Explainable BERT Embeddings for Veracity Assessment in Criminal Investigations

Thoha Ikhwanul Haq [1)], Chastine Fatichah [2,*)], and Anny Yuniarti [3)]

[1, 2, 3)] Department of Informatics, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia

E-mail: thoha.haq@gmail.com[1)], chastine@its.ac.id[2)], and anny@its.ac.id[3)]

ABSTRACT

The binary classification of truth and lies is often a detriment in criminal investigations as statements are intentionally not entirely true nor entirely false. This ambiguity in the veracity of their claims demands more extensive methods such as explainable models. Explainable models, particularly SHapley Additive exPlanations (SHAP), can help dissect statements and narrow down information for a more thorough investigation. Data from the Miami University Deception Database, comprising of various statements and their veracity, was analyzed for its linguistic features. This research utilizes Bidirectional Encoder Representations from Transformers (BERT) Embeddings to provide contextual understanding of statements and Sentiment Lexicons to provide domain specific knowledge. Results show that the R^2 (coefficient of determination) of the 2-Gram embedding performed the best at 0.39 by being able to capture more context than the 1-Gram embedding while being more general than the 3-Gram and 4-Gram embeddings. Each variant of the BERT Embedding was proven to be much more effective than sgeneral word embedding such as GloVe, Word2Vec and FastText. SHAP values were able to capture key points of interest in a statement by narrowing down pivotal and decision-making points. These results highlight potential indicators of either deceptive or truthful language such as the word ‘something’ and ‘our’. These points of interest can help humans focus on key points of investigation and intervention.

Keywords: BERT, deception, natural language processing, SHAP, word embedding.

1. Introduction

Criminal investigations are often a long and arduous process involving several parties while being contingent on the veracity of every claim made. According to [1], Artificial Intelligence (AI) can be beneficial for law enforcement such as to streamline tasks [2] and to validate the veracity of the claims made by all parties involved quickly [3]. The veracity of a claim can be verified in many ways such as cross-examination, witness testimony [4] or expert analysis. While Al is not yet a primary tool in the due process of legal proceedings [5], it can still be used to flag potential deception and efficiently direct resources into verifying the veracity of these specific statements.

Statements ranging from short to long each have their problems when it comes to determining their validity. The veracity of shorter statements can be difficult to assess due to their lack of context [6], which often leaves room for ambiguity or misinterpretation. In contrast, longer statements can be interwoven with subtle lies and truths to boost the overall veracity of a claim. Many deception detection methods such as [7] focus on the overall truthfulness of a statement, ignoring deceptive nuances in a statement which can often lead to problems with the latter. Therefore, it is useful to narrow down each part of a statements veracity rather than the statement as a whole. Once narrowed down, these individual segments can be used to guide investigative resources in the direction necessary.

One effective method for handling shorter statements in natural language processing tasks is the use of Bidirectional Encoder Representations from Transformers (BERT) [6]. BERT is an AI language model that enhances understanding of language by analyzing words in the context of both their preceding and following words. This

* Corresponding author.
Received: June 24th, 2025. Revised: July 30th, 2025. Accepted: December 1st, 2025.

Available online: January 15th, 2026.

DOI: https://doi.org/10.12962/j24068535.v24i1.a1327

bidirectional approach allows BERT to grasp the full semantic information of a word or phrase [8], even when it's in a shorter sentence. Whereas traditional models would often struggle due to their reliance on linear word sequences and limited context windows. Meanwhile, longer statements can contain multiple layers of information, including truthful and deceptive elements intertwined. These types of statements are a common tactic to mask lies. This disorganized information is useless unless properly dissected [9] as it will dilute the overall classification regardless of the outcome. Therefore, explainability techniques are necessary to gain a deeper understanding of the underlying truthfulness of statements such as in [10]. By segmenting the text, investigators can isolate individual ideas or claims and examine each one independently. This granular approach makes it easier to identify inconsistencies, contradictions, or areas that require further investigation. Moreover, this approach also ensures the AI models are more transparent [11] and understandable to a human.

While classification models can technically be used for deception detection, they are not useful practically. Classification models are designed to assign discrete labels, such as "true" or "false", causing them to overlook the subtle nuances and sentiments that exist within the human language. As noted in [12], such models tend to obscure sentiment or veracity intensity, thereby limiting their utility in real-world scenarios that demand a more detailed and context-aware analysis. To provide a more comprehensive and informative representation of the intensity of a sentiment, raw numbers are often better. For example, [12] enables analysts and investigators to make better-informed judgments. Rather than labeling a statement as entirely true or false, a model might indicate that the statement is "70% likely to be true," which is a more realistic reflection of how truth and deception are often interwoven in a criminals statement.

While BERT (Bidirectional Encoder Representations from Transformers) can be used directly for various NLP tasks, its performance is significantly enhanced when integrated with complementary models and techniques. For instance, combining BERT with Convolutional Neural Networks (CNNs) [13] can help in capturing local patterns within text, such as key phrases or syntactic structures, that BERT alone might overlook. Similarly, integrating BERT with word embedding techniques [14] can deepen a model’s understanding of semantic relationships by incorporating domain-specific lexical features or sentiment cues. Additionally, the combination of BERT with Long Short-Term Memory (LSTM) networks [15] allows for improved modeling of sequential dependencies and temporal patterns in text, which is particularly valuable in tasks involving narrative or conversational data. Among these, word embeddings are exceptional in both sentiment analysis and veracity detection tasks [16], as they can represent subtle nuances in tone, emotion, and truthfulness embedded within textual data. Using these combinations not only improves accuracy but also enhances the interpretability and generalizability of the model across various linguistic contexts. While pre-trained word embeddings like Global Vectors for Word Representation (GloVe) are valuable tools for a general-purpose natural language processing task, they are not always the best solution as shown by [16]. These embeddings often underperform when dealing with domain-specific tasks such as deception detection or sentiment intensity analysis. To create a domain specific embedding, [17] uses a sentiment lexicon alongside CNN for the purpose of sentiment analysis. The usage of a sentiment lexicon allowed the model to properly adapt to the use case of deception detection, proven by it outperforming other pretrained word embeddings namely GloVe and Word2Vec.

Research done in [10] claims that explainability plays a significant role in rationalizing AI models by understanding the underlying mechanisms. Explainable AI models are capable of converting black boxes [18] and opaque machine learning models into transparent ones and outlines how certain factors impact their outputs [19]. Explainable models are usually in the form of post hoc models such as local interpretable model-agnostic explanations (LIME) and SHapley Additive exPlanations (SHAP) [20]. These models analyze outputs by attributing importance to individual features. Therefore, in the context of veracity assessment, explainable models can be used to identify which specific words or phrases most influenced the model’s decision regarding the truthfulness of a statement. This gives potential investigators valuable understanding of why a model flagged a certain statements veracity as untruthful. As such, explainable models not only give more responsible culpability but also allows human analysts

to verify the models reasoning. This creates results that are more actionable in real-world criminal investigative scenarios.

Despite advancements in integrating BERT, CNNs, and explainability techniques, there remains a significant research gap in applying these models specifically to deception detection within real-world criminal investigations. As such, this research aims to provide a more practical understanding of the veracity of statements through explainable machine learning models while also handling potential contextual issues with BERT embeddings and CNN. These embeddings help capture nuanced semantic relationships within the text, while the explainability techniques ensure that users can comprehend how and why certain veracity judgments are made, thereby fostering trust and accountability in the decision-making process. The remainder of this paper is organized as follows: Section 2 reviews existing literature on AI in deception detection and highlights current approaches, section 3 exhibits the proposed framework integrating BERT embeddings and explanibility techniques for veracity assessment, section 4 reports findings and compares them with pre-trained word embeddings. Finally, section 5 summarizes the conclusion of this experiment.

2. Related Works

Lexicons such as in [17] and [21] have attempted to incorporate deep learning techniques into sentiment analysis. Specifically, their model utilizes BERT, a sentiment lexicon, CNN, BiGRU (Bidirectional Gated Recurrent Unit), and attention mechanisms. In this architecture, the sentiment lexicon enhances emotional feature representation, while CNN and GRU networks are used to extract both primary and contextual information. An attention mechanism is then used to apply appropriate weighting to these features. The result is a significant improvement in classification performance between 30-50% resulting in 90-93.5% accuracy. Notably, much of this improvement was attributed to the inclusion of CNN.

Despite these advances, lexicons alone are often insufficient because they do not inherently encode sentiment information into the embedding space. Addressing this limitation, [22] proposes a refined global word embedding method that enhances traditional embeddings by incorporating various positional features. These include internal positions (e.g., a word's placement within a sentence) and external positions (e.g., a word's co-occurrence with sentiment-bearing terms across broader contexts). This integration of positional signals results in embeddings that are not only semantically rich but also sentiment-aware, allowing the model to differentiate between words with similar meanings but contrasting emotional tones. Results show a consistent performance improvement across sentiment analysis tasks, with classification accuracy increasing by 1–5%. Another study by [23] also attempts a similar task of augmenting word embeddings. In this case, they did it through semantic lexicons and transfer learning. However, although it performed well, it still performed worse than a dictionary model. Therefore, there are potential pitfalls when using a sentiment lexicon. According to the researchers, this is due to the model being unable to capture specific contexts where one word can fit multiple definitions.

Challenges in sentiment analysis also arise when targeting a specific domain or domains with high linguistic variability, such as veracity detection or social media. For instance, [17] addresses problems of sparsity and high dimensionality in the context of social media by integrating sentiment lexicons into BERT-based embeddings and using CNN to reduce dimensionality. This dual approach enhances the computational efficiency of the sentiment model and produces more representative word vectors, thus improving predictive accuracy. However, their study acknowledges that further gains could potentially be achieved by exploring other neural networks such as LSTM or Bi-LSTM.

Based on a survey on explainable AI by [24], innovation on sentiment analysis stems from deep neural network. However, much of their outputs are obfuscated throughout this process. They suggest that more tangible explanations are required in explaining these models to understand it’s inner workings. Thus, to improve the real-world applicability of sentiment analysis, particularly in high-stakes domains, [10] and [25] introduce the concept of eXplainable Lexicons (XLex). Their work focuses on the financial and biology sector, where model transparency is critical. Built using SHAP values, XLex generates sentiment lexicons that are both data-driven and interpretable.

Table 1: Sample data.

VideoID	Transcription	Truth Prop
BF001_1PT	My best friend is a really nice person. Um. She's always kind to everyone. She continues to just be herself around everyone. Um. She has taught me so much throughout, like I've…	0.77
BF001_2NL	She's actually really two faced and not fun to be around. Um she's really negative. Um. I don't like the person that she's become. Um. she's just really not herself usually and…	0.6
BF001_3NT	So this specific person is actually just a really mean and negative person. Um. I'm not sure why she thinks she needs to be that way but, um, before I actually knew who she …	0.77
BF001_4PL	This person is actually a really kind person. She has so many friends. She's very popular. Everyone, um, watches her and looks up to her. She's really pretty actually. Um, yeah…	0.42

Each word in the lexicon is annotated with SHAP values, enabling users to trace sentiment predictions back to specific input features. This interpretability significantly reduces the manual effort typically required for domain-specific lexicon creation and validation. Compared to the widely used Loughran-McDonald (LM) lexicon, XLex achieved over 40% average improvement in classification accuracy. Moreover, in terms of performance, XLex was 3–10x faster compared to transformer models such as FinBERT and RoBERTa.

Studies on the particulars of veracity detection tend to have inconsistentices. Many works in [26] use visual and facial cues in order to detect deception. Research done by [27] suggest this as well. However, [28] objects the general notion that visual cues and emotion recognition play a favorable role in affecting the accuracy of a model. According to them, these factors only work on emotionally charged lies. Subtle lies that contradict the emotion being portrayed can heavily work against these types of models. They also suggest that veracity assessment is much more theoretically sound than traditional accuracy metrics.Other research such as [29] and [30] claim that humans themselves play a component into hindering the results of a veracity assessment. In their experiment, human judges were allowed to overrule the results of an automated deception detection. This is done because they claim human intervention is still required in many real world applications, particulary deception detection. However, the results showed otherwise. They claim that the human's emotions played a factor into lowering the accuracy of the experiment. It's assumed that the human's weren't able to properly understand the outcomes of the automated deception detection and as such tend to rely on gut feeling or personal biases to overrule the results. Failures such as these suggest that explainable AI is the proper path in handling these black box methods.

3. Methodology

This section describes the research flow shown in Fig. 1. Further detailed sub steps will be discussed in the subsections to follow.

3.1. Collecting Data

The dataset used was a deception detection database provided by Miami University totaling 320 videos which were then manually transcribed. Participants were students or affiliates of the university between the ages 18 to 26 and spoke in English. The data features a truth proposition feature to further explore the nuances between veracity. The truth proposition was determined by calculating the number of truthful statements compared to the total amount of statements made. Samples of the data can be seen in Table 1.

3.2. Data Preprocessing

To ensure the data can be fairly and accurately analyzed, preprocessing is required. Data preprocessing extracts important and non-trivial knowledge from unstructured text data [31]. The steps chosen for this research consist of the following, in order:

1. **Decontraction.** This expands contracted word forms into their long form.
2. **Lowercasing.** Words written in different capitalizations can affect the overall performance of the model. The same word with different capitalizations can cause the model to recognize them as different words. Therefore, it is important to convert all uppercase letters to lowercase and ensure that all words are equivalent.

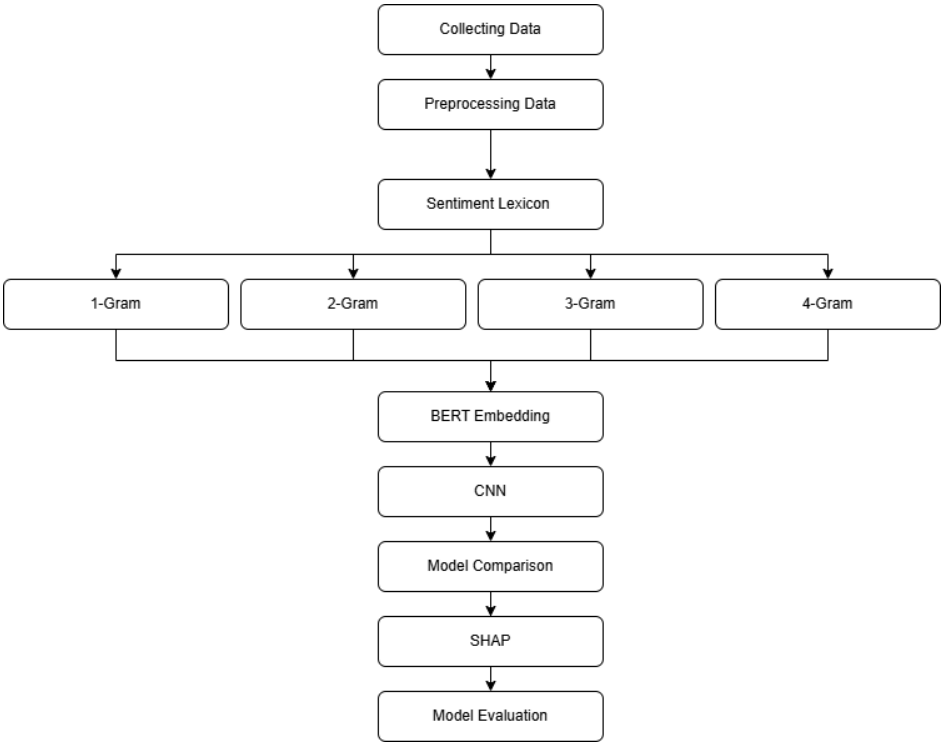

Fig. 1: Research process.

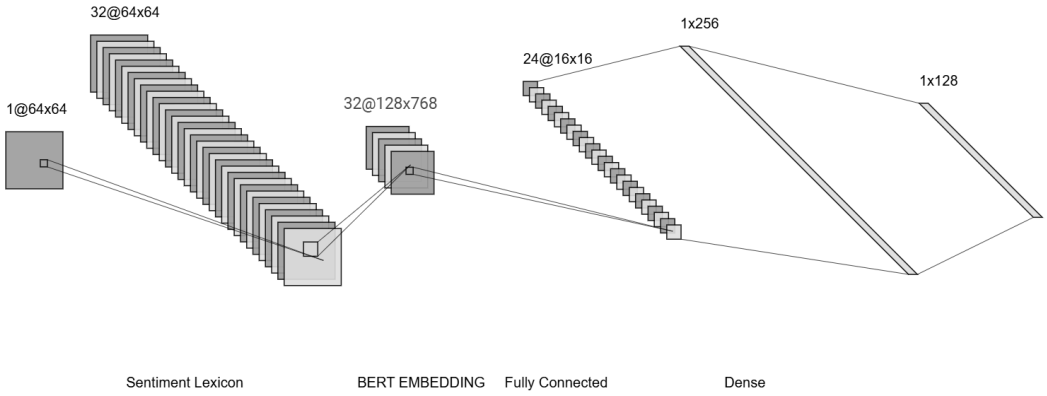

Fig. 2: CNN architecture for veracity assessment.

3. **Lemmatization.** The process of converting words into their dictionary form.
4. **Tokenization.** The process of breaking down text into small units called tokens, which can be words, phrases, symbols, or other meaningful elements.

3.3. BERT Sentiment Lexicon Embedding

In this research, this was done by using the VADER lexicon sentiment scores to create a BERT embedding. This hybrid representation combines BERT's contextual understanding with specific sentiment knowledge. Thus, this process can improve performance in sentiment analysis tasks by helping models focus on context specific information that BERT might not recognize on its own.

Different lengths of N-grams are used for each embedding to segment the input text. As different lengths provide potentially differing sentiments, context and understanding, this helps find the most appropriate n-gram for veracity assessment. The N-grams used were 1-gram, 2-gram, 3-gram and 4-gram.

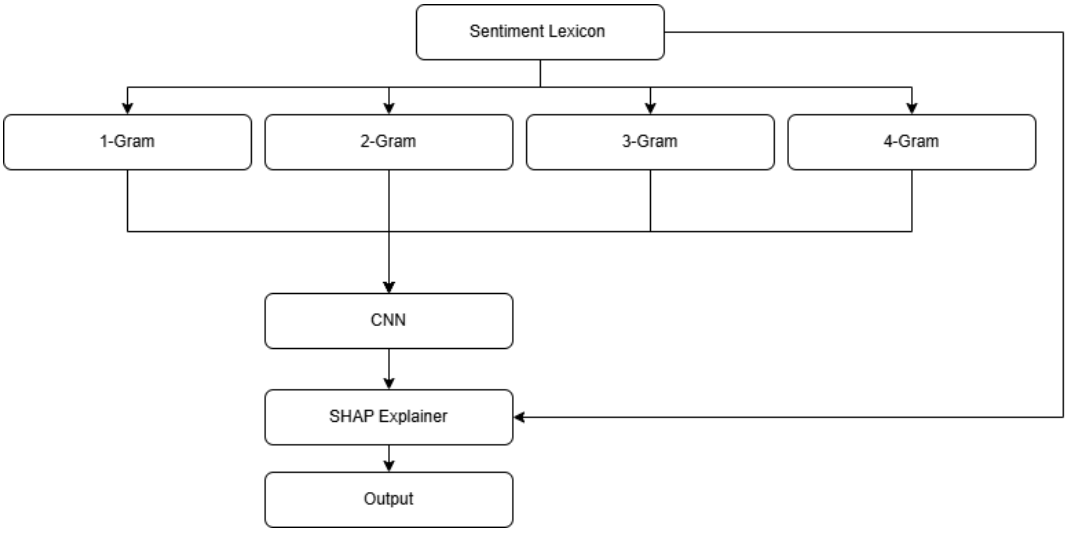

Fig. 3: SHAP process.

3.4. Convolutional Neural Network

3.5. Shapley Additive exPlanations

SHAP can be applied with BERT embeddings to interpret the influence of individual input tokens on a model's prediction. This requires the input tokens as well as the outputs of the model to be used as inputs for SHAP. As shown in Fig. 3, to apply SHAP in this research, the sentiment lexicon data is used as the first input and the results from the CNN as the second input. SHAP then estimates how each input token contributes to the final prediction by influencing the inputs and analyzing the corresponding changes in output.

This process enables a model-agnostic interpretability method where feature attributes are computed using game-theoretic principles. Specifically, SHAP assigns each token an importance value or SHAP value, which quantifies its contribution to the model's output. By influencing the input data and observing the variation in the output, SHAP simulates the effect of each token in the presence of all possible subsets of other tokens.

In this research, SHAP was applied to the CNN outputs over the sentiment lexicon N-grams. This setup allows for an examination of how individual words contribute to a prediction of the veracity of a sentence. The SHAP values provide token-level interpretability, making it possible to visualize which words increased or decreased the predicted probability of a claim being truthful.

Moreover, this approach helps identify potential biases or failure points within the model by revealing unexpected patterns in token contributions. These can then be used as points of human intervention if necessary. Thus, it serves both as a diagnostic tool for model refinement and a means to build trust in automated veracity classification systems.The formula for SHAP can be seen in Eq. (1) where i is the feature being predicted, F is the set of all features, S is a subset of features that does not include i, |S| is the number of features in subset S, |F| is the total number of features, $f_s(x_s)$ is the model prediction using only features in S, and $f_{(s\cup\{i\})}(x_{(s\cup\{i\})})$ is the model prediction using features in S plus feature i.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S) \right) \tag{1}$$

3.6. Model Evaluation Metrics

The model evaluation will be carried out based on several quantitative evaluation metrics that are widely used in regression models. These metrics include: R^2 (coefficient of determination), RMSE (Root Mean Squared Error),

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{2}$$

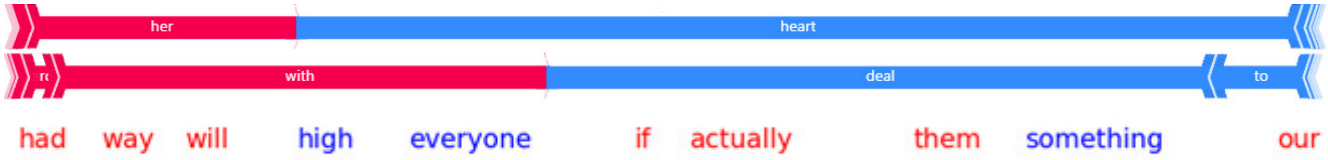

Fig. 4: SHAP single sentence analysis.

Table 2: Evaluation Metric Results

Word Embedding	R^2		RMSE		MSE		MAPE		SMAPE	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
1-Gram BERT Embedding	0.320	0.250	0.140	0.150	0.020	**0.020**	23%	25%	21%	22%
2-Gram BERT Embedding	**0.390**	**0.310**	**0.130**	**0.140**	**0.019**	0.021	**22%**	**23%**	**19%**	**21%**
3-Gram BERT Embedding	0.370	0.290	**0.130**	**0.140**	0.020	**0.020**	23%	24%	20%	**21%**
4-Gram BERT Embedding	0.310	0.220	0.140	0.150	0.030	**0.020**	23%	26%	21%	22%
GloVe	−0.052	−0.059	0.352	0.362	1.240	1.311	175%	96%	184%	185%
Word2Vec	−0.350	−0.360	0.232	0.251	1.210	1.250	87%	93%	95%	98%
FastText	−0.412	−0.389	0.250	0.267	1.130	1.270	65%	88%	84%	88%

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{5}$$

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)\ /\ 2} \tag{6}$$

The R^2 is used to measure how well a model can explain the variation in the target data. A higher R^2 value indicates that the model is better at capturing the relevant patterns between inputs and outputs. RMSE measures the average distance between predicted and actual values on the same scale. Lower RMSE values indicate that the model makes more accurate predictions. Meanwhile, MSE values provides the average squared error, which emphasizes the effect of outliers. This metric is useful for identifying whether a model is unstable when outliers are present. Next, MAPE expresses prediction accuracy as a percentage relative to the actual values. It is scale-independent, meaning it is not affected by whether the actual values are very large or very small. Lower percentages indicate higher accuracy. Finally, SMAPE is a variation of MAPE that balances errors symmetrically, reducing bias between overprediction and underprediction. This makes it useful for evaluating whether a model performs fairly in both directions. The formula for each of these can be seen in Eqs. (2) to (6) where y_i represents the actual value of the data, $\hat{y}_i$ represents the predicted value from the model, $\bar{y}$ is the mean of the actual values, and n is the total number of data points.

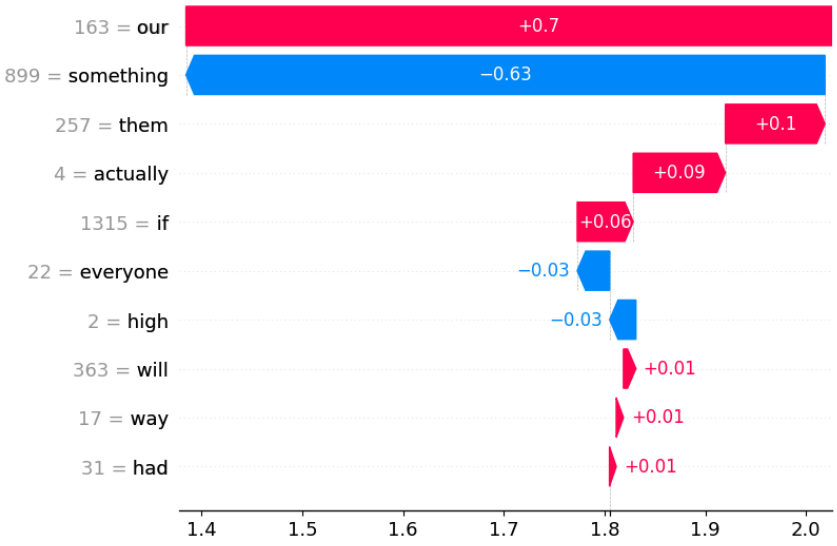

Fig. 5: Waterfall word impact on veracity of a single statement.

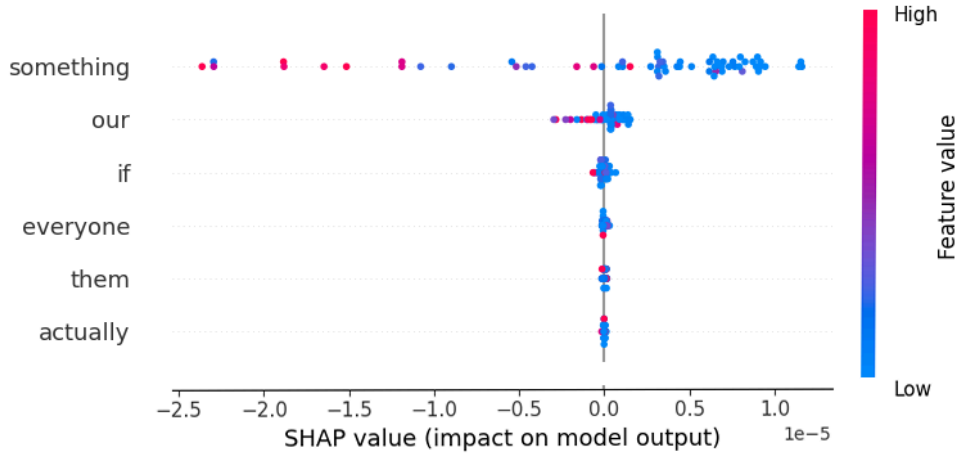

Fig. 6: SHAP veracity values.

4. Results and Discussion

Among all configurations, the 2-gram BERT embedding shows the best overall performance, achieving the highest R^2 scores (0.39 train, 0.31 test) and the lowest errors across most metrics (e.g., RMSE and MAPE). It slightly outperforms the 3-gram embedding, which also performs well. The 1-gram and 4-gram embeddings perform less effectively, with lower R^2 scores and slightly higher errors. 2-gram embedding performs the best by being able to capture more context than a 1-gram embedding. As the higher the n-gram used is, the more specific it is and the less samples it has, thus 2-gram also performed slightly better than 3-gram and marginally better than 4-gram. With more samples, 3-gram has the potential to be equal if not better than 2-gram.

In contrast, GloVe, Word2Vec and FastText performed significantly worse, with negative R^2 values and very high error metrics across the board. This indicates that the linear models using these embeddings explain less variance than simply predicting the mean. GloVe, despite its popularity, shows particularly poor generalization with test MAPE above 95% and SMAPE near 185%, suggesting that its static embeddings lack the contextual depth needed for this task. Similarly, Word2Vec embeddings, while slightly better in RMSE than GloVe, still yield unacceptably high error percentages and negative R^2 scores, reinforcing their unsuitability for the problem at hand. Each of these results indicate that general word embeddings are not well suited for the delicate task of veracity assessment.

Table 3: Most impactful words.

Word	**Veractiy Value**
Something	−0.532
Them	−0.432
Everyone	−0.180
Our	−0.140
If	−0.110
Sure	+0.134
Always	+0.163
Smart	+0.166
Kid	+0.184
Overall	+0.465

In Fig. 5, we can see the degree of veracity a certain word had in a single given sentence. In the example, the word ‘our’ had a high degree of deceptiveness while the word ‘something’ had a high degree of truthfulness in this specific context. Given the context, the word ‘our’ was likely used to deflect or share responsibility while the word ‘something’ expressed uncertainly by the speaker. These types of insights can be useful in further solidifying the veracity of a statement. However, as visualized in Fig. 6, tokens such as ‘something’ and ‘our’ were observed to exhibit high variability in SHAP values across different samples. This indicates that their contribution can significantly sway the model’s prediction towards either truthful or deceptive classifications depending on surrounding context. Insights such as this could prove valuable for identifying ambiguous language, highlighting cognitive dissonance and pinpointing key lexical triggers within deceptive or emotionally charged statements.

The analysis of the average veracity values in Table 3 shows the words that most often appear in deceptive and truthful statements. Negative values indicate that the word often appears in deceptive statements while positive values indivate that they appear in truthful statements. Several of these terms can act as subtle linguistic indicators of psychological distancing or uncertainty, common traits in deceptive communication. Words such as ‘Something’, ‘Them’, and ‘Everyone’ are strongly associated with uncertainty, avoidance, or shifting responsibility. Similarly, words like ‘Our’ and ‘If’ often indicate persuasive or speculative patterns. These findings suggest that linguistic markers of deception are characterized not only by word frequency but also by the negative or uncertain sentiment embedded within specific terms. When examined in context, such words may not definitively signal dishonesty, but they provide important leads for analysts or algorithms aiming to assess the trustworthiness of a statement.

In contrast, words linked to truthful statements display positive sentiment and emphasize clarity and certainty. For example, ‘Overall’ shows a strong positive association, reflecting comprehensive and consistent communication. Words such as ‘Kid’, ‘Smart’, and ‘Always’ highlight straightforwardness and reliability, while ‘Sure’ reflects speaker confidence. Collectively, these words illustrate that truthful statements are marked by sentiment patterns that reinforce credibility through certainty and clarity.

Overall, these results show that the combination of 2-gram BERT embeddings and CNN supplemented with SHAP interpretability, offers a robust framework for analyzing linguistic veracity with both predictive strength and interpretive transparency. Integrating this type of lexical insight into deception detection models can improve both interpretability and effectiveness.

5. Conclusions

This research uses the Miami University Deception Detection Database. The data was analyzed with a Lexicon BERT Embedding then compared with existing word embeddings. From this experiment, the 2-gram Embedding performed better than the 1-gram, 3-gram and 4-gram at an R^2 of 0.39. This was due to 2-gram being able to capture more context than 1-gram while being more general than 3-gram and 4-gram. Each iteration of the Lexicon BERT Embedding was a significant improvement over each of the standard word embeddings by a large margin.

By analyzing each individual word, it is possible to conduct further analysis for more confident decision making. This interpretability can trace which words or phrases most influence the model's final decision such as the words 'everyone', 'something' and 'our'. Due to the limited size of the dataset and the crucial nature of the context of deception detection, these results can't be taken at face value. Instead they should be used as a starting point for decision making. Further research can still be done to verify these results, particulary on a larger dataset or within the context of facial and visual cues instead.

CRediT Authorship Contribution Statement

T. I. Haq: Writing – Original Draft, Writing – Review & Editing, Validation, Software, Methodology, Conceptualization. **C. Fatichah:** Conceptualization, Methodology, Validation, Formal analysis, Resources, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition. **A. Yuniarti:** Conceptualization, Methodology, Validation, Formal analysis, Resources, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Thoha I. Haq thanks Muhammad Meftah Mafazy for acquiring dataset permissions and Miami University for providing the dataset used in this research. Each of their support was essential in providing a means to research the topic of this article.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Declaration of Generative AI and AI-assisted Technologies in The Writing Process

The authors used generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and take full responsibility for the final publication.

References

[1] N. S. Pandey, P. Rawat, S. Kathuria, R. Singh, G. Chhabra, and G. Pant, "Artificial Intelligence Assistance in the Domain of Law," in *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, 2023, pp. 1–4. doi: 10.1109/InC457730.2023.10263230.

[2] P. Raj, P. Rawat, J. Singh, S. Pandey, S. Aluvala, and V. Pachouri, "Law Enforcement and Dispensation of Judicial Equipoise: Convergence of Artificial Intelligence in Administration of Justice," in *2024 Parul International Conference on Engineering and Technology (PICET)*, 2024, pp. 1–5. doi: 10.1109/PICET60765.2024.10716181.

[3] M. Gogate, A. Adeel, and A. Hussain, "Deep learning driven multimodal fusion for automated deception detection," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–6. doi: 10.1109/SSCI.2017.8285382.

[4] B. V. Mendes, A. M. Tomé, I. M. Santos, and P. Bem-Haja, "Analysis of eyewitness testimony using electroencephalogram signals," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 894–897. doi: 10.1109/EMBC46164.2021.9630054.

[5] S. Raaijmakers, "Artificial Intelligence for Law Enforcement: Challenges and Opportunities," *IEEE Security & Privacy*, vol. 17, no. 5, pp. 74–77, 2019, doi: 10.1109/MSEC.2019.2925649.

[6] J. Zhu and Y. Cao, "Research on Performance Enhancement Strategies for Multimodal Short Text Classification by Fusing BERT and CNNs," in *2024 9th International Symposium on Computer and Information Processing Technology (ISCIPT)*, 2024, pp. 464–469. doi: 10.1109/ISCIPT61983.2024.10672913.

[7] S. Venkatesh, R. Ramachandra, and P. Bours, "Robust Algorithm for Multimodal Deception Detection," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 534–537. doi: 10.1109/MIPR.2019.00108.

[8] C. Yang, X. Wang, M. Li, and J. Li, "Research on fusion model of BERT and CNN-BiLSTM for short text classification," in *2023 4th International Conference on Computer Engineering and Application (ICCEA)*, 2023, pp. 525–529. doi: 10.1109/ICCEA58433.2023.10135222.

[9] M. Habibullah, M. S. Islam, F. Tuz Jahura, and J. Biswas, "Bangla Document Classification Based on Machine Learning and Explainable NLP," in *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*, 2023, pp. 1–6. doi: 10.1109/EICT61409.2023.10427766.

[10] A. Sankaranarayanan, D. Shetty, K. Chetwani, and B. R. Shambhavi, "Exploring BioClinical BERT's NLP Capabilities with Explainability Techniques," in *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, 2024, pp. 1–6. doi: 10.1109/ICETCS61022.2024.10544307.

[11] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali, and M. Abomhara, "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI," *IEEE Access*, vol. 12, no. , pp. 44462–44480, 2024, doi: 10.1109/ACCESS.2024.3381038.

[12] D. Lin, Y. Wen, W. Wang, and Y. Su, "Enhanced Sentiment Intensity Regression Through LoRA Fine-Tuning on Llama 3," *IEEE Access*, vol. 12, no. , pp. 108072–108087, 2024, doi: 10.1109/ACCESS.2024.3438353.

[13] C. Raj and P. Meel, "Microblogs Deception Detection using BERT and Multiscale CNNs," in *2021 2nd Global Conference for Advancement in Technology (GCAT)*, 2021, pp. 1–6. doi: 10.1109/GCAT52182.2021.9587698.

[14] A. M. Al-Zoubi, A. M. Mora, and H. Faris, "A Multilingual Spam Reviews Detection Based on Pre-Trained Word Embedding and Weighted Swarm Support Vector Machines," *IEEE Access*, vol. 11, no. , pp. 72250–72271, 2023, doi: 10.1109/ACCESS.2023.3293641.

[15] A. He and M. Abisado, "Text Sentiment Analysis of Douban Film Short Comments Based on BERT-CNN-BiLSTM-Att Model," *IEEE Access*, vol. 12, no. , pp. 45229–45237, 2024, doi: 10.1109/ACCESS.2024.3381515.

[16] D. Nam, J. Yasmin, and F. Zulkernine, "Effects of Pre-trained Word Embeddings on Text-based Deception Detection," in *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2020, pp. 437–443. doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00083.

[17] J. Mutinda, W. Mwangi, and G. Okeyo, "Sentiment analysis of text reviews using lexicon-enhanced bert embedding (LeBERT) model with convolutional neural network," *Applied Sciences*, vol. 13, no. 3, p. 1445, 2023, doi: 10.3390/app13031445.

[18] M. H. Fahim Siddiqui, D. Inkpen, and A. Gelbukh, "Towards Interpretable Emotion Classification: Evaluating LIME, SHAP, and Generative AI for Decision Explanations," in *2024 28th International Conference Information Visualisation (IV)*, 2024, pp. 1–6. doi: 10.1109/IV64223.2024.00053.

[19] S. Ahmed, M. S. Kaiser, M. Shahadat Hossain, and K. Andersson, "A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions," *IEEE Access*, vol. 13, no. , pp. 37370–37388, 2025, doi: 10.1109/ACCESS.2024.3422319.

[20] D. K. Jain, A. Rahate, G. Joshi, R. Walambe, and K. Kotecha, "Employing Co-Learning to Evaluate the Explainability of Multimodal Sentiment Analysis," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 4673–4680, 2024, doi: 10.1109/TCSS.2022.3176403.

[21] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access*, vol. 8, no. , pp. 23522–23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

[22] Y. Wang, G. Huang, J. Li, H. Li, Y. Zhou, and H. Jiang, "Refined Global Word Embeddings Based on Sentiment Concept for Sentiment Analysis," *IEEE Access*, vol. 9, no. , pp. 37075–37085, 2021, doi: 10.1109/ACCESS.2021.3062654.

[23] T. Alshaabi, C. M. Van Oort, M. I. Fudolig, M. V. Arnold, C. M. Danforth, and P. S. Dodds, "Augmenting semantic lexicons using word embeddings and transfer learning," *Frontiers in Artificial Intelligence*, vol. 4, p. 783778, 2022, doi: 10.3389/frai.2021.783778.

[24] A. Diwali, K. Saeedi, K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 837–846, 2024, doi: 10.1109/TAFFC.2023.3296373.

[25] M. Rizinski, H. Peshov, K. Mishev, M. Jovanovik, and D. Trajanov, "Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex)," *IEEE Access*, vol. 12, no. , pp. 7170–7198, 2024, doi: 10.1109/ACCESS.2024.3349970.

[26] A. S. Constâncio, D. F. Tsunoda, H. d. F. N. Silva, J. M. d. Silveira, and D. R. Carvalho, "Deception detection with machine learning: A systematic review and statistical analysis," *Plos one*, vol. 18, no. 2, p. e281323, 2023, doi: 10.1371/journal.pone.0281323.

[27] S. Salah *et al.*, "Explainable AI for Unraveling the Significance of Visual Cues in High Stakes Deception Detection," *IEEE Access*, vol. 13, no. , pp. 65839–65862, 2025, doi: 10.1109/ACCESS.2025.3558875.

[28] M. Zloteanu, P. Bull, E. G. Krumhuber, and D. C. Richardson, "Veracity judgement, not accuracy: Reconsidering the role of facial expressions, empathy, and emotion recognition training on deception detection," *Quarterly Journal of Experimental Psychology*, vol. 74, no. 5, pp. 910–927, 2021, doi: 10.1177/1747021820978851.

[29] A. Turi, M.-R. Rebeleș, and L. Visu-Petra, "The tangled webs they weave: A scoping review of deception detection and production in relation to Dark Triad traits," *Acta Psychologica*, vol. 226, p. 103574, 2022, doi: 10.1016/j.actpsy.2022.103574.

[30] B. Kleinberg and B. Verschuere, "How humans impair automated deception detection performance," *Acta psychologica*, vol. 213, p. 103250, 2021, doi: 10.1016/j.actpsy.2020.103250.

[31] S. Kannan *et al.*, "Preprocessing techniques for text mining," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2014.

[32] Y. Wang, L. Cui, and Y. Zhang, "Improving Skip-Gram Embeddings Using BERT," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, no. , pp. 1318–1328, 2021, doi: 10.1109/TASLP.2021.3065201.

[33] I. Üveges and O. Ring, "HunEmBERT: A Fine-Tuned BERT-Model for Classifying Sentiment and Emotion in Political Communication," *IEEE Access*, vol. 11, no. , pp. 60267–60278, 2023, doi: 10.1109/ACCESS.2023.3285536.

[34] M. Huang, H. Xie, Y. Rao, Y. Liu, L. K. M. Poon, and F. L. Wang, "Lexicon-Based Sentiment Convolutional Neural Networks for Online Review Analysis," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1337–1348, 2022, doi: 10.1109/TAFFC.2020.2997769.

[35] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022, doi: 10.1109/TNNLS.2021.3084827.