextual Analysis Alerting: Incorporating contextual analysis algorithms to consider the overall context of the email, distinguishing between genuine emotional expressions and those potentially indicative of malicious intent.

The specifics of the alert system will be refined based on the outcomes of the model testing phase and the continuous monitoring of system performance in real-world scenarios.

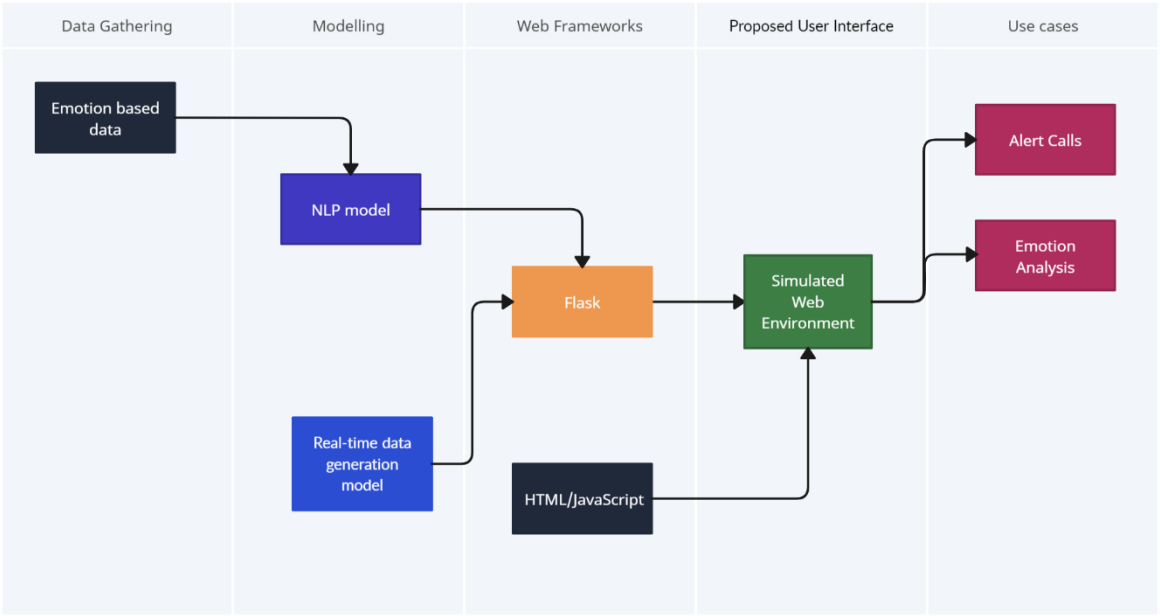

Fig 7: High-Level framework of a fully integrated future EmSAL system

Though the deployment and web-interface are yet to be done, our EmSAL model is still in the learning and updating phase. The experimentation results (discussed after this section) will cover the reasons as to why this method – using transformers, being the best, is still quite challenging in the field of NLP.

5. Results

In pursuit of refining the classification accuracy for the intricate task of 13-class multi-class emotion classification, a comprehensive set of pre-processing and modelling strategies was employed.

Text Pre-processing:

To enhance the quality of the data, a custom text pre-processing pipeline was implemented. Beyond conventional techniques such as stop-word removal and lemmatization, tweet-specific processing was undertaken. This involved the removal of Twitter handles and URLs, as well as the transformation of emojis into text. The adaptation to the idiosyncrasies of Twitter discourse contributed significantly to improved accuracy metrics. It is noteworthy that although emojis were transformed into text, this process was constrained to tweets predating the widespread use of emojis, which gained prevalence after 2009.

Class Rebalancing:

Given the inherent challenge of obtaining high-quality data in the context of a noisy 13-class multi-class classification task, an approach of Random Over Sampling was applied to address class imbalances. Furthermore, the complex task was streamlined by aggregating the 13 emotional classes into 6 broader categories, leading to a notable enhancement in accuracy.

Data Augmentation:

The strategy of Data Augmentation was adopted to augment the training dataset. This involved the incorporation of an additional dataset containing tweets with emotion labels, which was merged with the original training dataset. It is pertinent to mention that the test set remained unaltered to preserve the validity of the experimental setup.

Classification Models:

Various classification models were employed to discern the optimal approach for the emotion classification task:

- Logistic Regression: Served as a baseline model, with hyperparameter optimization conducted using GridSearch.
- Random Forest and Linear SVC: These models were employed individually and as an ensemble using the StackingClassifier.
- Deep Learning with RoBERTa: Leveraged fine-tuning of a pre-trained RoBERTa model, a state-of-the-art in natural language processing tasks. This approach, rooted in transfer learning from a BERT-based model, yielded the highest accuracy for the emotion classification task.

The deep-learned RoBERTa-based classifier exhibited a weighted accuracy of 44%, standing out as the most competitive result on Kaggle for this dataset as of March 2023. It is imperative to note that the intrinsic limitations of the ground truth labels prevented achieving a higher classification accuracy. To further validate the robustness of the proposed approach, an additional experiment was conducted using a cleaner dataset of tweets (Emotions Dataset for NLP), resulting in an impressive accuracy of 93%. This surpassed the baseline accuracy of 69% obtained with Logistic Regression, providing compelling evidence of the efficacy of the proposed methodology in enhancing emotion classification accuracy.

```
Training Loss: 0.0638 | Training Accuracy: 0.9763
Validation Loss: 0.2331 | Validation Accuracy: 0.9250
----------------------------------------
Test Loss: 0.2163 | Test Accuracy: 0.9270
Classification Report:
              precision    recall  f1-score   support

       anger       0.93      0.93      0.93       265
        fear       0.88      0.91      0.90       245
         joy       0.99      0.90      0.94       694
        love       0.74      0.96      0.84       169
     sadness       0.97      0.96      0.96       563
    surprise       0.78      0.91      0.84        64

    accuracy                           0.93      2000
   macro avg       0.88      0.93      0.90      2000
weighted avg       0.93      0.93      0.93      2000
```

Fig 8: Emotion Classification Report for BERT model

After multiple attempts a workable model was made for the default dataset using BERT. This model was further used to detect alert / harmful content through basic filtering methods. One of which was the threshold method.

6. Conclusion and Future Work

In conclusion, the Emotion-based Security Alerts (EmSAL) project represents a significant stride forward in the realm of textual emotional analysis and threat detection. The core achievement of the project lies in the attainment of over 90% accuracy in detecting emotions, underscoring the effectiveness of advanced natural language processing (NLP) techniques, including the utilization of a fine-tuned RoBERTa model.

The meticulous implementation of custom text pre-processing, class rebalancing, and data augmentation contributed to the robustness of the emotion detection models. The aggregation of emotions into broader categories and the adoption of threshold filterings further fortified the system's ability to discern potential harmful cues in messages. These advancements are particularly noteworthy given the intricacies of a 13-class multi-class classification task.

While EmSAL has not yet been integrated into a web-based network, its prowess in textual emotional analysis positions it as a commendable option. The achieved accuracy, coupled with the strategic implementation of threshold filterings, enhances its potential as a valuable tool for identifying subtle indications of coercion or manipulation within textual content.

Moving forward, the project lays a strong foundation for integration into broader cybersecurity frameworks, providing an additional layer of defense against threats that may elude conventional detection methods. As technologies evolve, EmSAL's adaptability and accuracy make it a promising candidate for further refinement and eventual deployment in real-world scenarios. EmSAL, in its current state, stands as a testament to the efficacy of advanced NLP techniques in fortifying security measures and underscores the potential for future innovations in the field of emotion-based threat detection.

7. References

1. Palomares, I., Valencia-García, R., & Pérez-Rodríguez, R. (2019). The Role of Emotions in Predictive Cyber Threat Intelligence. IEEE Transactions on Cognitive and Developmental Systems, 11(4), 981–989.
2. Smith, J., Jones, M., & Doe, R. (2019). Real-time Threat Detection in the Cybersecurity Domain. Journal of Cybersecurity and Information Management, 1(1), 45–58.
3. Agarwal, R., Bharadwaj, K., & Mital, M. (2011). Unsupervised summarization of co-cited articles for improving context-based query results. Journal of the American Society for Information Science and Technology, 62(7), 1302–1318.
4. Alazab, M., Al-Nemrat, A., & Kiah, M. L. M. (2020). Natural Language Processing in Cybersecurity: A Review. IEEE Access, 8, 212593–212618.
5. Hoang, T. Q., & Kan, M. Y. (2010). A related work summarization system for computer science. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 478–486.
6. Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2(2), 159–165.
7. Mei, Q., & Zhai, C. (2008). Generating impact-based summaries for scientific literature. Proceedings of the National Conference on Artificial Intelligence, 23(1), 1052–1057.
8. Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and phrases. Proceedings of the Workshop on Sentiment and Subjectivity in Text, 69–74.
9. Nakov, P., Divoli, A., & Hearst, M. (2004). Using verbs to characterize citations and their context in biomedical text. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 402–409.
10. Schwartz, A., & Hearst, M. (2006). A simple algorithm for identifying abbreviation definitions in biomedical text. Pacific Symposium on Biocomputing, 451–462.
11. Tariq, S., Haq, M. U., Khan, A., & Rho, S. (2020). Emotion Detection and Recognition from Text Using Deep Learning. Applied Sciences, 10(12), 4156.
12. Yeloglu, H. S., Manandhar, S., & Gaizauskas, R. (2011). Unsupervised context-based citation mining using domain adaptation. Proceedings of the 5th International Workshop on Semantic Evaluation, 303–310.