Sarcasm Detection Using Word Embeddings: A Comparative Study of Word2Vec, GloVe, and BERT

Breeha Qasim, Namel Shahid, Ashbah Faisal
Habib University
Karachi, Pakistan

Abstract

Sarcasm detection in text remains a challenging natural language processing task due to the subtle linguistic cues and contextual dependencies that characterize sarcastic expressions. This study presents a comprehensive comparative analysis of three word embedding approaches for sarcasm detection: Word2Vec trained from scratch using skip-gram with negative sampling, GloVe pre-trained embeddings, and BERT transformer-based contextual embeddings. We evaluate these approaches on two distinct datasets representing different text domains: news headlines and Twitter posts. Our experiments employ a Bidirectional LSTM architecture for all embedding models, including a BERT + BiLSTM combination for contextual embeddings. Results demonstrate that BERT achieves the highest F1-score of 0.900 on the headlines dataset, while Word2Vec performs competitively on the tweets dataset with an F1-score of 0.776. The study reveals significant domain-specific performance variations, with all models exhibiting decreased precision on the tweets dataset due to its informal and noisy nature.

1 Introduction and Literature Review

Sarcasm detection has emerged as a critical yet challenging task in natural language processing, requiring models to capture subtle linguistic nuances and contextual incongruities that signal ironic intent. The research community has explored various approaches ranging from feature engineering to deep learning methods, with recent focus on multi-modal and multi-task learning frameworks.

Chauhan et al. [2020] proposed a multi-task learning framework for multi-modal sarcasm, sentiment, and emotion analysis. Their work demonstrated that sentiment and emotion can serve as auxiliary tasks to improve sarcasm detection performance, leveraging the inherent relationships between these affective computing tasks. This approach recognizes that sarcastic expressions often involve emotional incongruity, where the expressed sentiment contradicts the intended meaning. The authors showed that joint learning across these related tasks enables the model to capture richer semantic representations that benefit sarcasm classification.

Pan et al. [2020] addressed the multi-modal aspect of sarcasm detection by modeling intra and inter-modality incongruity. Their work highlighted that sarcasm often manifests through mismatches between different modalities, such as text and images, or between different parts of the same modality. The incongruity-based approach provides a principled framework for understanding how sarcastic intent is conveyed through contrasting signals. While our study focuses on text-only sarcasm detection, the concept of incongruity remains relevant as sarcastic text often contains internal contradictions between literal and implied meanings.

Savini and Caragea [2020] further explored multi-task learning for sarcasm detection, demonstrating that auxiliary tasks can enhance the primary sarcasm classification objective. Their work provided evidence that shared representations learned across related tasks capture complementary linguistic features. Similarly, Tay et al. [2020] developed a comprehensive multi-task learning framework that jointly models sarcasm with sentiment and emotion recognition, achieving strong performance on benchmark datasets. This supports the broader trend of leveraging transfer learning and multi-task approaches in NLP, suggesting that sarcasm detection benefits from learning general language understanding capabilities alongside task-specific features.

More recently, Jang and Frassinelli [2024] addressed the critical issue of generalizability in sarcasm detection. Their work highlighted that many existing sarcasm detection models overfit to dataset-specific patterns and fail to generalize across different domains and data sources. This research is particularly relevant to our study as we evaluate models on two distinct domains: news headlines and tweets. Their findings underscore the importance of cross-domain evaluation and the development of more robust sarcasm detection

approaches.

Building on these foundational works, our study contributes to the literature by providing a systematic comparison of three prominent word embedding approaches across multiple domains. We investigate how different text representations, including Word2Vec, GloVe, and BERT, affect sarcasm detection performance. By comparing embeddings trained from scratch versus pre-trained representations, we provide insights into the trade-offs between model complexity, computational requirements, and classification accuracy across varied text domains.

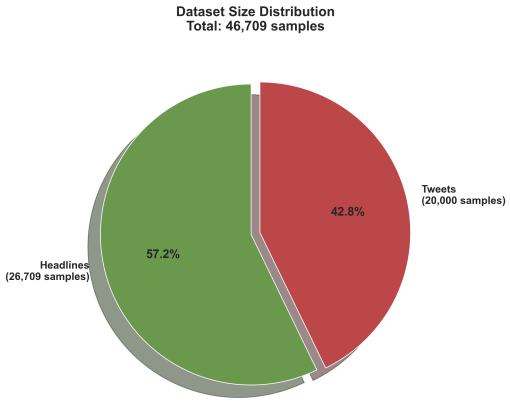

Figure 1: Dataset size distribution showing the proportion of headlines and tweets samples in the combined corpus.

2 Methodology

2.1 Datasets

We conducted experiments on two publicly available sarcasm detection datasets representing distinct text domains. The first dataset consists of 26,709 news headlines collected from The Onion (sarcastic) and HuffPost (non-sarcastic), with a sarcastic ratio of approximately 43.9%. News headlines represent formal, edited text with consistent grammatical structure. The second dataset comprises Twitter posts representing informal social media text characterized by abbreviations, hashtags, and unconventional grammar. The tweets dataset was originally provided in CSV format, which we converted to JSON format for consistency with our data pipeline. The original dataset contained over 80,000 entries with four sentiment classes; however, we reduced it to 20,000 samples due to computational constraints and training time considerations. We also modified the classification scheme from four classes to binary classification by mapping the original labels to sarcastic and not sarcastic categories, resulting in a sarcastic ratio of approximately 25.1%.

Figure 1 illustrates the size distribution between the two datasets, with headlines comprising 57.2% (26,709 samples) and tweets accounting for 42.8% (20,000 samples) of the total corpus of 46,709 samples. Figure 2 presents the class distribution within each dataset, revealing a notable imbalance difference. The headlines dataset exhibits a more balanced distribution with 43.9% sarcastic content, while the tweets dataset shows greater imbalance with only 25.1% sarcastic samples. This class distribution disparity poses challenges for model training and evaluation across domains.

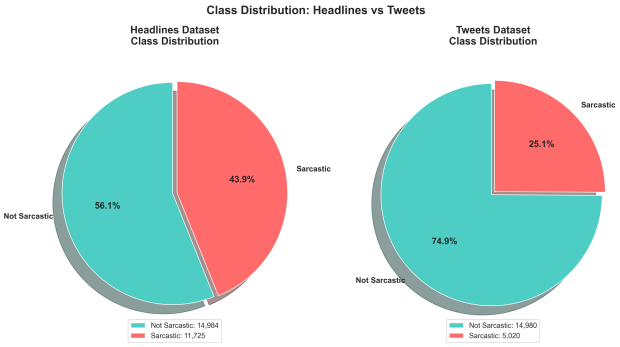

Figure 2: Class distribution comparison between headlines (43.9% sarcastic) and tweets (25.1% sarcastic) datasets, highlighting the class imbalance variation across domains.

For both datasets, we applied a 70-30 train-test split with stratified sampling to maintain class distribution. Text preprocessing included lowercase conversion and removal of special characters while preserving letters, numbers, and spaces. For the BERT model, we utilized the built-in tokenization with a maximum sequence length of 100 tokens.

2.2 Word2Vec Model

We implemented Word2Vec from scratch using the skip-gram architecture with negative sampling. The model learns distributed word representations by predicting context words given a center word. The skip-gram objective with negative sampling is defined as:

$$\mathcal{L} = \log \sigma(v'_{w_O} \cdot v_{w_I}) + \sum_{i=1}^{k} E_{w_i \sim P_n(w)} [\log \sigma(-v'_{w_i} \cdot v_{w_I})] \quad (1)$$

where v_{w_I} is the input vector of the center word, v'_{w_O} is the output vector of the context word, σ is the

sigmoid function, k is the number of negative samples, and $P_n(w)$ is the noise distribution. We use the smoothed unigram distribution $P_n(w) \propto U(w)^{0.75}$ where $U(w)$ is the unigram frequency.

The Word2Vec model was configured with an embedding dimension of 128, a context window size of 10, and 10 negative samples per positive example. Vocabulary was built from the training corpus with a minimum word frequency threshold of 2 to filter rare tokens. The embedding matrices were initialized with uniform random values scaled by the embedding dimension. Training proceeded for 25 epochs with an initial learning rate of 0.025 and a decay factor of 0.95 per epoch. Negative sampling probabilities were computed using the smoothed unigram distribution with an exponent of 0.75.

We chose to combine Word2Vec with BiLSTM because Word2Vec produces static word embeddings that lack sequential context awareness. While Word2Vec captures semantic relationships between words, it cannot model the order-dependent patterns crucial for sarcasm detection. The BiLSTM component addresses this limitation by processing the embedded sequence bidirectionally, capturing both preceding and following context that helps identify sarcastic intent through word arrangement and phrase-level patterns.

2.3 GloVe Model

For GloVe embeddings, we utilized pre-trained vectors from the GloVe 6B corpus trained on Wikipedia and Gigaword data. We employed 100-dimensional embeddings containing 400,000 vocabulary words. Unlike Word2Vec which we trained from scratch, GloVe provides general-purpose embeddings that capture global co-occurrence statistics from a large external corpus. This approach evaluates the effectiveness of transfer learning from pre-trained embeddings to the sarcasm detection task.

The rationale for pairing GloVe with BiLSTM follows similar reasoning to Word2Vec. GloVe embeddings, while trained on massive corpora and encoding rich semantic information, are context-independent representations. Each word receives the same vector regardless of its surrounding context. The BiLSTM layer compensates for this limitation by learning sequential dependencies specific to sarcasm patterns, enabling the model to distinguish between literal and sarcastic usage of the same words based on contextual cues.

2.4 BERT Model

We implemented a BERT + BiLSTM model leveraging the pre-trained bert-base-uncased model with 110 million parameters. The architecture consists of four main components: the BERT encoder producing 768-dimensional contextualized representations, a sentence encoding layer with a dense transformation and ReLU activation, a context summarization layer using masked mean pooling to aggregate token representations into a single vector, and a BiLSTM layer with 128 hidden units. The final classification is performed through a fully connected layer with 32 units followed by a sigmoid output layer.

Although BERT already produces contextualized embeddings through its self-attention mechanism, we augmented it with a BiLSTM layer for consistency with our other models. However, it should be noted that the BiLSTM processes the pooled BERT representation (a single vector per sample) rather than the full token sequence, meaning its sequential modeling contribution is limited in this architecture. The design primarily leverages BERT's contextual understanding with additional dense transformations for classification.

The BERT model was fine-tuned for 5 epochs using the Adam optimizer with a learning rate of 2e-5. We used a batch size of 32 and maximum sequence length of 100 tokens. The model was trained with binary cross-entropy loss and employed dropout for regularization.

2.5 BiLSTM Classifier

For Word2Vec and GloVe embeddings, we employed a shared Bidirectional LSTM classification architecture. The BiLSTM processes input sequences in both forward and backward directions:

$$\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}}), \quad \overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}}) \quad (2)$$

The bidirectional hidden states are concatenated as $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. Pooling is essential because the BiLSTM produces a sequence of hidden states (one per time step), but classification requires a fixed-size representation. We apply both mean and max pooling over the sequence:

$$h_{mean} = \frac{1}{T}\sum_{t=1}^{T} h_t, \quad h_{max} = \max_{t=1}^{T}(h_t) \quad (3)$$

Mean pooling computes the average of all hidden states, capturing the overall semantic content distributed across the entire sequence. Max pooling selects the maximum activation for each dimension, highlighting the most salient features regardless of their position, which is particularly useful for detecting strong

sarcasm indicators that may appear at specific locations. We concatenate both pooling outputs as $h_{final} = [h_{mean}; h_{max}]$ because they capture complementary information: mean pooling provides a smooth summary of the entire context while max pooling preserves peak activations that might correspond to key sarcastic cues. This dual pooling strategy has been shown to outperform using either method alone.

The final representation is passed through dense layers for classification. The model consists of embedding dropout (0.2), a two-layer BiLSTM with 64 hidden units per direction and inter-layer dropout of 0.3, dual pooling concatenated to form a 256-dimensional representation, and a three-layer classification head with dimensions 128, 32, and 1, incorporating batch normalization, ReLU activations, and dropout.

The classifier was trained for 40 epochs using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 100. Input sequences were padded or truncated to a maximum length of 50 tokens. Gradient clipping with a maximum norm of 1.0 was applied to prevent exploding gradients. Binary cross-entropy with logits loss was used as the training objective. All experiments used random seed 42 for reproducibility.

2.6 Evaluation Metrics

Model performance was evaluated using four standard classification metrics. Accuracy measures overall correctness, precision quantifies the quality of positive predictions, recall captures coverage of actual positives, and F1-score provides a balanced measure:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively. We also report the complete confusion matrix for detailed error analysis.

3 Results

3.1 Overall Performance Comparison

As visualized in Figure 3 and Figure 5, BERT achieves the highest overall performance on the headlines dataset with an F1-score of 0.900, followed by GloVe (0.864) and Word2Vec (0.854). The performance gap narrows considerably on the tweets dataset, where Word2Vec achieves the highest F1-score (0.776), marginally outperforming BERT (0.772) and GloVe (0.760). On headlines, Word2Vec achieves accuracy of 0.865, precision of 0.816, and recall of 0.895, while on tweets it achieves accuracy of 0.859, precision of 0.645, and recall of 0.975. GloVe performs with accuracy of 0.881, precision of 0.866, and recall of 0.863 on headlines, and accuracy of 0.854, precision of 0.647, and recall of 0.920 on tweets. BERT leads on headlines with accuracy of 0.913, precision of 0.911, and recall of 0.889, while on tweets it achieves accuracy of 0.852, precision of 0.630, and near-perfect recall of 0.999.

3.2 Confusion Matrix Analysis

The confusion matrices, as illustrated in Figure 4, reveal interesting patterns in model behavior across domains. On the headlines dataset, all models exhibit relatively balanced distributions. BERT achieves TN=2095, FP=153, FN=195, and TP=1564, indicating robust discrimination. Word2Vec shows TN=3784, FP=712, FN=368, and TP=3149, while GloVe achieves TN=4027, FP=469, FN=483, and TP=3034.

On the tweets dataset, all models shift toward high recall at the expense of precision. BERT demonstrates an extreme case with only 1 false negative (TN=1805, FP=442, FN=1, TP=752). Word2Vec shows TN=3688, FP=807, FN=38, and TP=1467, while GloVe achieves TN=3741, FP=754, FN=120, and TP=1385. This pattern suggests models over-predict sarcasm on the noisier tweets dataset.

3.3 Visual Analysis

Figure 3 presents bar charts breaking down all evaluation metrics. The accuracy subplot (top-left) shows consistent performance across models regardless of dataset. The precision subplot (top-right) reveals the most significant cross-domain variation, with all models achieving strong precision on headlines but experiencing dramatic drops on tweets due to increased false positives. The recall subplot (bottom-left) demonstrates the inverse pattern, with tweets achieving higher recall as models rarely miss sarcastic content. The F1-score subplot (bottom-right) balances these trade-offs, showing the performance ranking shifts between datasets.

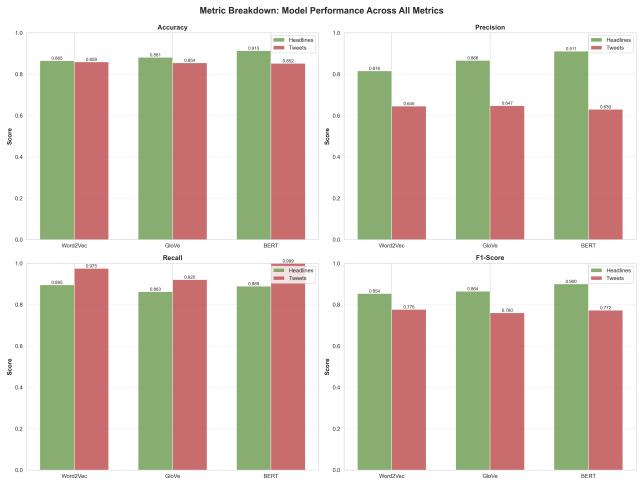

Figure 3: Detailed metric breakdown showing Accuracy, Precision, Recall, and F1-Score for all models across both datasets.

Figure 4 displays confusion matrix heatmaps in a 2x3 grid (datasets as rows, models as columns). Color intensity represents prediction counts, with darker blue indicating higher values. The headlines row (top) shows balanced distributions with strong diagonal elements (TN and TP) across all models. The tweets row (bottom) reveals a striking pattern: intensely colored TP cells contrast with nearly empty FN cells, while substantial FP cells indicate over-prediction bias. This visual pattern demonstrates that informal language patterns may be misinterpreted as sarcastic cues.

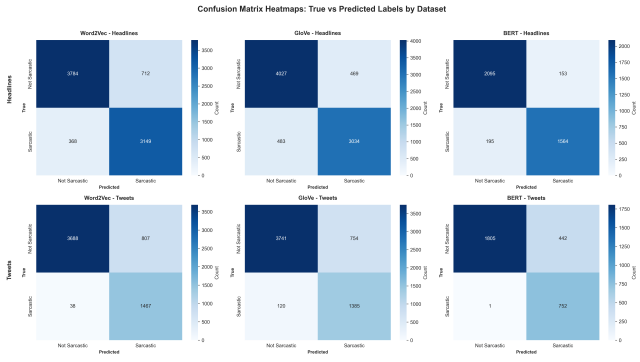

Figure 4: Confusion matrix heatmaps organized by dataset (rows) and model (columns).

Figure 5 illustrates the domain gap through grouped bar charts. Every model shows performance degradation from headlines to tweets. BERT experiences the largest drop (0.128), suggesting its contextual understanding may not transfer effectively to informal social media language. Word2Vec shows the smallest drop (0.078), while GloVe falls by 0.104. Notably, the model ranking reverses between domains: BERT dominates on headlines while Word2Vec leads on tweets, indicating simpler embeddings may be more robust to domain shift.

Figure 5: Cross-domain performance comparison showing F1-scores for each model on both datasets.

4 Discussion

The experimental results reveal several important insights about embedding approaches for sarcasm detection. BERT's superior performance on the headlines dataset aligns with expectations, as contextual embeddings can capture the nuanced semantic relationships and incongruities characteristic of sarcastic headlines. The BERT + BiLSTM architecture proves effective for this formal text domain. However, BERT's advantage diminishes on the tweets dataset, where the simpler Word2Vec model achieves comparable or slightly better F1-scores. This finding aligns with the concerns raised by Jang and Frassinelli [2024] regarding the generalizability of sarcasm detection models across domains.

The precision-recall trade-off observed across datasets warrants careful consideration. All models exhibit high recall but substantially lower precision on tweets, indicating a tendency to over-classify content as sarcastic. This behavior likely stems from the noisy, ambiguous nature of social media text where informal language patterns may superficially resemble sarcastic expressions. The class imbalance in the tweets dataset (25.1% sarcastic) compared to headlines (43.9%) may also contribute to this bias.

Several challenges emerged during model development and experimentation. The most significant challenge was the computational burden of training on large datasets. The original tweets dataset contained over 80,000 samples, which we reduced to 20,000 to make training feasible within reasonable time constraints. Even with this reduction, training took several days on local hardware. We attempted to leverage Google Colab for GPU acceleration, but encountered compatibility and session timeout issues that prevented successful execution. Consequently, the Word2Vec embedding training phase consumed more time than the

subsequent BiLSTM classifier training, as the skip-gram with negative sampling algorithm requires iterating through millions of word pairs across multiple epochs. Training Word2Vec from scratch also required careful hyperparameter tuning to achieve competitive performance, with the vocabulary size and embedding dimension significantly impacting downstream classification accuracy. The domain shift between datasets presented another significant challenge, as evidenced by the consistent performance degradation on tweets, supporting the multi-task learning motivations discussed by Chauhan et al. [2020], Tay et al. [2020], and the incongruity modeling approaches of Pan et al. [2020].

Our study has several limitations that should be acknowledged. The evaluation is limited to two datasets from specific domains, and generalization to other sarcasm detection contexts remains to be validated. We focused on binary sarcasm classification and did not explore fine-grained sarcasm types or intensity levels. The BERT experiments used fewer training epochs due to computational constraints, potentially underestimating its full capacity. The reduced dataset size for tweets may also limit the model's ability to capture the full diversity of sarcastic expressions in social media text.

Future work should explore several promising directions. First, we plan to evaluate standalone BERT (without the BiLSTM layer) and compare its performance against our current BERT + BiLSTM architecture to determine whether the additional layers provide meaningful benefits. Second, we intend to compare Word2Vec word-level embeddings with sentence and paragraph embedding models such as Doc2Vec and Sentence-BERT to assess whether capturing longer contextual dependencies improves sarcasm detection. Additionally, multi-task learning approaches incorporating sentiment and emotion as auxiliary tasks, domain adaptation techniques for cross-domain robustness, and integration of external knowledge sources such as commonsense knowledge graphs could help better capture the incongruity underlying sarcastic expressions.

5 Conclusion

This study compared Word2Vec, GloVe, and BERT embeddings combined with BiLSTM classifiers for sarcasm detection across two domains. Three key findings emerged: (1) BERT achieves superior performance on formal headlines (F1=0.900) but loses its advantage on informal tweets where Word2Vec leads (F1=0.776), (2) all models exhibit a precision-recall trade-off on tweets, sacrificing precision for near-perfect recall due to over-prediction bias, and (3) the domain gap challenge persists across all embedding approaches, with performance drops ranging from 0.078 to 0.128 F1-score. These results suggest that model complexity does not guarantee cross-domain generalization, and simpler embeddings trained on domain-specific data can match or outperform sophisticated contextual models on informal text. Future work will evaluate standalone BERT against BERT + BiLSTM and explore sentence-level embeddings for improved sarcasm detection.

References

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.401/.

Hyewon Jang and Diego Frassinelli. Generalizable sarcasm detection is just around the corner, of course! In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2024)*. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.naacl-long.238.pdf.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.findings-emnlp.124.pdf.

Edoardo Savini and Cornelia Caragea. A multi-task learning approach to sarcasm detection (student abstract). In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*. Association for the Advancement of Artificial Intelligence, 2020. URL https://cdn.aaai.org/ojs/7226/7226-13-10455-1-10-20200526.pdf.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. A multi-task learning framework for sarcasm detection with sentiment and emotion. In *Proceedings of the 2020 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 4408–4417. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.emnlp-main.703/.