www.editada.org

*International Journal of Combinatorial Optimization Problems and Informatics, 17(2), 2026, 378-390. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2026.v17i2.853*

# A Comparative Study of BERT-Based Models for Sarcasm Detection in Social Media Texts

*Rafael Jiménez Castro, J. Patricia Sánchez-Solís, Vicente García Jiménez, Gilberto Rivera Zárate, Rogelio Florencia Juárez*

Universidad Autónoma de Ciudad Juárez. Departamento de Ingeniería Eléctrica y Computación
E-mails: rafael.jimenez.cstr@gmail.com; julia.sanchez@uacj.mx; vicente.jimenez@uacj.mx; gilberto.rivera@uacj.mx; rogelio.florencia@uacj.mx

**Abstract.** Social media has transformed communication, facilitating the rapid exchange of emotions and ideas between users. This shift has created the necessity for the development of sentiment analysis tools, enabling businesses to gain insights into audience reactions. However, detecting sentiment remains a challenging task due to the presence of informal language, abbreviations, and, notably, sarcasm, which can modify the intended message. Sarcasm, often conveyed through ironic or contrary statements, is particularly difficult to identify in text as it lacks the non-verbal cues typically present in face-to-face communication. Recent advancements in deep learning, particularly the emergence of BERT (Bidirectional Encoder Representations from Transformers) based models, have significantly enhanced sarcasm detection by capturing nuanced contextual meanings of words. This paper compares several BERT-based models—BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa—to assess their effectiveness in sarcasm detection on iSarcasmEval and Sarcasm Corpus V2 datasets. Key performance metrics, including accuracy, computational efficiency, and the ability to capture complex contextual relationships, are analyzed to identify the most suitable model for sarcasm detection tasks. DeBERTa achieved the best performance on both datasets in this challenging task.

**Keywords:** Natural Language Processing, Transformers, BERT, Sarcasm Detection.

Article Info
*Received February 2, 2025*
*Accepted Dec 11, 2025*

## 1 Introduction

Social media and communication platforms have revolutionized how individuals interact with each other, facilitating the rapid and accessible exchange of ideas and emotions (Pandey, et al., 2017). Users now express their opinions on social networks through short text messages or comments, generating a large amount of information online. This shift has led to the emergence of sentiment analysis, an area of natural language processing. Sentiment analysis enables a better understanding of people's emotions and opinions on a variety of topics, which can be used by businesses and organizations for strategic decision-making. However, identifying sentiments in social media texts faces several challenges, mainly due to informal language and abbreviations, sarcasm, which can significantly distort the intended message (Bouazizi & Otsuki, 2016). Sarcasm is frequently conveyed through ironic or mocking remarks, obscuring the emotional tone of a message (Lunando & Purwarianti, 2013; Kenneth, et al., 2024). Detecting sarcasm is crucial for accurate sentiment analysis. However, it is a difficult task due to its subjective, implicit, and context-dependent nature.

Over the years, various approaches have emerged to address this challenge. Initially, linguistic studies focused on the complexities of language (Rajadesingan, et al., 2015), but more recently, technological advancements have led to the development of machine learning models (Wang, et al., 2015), neural networks (Majumder, et al., 2019), and deep learning techniques. One of the key difficulties in sarcasm detection lies in its dependence on contextual information (Kumar & Garg, 2019).

The Bidirectional Encoder Representations from Transformers (BERT) model (Devlin, et al., 2019) has led to promising advances in sarcasm and irony detection. BERT's bi-directional architecture allows for semantically contextualizing words within a sentence, which has helped improve sarcasm detection, increasing the accuracy of sentiment analysis in written texts (Jihang & Wanli, 2019).

The original BERT model, known as BERT Base, was primarily designed for general-domain text analysis. Subsequently, several BERT-based models, such as RoBERTa, DistilBERT, and DeBERTa, were developed. Some of these models have proven to be effective in sarcasm detection.

In this paper, several BERT-based models were implemented to address the task of detecting sarcasm in social media text. Two distinct datasets were used to train these models: the iSarcasmEval dataset and the Sarcasm Corpus V2 dataset. These datasets were selected for their relevance to detecting sarcasm in social media and other textual contexts. The main contribution of this work is to evaluate the performance of the most commonly used BERT-based models and identify their limitations, identifying the most effective model for sarcasm detection considering the unique challenges of the task (Joshi, et al., 2016).

The paper is organized as follows: Section 2 provides the background outlining the relevant literature and theoretical foundations that support the research. Section 3 details the experimental design, describing the methodology and experimental setup employed for the study. Section 4 discusses the results, highlighting key findings and their implications. Lastly, Section 5 concludes with a summary of the findings and suggestions for future research.

## 2 Background

### 2.1 Sentiment Analysis

The concept of sentiment analysis can be said to be recent as the first related case is from 2013 when Brun and Hagege (Brun & Hagege, 2013) noticed the growing sources of online information such as review sites and personal blogs. To better understand people's opinions and subjective feedback they developed an automatic system for extracting emotions and sentiments. While building this system, they discovered that within the expressed opinions and emotions, there was valuable customer feedback that often went unnoticed in traditional sentiment analysis. Capturing such insights could help businesses enhance their products by adding requested features or addressing unmet customer needs.

Social media users frequently express frustration, irony, or criticism indirectly, using comments that appear positive or neutral but communicate the opposite. This type of communication is known as sarcasm. Detecting sarcasm in social media text is especially complex due to the lack of non-verbal cues, such as tone of voice or facial expressions, which typically help interpret this humor in face-to-face communication. Moreover, sarcasm in social media is influenced by cultural and social context, making it even more challenging for machine learning models.

In literature, sarcasm has been defined as: an inconsistency between the text and the context (Wilson, 2006), the difference between the literal proposition and the intended proposition (Ivanko & Pexman, 2003), and as a form of indirect denial where there is no explicit denial (Giora, 1995).

Misinterpreting sarcasm can lead to misunderstandings, highlighting the need to develop more sophisticated models that understand not only the text but also the context and intentions behind online posts.

### 2.2 BERT Models

BERT (Bidirectional Encoder Representations from Transformers) is a language model based on the transformer architecture developed by Google in 2018. It can analyze the context of a word by considering both the words before and after it, which allows it better to understand the meaning of sentences by using a technique called bidirectional attention (Devlin, et al., 2019). BERT is effective at detecting sarcasm because of its attention mechanism which allows it to determine if there’s a mismatch between the literal meaning of a sentence and the context it’s used in. For example, in the following sentence:

"Fantastic! A flat tire."

At first glance, the word "Fantastic" may suggest that the sentence is positive but "a flat tire" denotes a frustrating situation for most people. BERT analyzes the full context of the sentence, looking at the words simultaneously, understanding how they relate to each other to capture the overall meaning of the phrase. This mismatch between words and context is a big clue that the sentence contains sarcasm (Wilson, 2006).

The models compared in this work are shown in Table 1 (Qiu, et al., 2020; He, et al., 2021), along with their respective developers and training data sizes.

**Table 1**. Models used, their developer and the training sized used for their base model.

| Model | Developer | Parameters | Advantages |
|---|---|---|---|
| BERT | Google AI | 110M (Base), 340M (Large) | - Pre-training on a large corpus<br>- Good performance on specific tasks with fine-tuning |
| RoBERTa | Facebook AI Research | 125M (Base), 355M (Large) | - Better performance than BERT on many tasks<br>- Focuses on large-scale data and more robust pre-training |
| ALBERT | Google & Toyota Technological Institute at Chicago | 12M (Base), 18M (Large) | - More efficient with fewer parameters<br>- Focuses on reducing model size while retaining performance on tasks |
| DistilBERT | Hugging Face | 66M | - Faster and more efficient than BERT<br>- Focuses on reducing size and inference time with minimal loss in accuracy |
| DeBERTa | Microsoft Research | 140M (Base), 400M (Large) | - Improved understanding of token relationships and position<br>- Outperforms BERT and RoBERTa in many NLP tasks due to more advanced attention mechanisms and model architecture |

To date, various models based on BERT have been developed to improve its performance. The BERT-based models used in our study are as follows:

- BERT, as the original BERT base model, serves as the benchmark that all other models aim to outperform. This model is typically used for fine-tuning with additional contextual cues as parameters (Zhuang, et al., 2021).
- RoBERTa, initially described in the paper by Liu (Liu, et al., 2019), is characterized by its much larger training data, which includes books, Wikipedia, and general web text. Notably, it introduced dynamic masking of words, making training more effective as the model learns to predict masked words. This model has been used in studies by Hercog (Hercog, et al., 2022) and Shu (Shu, 2024).
- ALBERT is a lighter version of BERT. While in BERT each transformer layer has its own set of unique weights, increasing the number of parameters as the layers are stacked, ALBERT reuses the same weights across all layers, reducing the model's size without compromising its ability to learn complex patterns (Lan, et al., 2020).
- DistilBERT is a lightweight version of BERT created by Sanh (Sanh et al., 2019). It uses a teacher-student machine learning framework, where a smaller, more efficient model (the student) learns from a larger, more complex model (the teacher) (Hinton et al., 2015). In this case, the student model (DistilBERT) learns the behavior of BERT during training, using only 40% of the parameters of the original BERT model. It has been used to detect sarcasm by utilizing CoMet sequences to capture the contrast between intent and action of the subject (Basu Roy Chowdhury & Chaturvedi, 2021) and by recognizing four types of humor (self-enhancing, self-deprecating, affiliative, and aggressive) (Kenneth, et al., 2024).

- DeBERTa was first introduced by He (He, et al., 2021) as a BERT model with disentangled attention, which separates the attention for a word and its position within a sentence. By doing so, it contextualizes information for each word more accurately. It has been used to detect sarcasm in multilingual settings (Han, et al., 2022).

## 3 Experimental Design

In this section, we explain how the comparison tests were conducted. The experimental design of this study was carefully crafted to ensure a fair and balanced comparison of various BERT-based models for sarcasm detection in social media texts. This approach was designed to make the results reproducible and establish a reliable foundation for interpreting the findings, particularly in the context of sarcasm detection in social media text. Section 3.1 describes the features of the datasets used in the study. Section 3.2 provides details on the hardware setup required to run the experiments. Lastly, Section 3.3 presents the validation process, and the metrics used.

### Datasets

This study utilized two widely recognized datasets for sarcasm detection to evaluate the performance of BERT-based models: the *iSarcasmEval* and *Sarcasm Corpus V2* datasets. The *iSarcasmEval*[1] dataset, derived from Task 6 at SemEval 2022, is a new collection where sarcasm labels are provided by the authors. Although the dataset contains both Arabic and English text, only the English portion was used for this comparison, it has been used in works by several authors like Han, et al (Han, et al., 2022), Grover and Banati (Grover & Banati, 2022), Krishnan, et al. (Krishnan, et al., 2022), Du, et al. (Du, et al., 2022), and Abu Farha, et al. (Abu Farha, et al., 2022). This dataset offers a valuable benchmark for evaluating sarcasm detection in social media contexts. The *Sarcasm Corpus V2*[2], created by the University of California, Santa Cruz, is a dataset that includes three categories of sarcasm: *general sarcasm*, *hyperbole*, and *rhetorical questions*. However, only the general sarcasm category was used. This dataset has been used in the works by Ghosh, et al. (Ghosh, et al., 2018). Jang and Frassinelli (Jang & Frassinelli, 2024), and Najafabadi, et al. (Najafabadi, et al., 2024). As shown in Table 2, the iSarcasmEval dataset contains 3,469 instances, with 2,402 instances labeled as 'No sarcasm' (class 0) and 1,067 instances labeled as 'Sarcasm' (class 1). Equally, the Sarcasm Corpus V2 dataset has 6,520 instances, evenly distributed between the two classes.

**Table 2**. Class distribution of both datasets used.

| Dataset | Instances | No sarcasm (class 0) | Sarcasm (class 1) |
|---|---|---|---|
| iSarcasmEval | 3,469 | 2,402 | 1,067 |
| Sarcasm Corpus V2 | 6,520 | 3,260 | 3,260 |

### Validation process

The Hold-out validation technique was used to evaluate the performance of the models. This technique is commonly used to assess a model's performance (Joshi, et al., 2016; Rajadesingan, et al., 2015). This method consists of splitting the dataset into two parts: generally, 80% for training all models and 20% for testing the models. Thus, from the iSarcasmEval dataset, 2,775 instances were used for training and 694 for testing. From the Sarcasm Corpus V2 dataset, 5,216 instances were used for training and 1,304 for testing.

The evaluation of the classification models is carried out based on popular metrics such as precision, accuracy, sensitivity, and F1 score (D'Andrea, et al., 2019). These metrics are defined in Table 3, where: i) true positives (TP) are the number of positive instances correctly classified (instances correctly classified as Sarcasm); ii) true negatives (TN) are the number of negative instances correctly classified (instances correctly classified as No sarcasm); iii) false positives (FP) are the number of negative instances incorrectly classified as positive (instances incorrectly classified as Sarcasm); iv) false negatives (FN) are the number of positive instances incorrectly classified as negative (instances incorrectly classified as No sarcasm).

[1] https://sites.google.com/view/semeval2022-isarcasmeval

[2] https://nlds.soe.ucsc.edu/sarcasm2

**Table 3**. Definition of evaluation metrics.

| Metric | Formula |
|---|---|
| Precision | $Precision = \frac{TP}{TP+FP}$ |
| Accuracy | $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$ |
| Recall | $Recall = \frac{TP}{TP+FN}$ |
| F1-score | $F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall}$ |

### Hardware testbench

BERT delivers outstanding performance on NLP tasks but comes with high resource demands, including memory, computational power, and storage. The resource consumption level depends on the size of the BERT model and the training dataset. To run BERT efficiently and significantly reduce training time, using GPUs or TPUs is essential. For this reason, Google Colab was chosen as the testing environment. It offers an online runtime that supports free TPU usage making it ideal for experimentation. Table 4 provides the hardware specifications of Google Colab's free-tier runtime setup.

**Table 4.** Hardware specifications

| Category | Specification |
|---|---|
| CPU | Intel Xeon 2 Cores |
| GPU | NVIDIA Tesla T4 |
| RAM | 12.6GB |
| vRAM | 16GB |
| Storage | 50GB |

## 4 Results

This section describes the evaluation process performed. The five models were trained on the two datasets, iSarcasmEval and Sarcasm Corpus V2. The models were trained using the default values of the hyperparameters in 1 epoch and 2 epochs. Section 4.1 shows the results obtained on the iSarcasmEval dataset. Section 4.2 shows the results obtained on the Sarcasm Corpus V2 dataset. Lastly, Section 4.3 presents a discussion about the performance of the models.

The models were trained using 1 epoch and 2 epochs. This decision was based on two factors: first, BERT is computationally expensive to run, and second, BERT is already pre-trained in massive amounts of text. Furthermore, as noted by Devlin (Devlin, et al., 2019), 1 epoch or 2 epochs are enough to train BERT in small datasets, since more epochs could generate overfitting.

### iSarcasmEval dataset results

The evaluation results using the iSarcasmEval dataset in 1 epoch are summarized in Table 5 (1 epoch), highlighting the computational resources used and the resulting metrics. This table includes the columns 'RAM' and 'GPU RAM' indicating the amount of system memory and GPU memory consumed, respectively. 'Storage' is the gigabytes of storage data used to train the model. 'Time' shows the total number of minutes required to complete the training process. 'Accuracy' and 'F1 scores' show information about the model's performance in class 0 and class 1.

In 1 epoch (as shown in Table 5), BERT consumed the least amount of RAM, while ALBERT utilized the least amount of GPU RAM and storage. DistilBERT was the most time-efficient model and was also the model that achieved the highest accuracy at 71.76%. However, BERT was the model that achieved the best F1 score for class 1, that is, the best model in sarcasm detection. The confusion matrices for each model are provided in Table 6 for the BERT, RoBERTa and ALBERT, DistilBERT, and DeBERTa models. Based on these results, BERT was the best model at detecting sarcasm, correctly identifying 81 out of 206 instances.

**Table 5.** Performance of the models on the iSarcasmEval dataset in 1 epoch

|  | Computational Resources |  |  | iSarcasmEval - 1 epoch |  |  |  |
|---|---|---|---|---|---|---|---|
| Model | RAM | GPU RAM | Storage | Time | Accuracy | F1 Class 0 | F1 Class 1 |
| BERT | **2.4GB** | 3.2GB | 32.7GB | 1:55 | 65.99% | 76.16% | **40.70%** |
| RoBERTa | 2.6GB | 3.5GB | 32.8GB | 2:02 | 70.32% | 82.57% | 0% |
| ALBERT | 2.5GB | **2.7GB** | **32.4GB** | 2:02 | 70.32% | 82.57% | 0% |
| DistilBERT | 2.8GB | 2.9GB | 32.6GB | **0:41** | **71.76%** | **82.99%** | 16.95% |
| DeBERTa | 2.6GB | 6.2GB | 32.9GB | 2:57 | 70.46% | 82.61% | 1.91% |

**Table 6.** Confusion Matrices for BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa Models using the iSarcasmEval dataset and 1 epoch.

|  |  | Actual Values |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | BERT |  | RoBERTa |  | ALBERT |  | DistilBERT |  | DeBERTa |  |
|  |  | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 |
| Predicted Values | Class 0 | 377 | 111 | 488 | 0 | 488 | 0 | 478 | 10 | 487 | 1 |
|  | Class 1 | 125 | 81 | 206 | 0 | 206 | 0 | 186 | 20 | 204 | 2 |

In 2 epochs (as shown in Table 7), the models that required less computational resources remained the same as those described in Table 5. DeBERTa emerged as the best performing model, achieving the highest accuracy of 76.37% and the best F1 scores in class 0 and class 1, with 84.56% and 49.69%, respectively. It is worth noting that, although all models showed an improvement in F1 score in class 1 compared to 1 epoch, RoBERTa obtained a smaller increase from 0% to 0.97%. On the other hand, although BERT also improved its performance, its improvement was not as noticeable compared to the ALBERT, DistilBERT, and DeBERTa models. The confusion matrices for each model are provided in Table 8 for the BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa models. Based on these results, BERT detected the highest number of sarcastic instances with 85, it also incorrectly classified the highest number of non-sarcastic instances with 81, which affected the F1 score in class 1.

**Table 7.** Performance of the models on the iSarcasmEval dataset in 2 epoch

|  | Computational Resources |  |  | iSarcasmEval - 2 epoch |  |  |  |
|---|---|---|---|---|---|---|---|
| Model | RAM | GPU RAM | Storage | Time | Accuracy | F1 Class 0 | F1 Class 1 |
| BERT | **2.2GB** | 3.2GB | 32.7GB | 7:57 | 70.89% | 80.12% | 45.7% |
| RoBERTa | 2.7GB | 3.5GB | 32.8GB | 3:57 | 70.46% | 82.64% | 0.97% |
| ALBERT | 2.4GB | **2.8GB** | **32.4GB** | 4:09 | 73.49% | 83.66% | 29.77% |
| DistilBERT | 2.9GB | 2.9GB | 32.6GB | **1:21** | 72.91% | 82.36% | 41.61% |
| DeBERTa | 2.6GB | 6.4GB | 32.9GB | 5:53 | **76.37%** | **84.56%** | **49.69%** |

**Table 8.** Confusion Matrices for BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa Models using the iSarcasmEval dataset and 2 epoch

|  |  | Actual Values |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | BERT |  | RoBERTa |  | ALBERT |  | DistilBERT |  | DeBERTa |  |
|  |  | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 |
| Predicted Values | Class 0 | 407 | 81 | 488 | 0 | 471 | 17 | 439 | 49 | 449 | 39 |
|  | Class 1 | 121 | 85 | 205 | 1 | 167 | 39 | 139 | 67 | 125 | 81 |

## Sarcasm Corpus V2 dataset results

The computational resource usage for the Sarcasm Corpus V2 dataset in 1 epoch, shown in Table 9, remains consistent with those shown in Table 5. BERT and RoBERTa used the least amount of RAM, while ALBERT required the least amount of RAM and GPU storage and DistilBERT remained the fastest model. ALBERT achieved the highest accuracy of 86.63% and the highest F1 score of 87.05% and 86.18% in class 0 and class 1, respectively.

In 1 epoch (Table 9), ALBERT achieved the best results with an accuracy of 86.63% and an F1 score of 87.05% and 86.18% in class 0 and class 1, respectively. While all models, except for RoBERTa, showed higher accuracy scores compared to those obtained using the iSarcasmEval dataset and 1 epoch in Table 5. The confusion matrices for each model are provided in Table 10 for the BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa models. Based on these results, BERT detected the highest number of sarcastic instances with 85, it also incorrectly classified the highest number of non-sarcastic instances with 81, which affected the F1 score in class 1. All models achieved a correct classification rate for sarcasm above 70%. RoBERTa stood out by correctly identifying the highest number of instances, with 592 out of 633 classified accurately. However, it also had the highest number of non-sarcastic instances misclassified as sarcastic. In contrast, ALBERT performed the best in correctly classifying non-sarcastic instances and ranked second in accurately identifying sarcastic sentences.

**Table 9.** Performance of the models on the Sarcasm Corpus V2 dataset in 1 epoch

| | Computational Resources | | | Sarcasm Corpus V2 - 1 epoch | | | |
|---|---|---|---|---|---|---|---|
| Model | RAM | GPU RAM | Storage | Time | Accuracy | F1 Class 0 | F1 Class 1 |
| BERT | **2.5GB** | 4.1GB | 34.0GB | 3:59 | 78.45% | 80.39% | 76.09% |
| RoBERTa | **2.5GB** | 3.5GB | 32.8GB | 3:53 | 62.88% | 48.51% | 70.98% |
| ALBERT | 2.6GB | **2.9GB** | **32.4GB** | 3:56 | **86.63%** | **87.05%** | **86.18%** |
| DistilBERT | 2.9GB | 3.2GB | 32.6GB | **1:17** | 76.38% | 79.47% | 72.20% |
| DeBERTa | 2.7GB | 7.2GB | 32.9GB | 5:36 | 80.67% | 81.66% | 79.58% |

**Table 10.** Confusion Matrices for BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa Models using the Sarcasm Corpus V2 dataset and 1 epoch

| | | Actual Values | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BERT | | RoBERTa | | ALBERT | | DistilBERT | | DeBERTa | |
| | | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 |
| Predicted Values | Class 0 | 576 | 95 | 228 | 443 | 541 | 30 | 596 | 75 | 561 | 110 |
| | Class 1 | 186 | 447 | 41 | 592 | 131 | 502 | 233 | 400 | 142 | 491 |

In 2 epochs (Table 11), the models showed improvements in F1 score in class 1 compared to the results shown in Table 16, the only exceptions being ALBERT and DeBERTa. DeBERTa had the best accuracy with 79.91% and the best F1 score in class 0 with 81.68%. DistilBERT had the best F1 score in class 1 with 77.98%. The confusion matrices for each model are provided in Table 12 for the BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa models. Based on the results, although BERT had the highest number of correctly classified sarcastic examples, it also misclassified the highest number of non-sarcastic examples as sarcastic. BERT showed the least amount of bias for this test, achieving 77.13% and 77.92% in F1 score in class 0 and class 1, respectively.

**Table 11**. Performance of the models on the Sarcasm Corpus V2 dataset in 2 epoch

| | Computational Resources | | | Sarcasm Corpus V2 - 2 epoch | | | |
|---|---|---|---|---|---|---|---|
| Model | RAM | GPU RAM | Storage | Time | Accuracy | F1 Class 0 | F1 Class 1 |
| BERT | **2.4GB** | 4.1GB | 34.0GB | 7:59 | 77.53% | 77.13% | 77.92% |
| RoBERTa | 2.6GB | 3.5GB | 32.8GB | 7:45 | 74.16% | 76.48% | 71.32% |
| ALBERT | 2.6GB | **2.9GB** | **32.4GB** | 7:52 | 79.14% | 81.29% | 76.43% |
| DistilBERT | 2.9GB | 3.2GB | 32.6GB | **2:34** | 78.91% | 79.76% | **77.98%** |
| DeBERTa | 2.7GB | 7.2GB | 32.9GB | 11:11 | **79.91%** | **81.68%** | 77.76% |

**Table 12.** Confusion Matrices for BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa Models using the Sarcasm Corpus V2 dataset and 2 epoch

| | | Actual Values | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BERT | | RoBERTa | | ALBERT | | DistilBERT | | DeBERTa | |
| | | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 |
| Predicted Values | Class 0 | 494 | 177 | 548 | 123 | 591 | 80 | 542 | 129 | 584 | 87 |
| | Class 1 | 116 | 517 | 214 | 419 | 192 | 441 | 146 | 487 | 175 | 458 |

## Discussion

To analyze the performance of the models, the t-SNE (t-Distributed Stochastic Neighbor Embedding) dimensionality reduction technique, proposed by van der Maaten and Hinton (van der Maaten & Hinton, 2008), was used. This technique was used to reduce the high dimensionality of the embeddings to only 2 dimensions, facilitating their interpretation and visual analysis. The reduced dimensions generated by t-SNE have no direct meaning in terms of the original characteristics of the data but rather represent a projection of the data into a lower-dimensional space where local relationships are preserved.

Considering the recommendations and observations of van der Maaten and Hinton (van der Maaten & Hinton, 2008; Kobak, et al., 2020) on how to analyze and interpret t-SNE results, the following is presented:

- Clusters: Clusters are distinct and well-defined, or they may overlap. Clear separation suggests that there are distinct groupings in the data, while overlap may indicate similarities between features.
- Outliers: Points that are far from a cluster represent anomalies or noisy data.
- Density: Clusters with a high degree of density suggest similarity between their points; on the other hand, a scattered cluster may indicate more variability within its points.
- Distribution: Even distribution of points within clusters or concentrations of points in certain regions of the cluster may indicate subgroups.

Figure 1 presents the t-SNE diagrams of each of the models in the iSarcasmEval dataset. From these diagrams, it can be seen that none of the models generate distinct categorical clusters from the dataset. This, combined with the class imbalance, makes it difficult for the models to accurately detect sarcasm without relying on additional contextual information. In the diagrams, most of the models produce clusters where data points from different features overlap. However, the BERT model stands out slightly, as its classes tend to cluster better. This pattern could explain why BERT was able to correctly identify more instances of sarcasm compared to the other models.

Figure 2 shows the t-SNE plots of the models on the Sarcasm Corpus V2 dataset. Here, the diagrams reveal that all models generate large, dense clusters with overlapping categorical data points. Despite the overlap, the classes are concentrated in different regions of the clusters, allowing for some degree of separation between the categories. This separation appears to be a key factor in allowing the models to achieve better sarcasm detection performance compared to their results on the iSarcasmEval dataset.

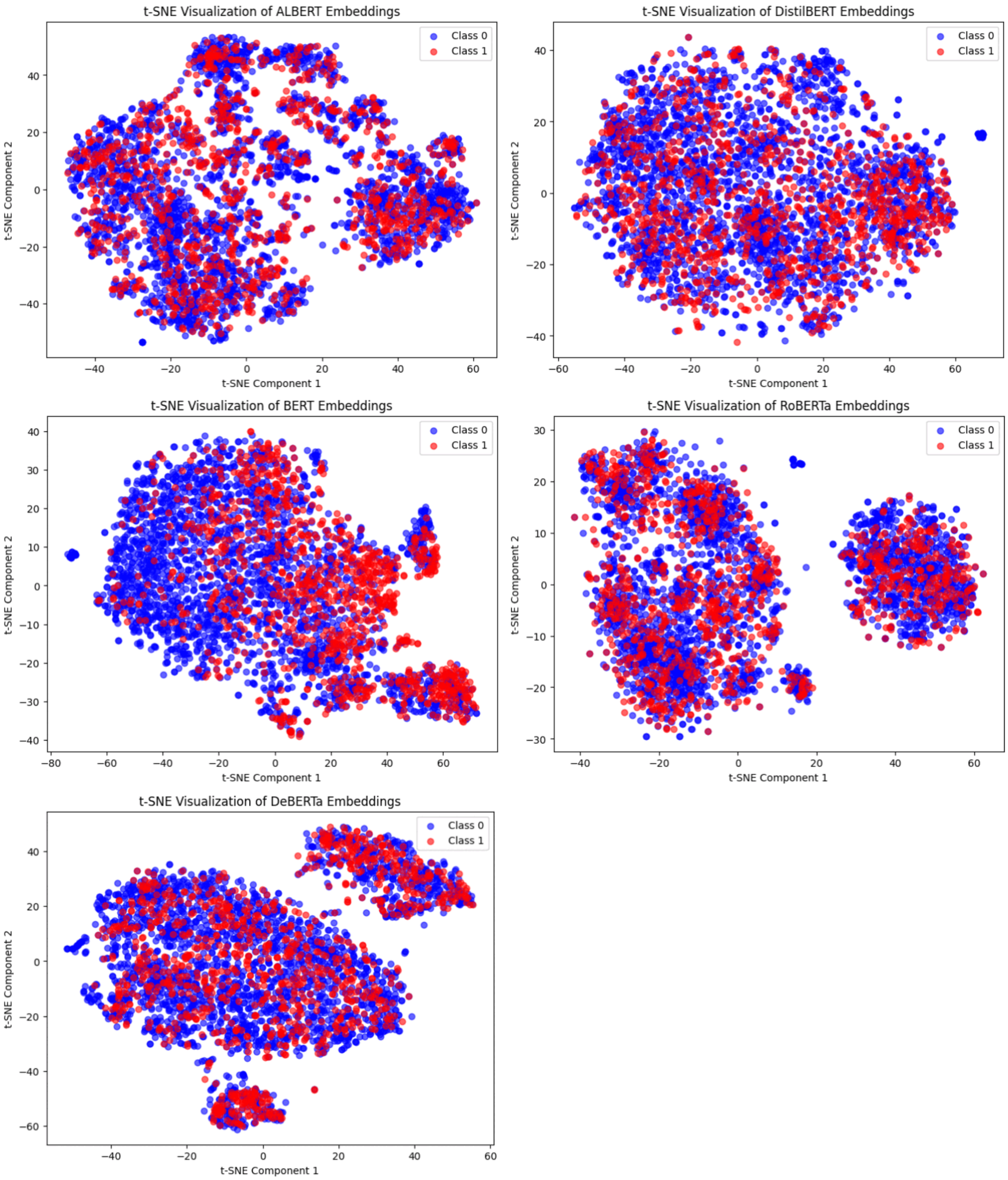

Fig. 1 t-SNE plot diagrams of the models in the iSarcasmEval dataset.

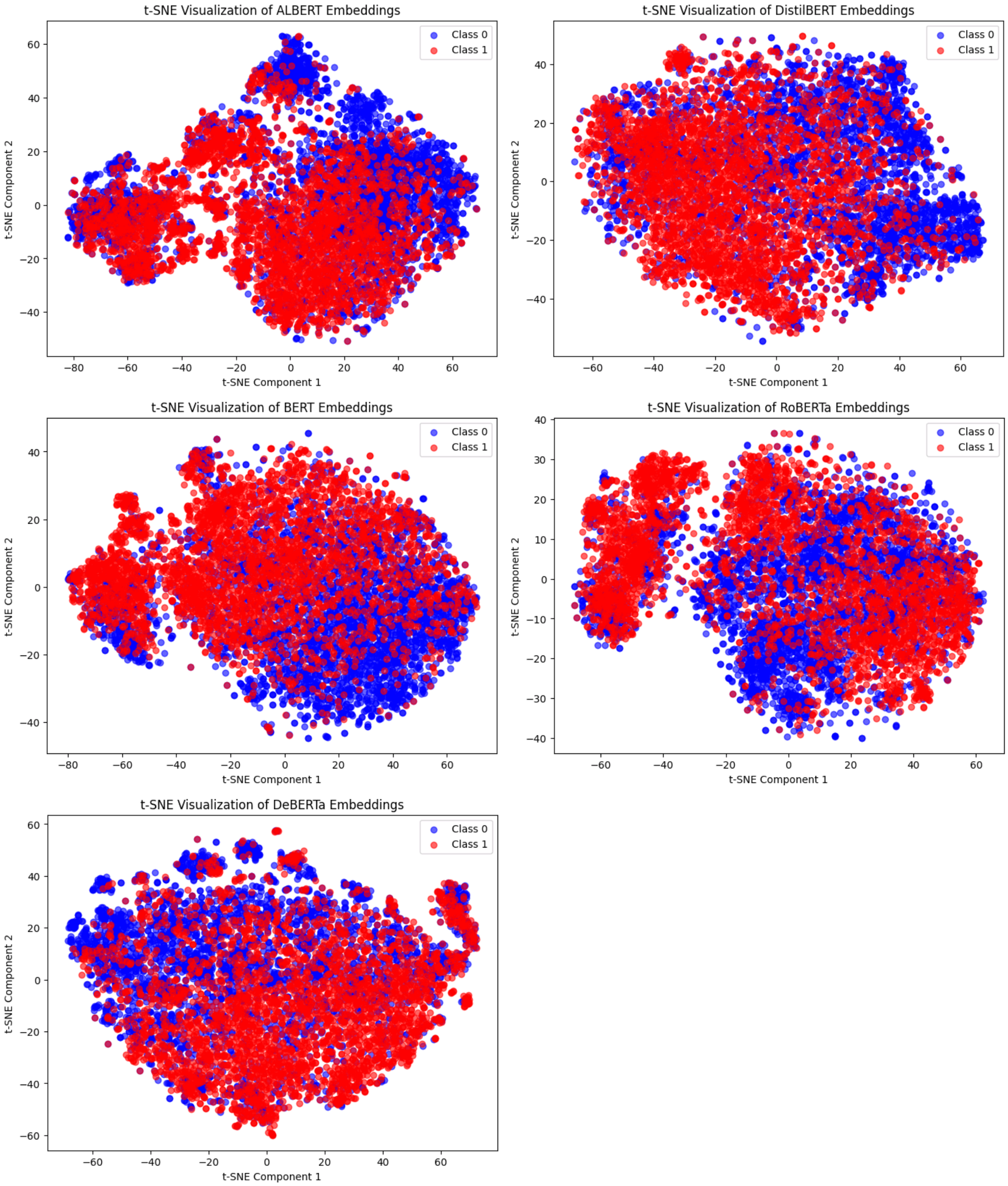

Fig. 2 t-SNE plot diagrams of the models in the Sarcasm Corpus V2.

## 5 Conclusion

This work compared the performance of different BERT-based models for sarcasm detection in social media text. The models used were BERT, RoBERTa, ALBERT, DistilBERT, and DeBERTa. These models were trained and evaluated on the iSarcasmEval and Sarcasm Corpus V2 datasets.

Experiments performed on the iSarcasmEval and Sarcasm Corpus V2 datasets provided key insights for sarcasm detection. Although BERT performed the best on the iSarcasmEval dataset trained in 1 epoch achieving an F1 score of 40.70% in class 1, its ability to detect sarcasm was still limited. On 2 epochs, DeBERTa emerged as the best model with an accuracy of 76.37% and an F1 scores of 84.56% and 49.69% in class 0 and class 1, respectively. Although it was the best model, it still struggled to correctly classify sarcastic comments.

On the Sarcasm Corpus V2 dataset, ALBERT outperformed the other models on 1 epoch, achieving the highest accuracy of 86.63% and the highest F1 score of 87.05% and 86.18% in class 0 and class 1, respectively. The confusion matrices show that sarcasm detection improved compared to the iSarcasmEval dataset with fewer misclassifications. The best performance on 2 epochs was obtained by DistilBERT achieving an F1 score of 77.98% in class 1; however, there were also a large number of non-sarcastic instances misclassified as sarcastic.

We used t-SNE to reduce the dimensionality of the model's embeddings and visualize how the models distinguish between sarcastic and non-sarcastic examples. The biggest takeaway here is the difference between the datasets themselves. The iSarcasmEval dataset, for example, showed more separation between clusters but with an even distribution of data points within them which created an overlap between classes. This made it harder for the models to accurately detect sarcasm. In contrast, the Sarcasm Corpus V2 dataset produced large, dense clusters with clearer separation between classes. This type of structure made it easier for the models to identify sarcasm, as the data points were more concentrated in distinct regions, providing a stronger basis for differentiation.

Overall, while most models showed they produced good, or the best, results in some tests, DeBERTa achieved better performance metrics most of the time, but the lack of clear separation in its embeddings may explain its difficulty in accurately detecting sarcasm. To fine-tune these models, particularly DeBERTa, additional features such as sentiment scores or custom embeddings should be incorporated to contextualize each comment in these datasets better and improve detection.

In future work, several strategies could be explored to improve sarcasm detection in social media texts. One of the most important strategies is identifying contextual information that could help enhance the models' performance. Additionally, fine-tuning the models' hyperparameters could yield better results. Finally, training the models on larger and more diverse datasets that include data from various social media platforms could help the models generalize better across different contexts.

## References

Abu Farha, I., Oprea, S. V., Wilson, S., & Magdy, W. (2022). SemEval-2022 Task 6: iSarcasmEval, Intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 802–814). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.111

Basu Roy Chowdhury, S., & Chaturvedi, S. (2021). Does commonsense help in detecting sarcasm? In *Proceedings of the Second Workshop on Insights from Negative Results in NLP* (pp. 9–15). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.insights-1.2

Bouazizi, M., & Otsuki, T. (2016). A pattern-based approach for sarcasm detection on Twitter. *IEEE Access, 4*, 5477–5488. https://doi.org/10.1109/ACCESS.2016.2594194

Brun, C., & Hagège, C. (2013). Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computing Science, 70*, 199–209.

D'Andrea, E., Ducange, P., Bechini, A., & Renda, A. (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications, 116*, 209–226.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

Du, X., Hu, D., Zhi, J., Jiang, L., & Shi, X. (2022). PALI-NLP at SemEval-2022 Task 6: iSarcasmEval- Fine-tuning the pre-trained model for detecting intended sarcasm. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 815–819). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.112

Ghosh, D., Fabbri, A. R., & Muresan, S. (2018). Sarcasm analysis using conversation context. *Computational Linguistics, 44*(4), 755–792. https://doi.org/10.1162/coli_a_00336

Giora, R. (1995). On irony and negation. *Discourse Processes, 19*(2), 239–264. https://doi.org/10.1080/01638539509544916

Grover, V., & Banati, H. (2022). DUCS at SemEval-2022 Task 6: Exploring emojis and sentiments for sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 1005–1011). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.141

Han, Y., Chai, Y., Wang, S., Sun, Y., Huang, H., Chen, G., Xu, Y., & Yang, Y. (2022). X-PuDu at SemEval-2022 Task 6: Multilingual learning for English and Arabic sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 999–1004). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.140

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv*. https://arxiv.org/abs/2006.03654

Hercog, M., Jaroński, P., Kolanowski, J., Mieczyński, P., Wiśniewski, D., & Potoniec, J. (2022). Sarcastic RoBERTa: A RoBERTa-based deep neural network detecting sarcasm on Twitter. In *Big Data Analytics and Knowledge Discovery* (pp. 46–52). Springer. https://doi.org/10.1007/978-3-031-12670-3_4

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*. https://arxiv.org/abs/1503.02531

Ivanko, S. L., & Pexman, P. M. (2003). Context incongruity and irony processing. *Discourse Processes, 35*(3), 241–279. https://doi.org/10.1207/S15326950DP3503_2

Jang, H., & Frassinelli, D. (2024). 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 (Volume 1: Long Papers)* (pp. 4238–4249). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.238

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). *Automatic sarcasm detection: A survey. ACM Computing Surveys, 50*(5), Article 73. https://doi.org/10.1145/3124420

Kenneth, M. O., Khosmood, F., & Edalat, A. (2024). A two-model approach for humour style recognition. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* (pp. 259–274). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.nlp4dh-1.25

Kobak, D., Linderman, G., Steinerberger, S., Kluger, Y., & Berens, P. (2020). Heavy-tailed kernels reveal a finer cluster structure in t-SNE visualisations. In *Machine Learning and Knowledge Discovery in Databases* (pp. 124–139). Springer. https://doi.org/10.1007/978-3-030-46150-8_8

Krishnan, D., Mahibha C, J., & Durairaj, T. (2022). GetSmartMSEC at SemEval-2022 Task 6: Sarcasm detection using contextual word embedding with Gaussian model for irony type identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 827–833). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.114

Kumar, A., & Garg, G. (2019). Empirical study of shallow and deep learning models for sarcasm detection using context in benchmark datasets. *Journal of Ambient Intelligence and Humanized Computing*. Advance online publication. https://doi.org/10.1007/s12652-019-01419-7

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for self-supervised learning of language representations. *arXiv*. https://doi.org/10.48550/arXiv.1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692

Majumder, N., Poria, S., Peng, H., Chhaya, N., Cambria, E., & Gelbukh, A. (2019). Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems, 34*(3), 38–43. https://doi.org/10.1109/MIS.2019.2904691

Mao, J., & Liu, W. (2019). A BERT-based approach for automatic humor detection and scoring. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)* (pp. 197–202). CEUR Workshop Proceedings.

Najafabadi, M. K., Ko, T. Z. C., Chaeikar, S. S., & Shabani, N. (2024). A multi-level embedding framework for decoding sarcasm using context, emotion, and sentiment feature. *Electronics, 13*(22), Article 4429. https://doi.org/10.3390/electronics13224429

Pandey, A., Rajpoot, D., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management, 53*, 764–779. https://doi.org/10.1016/j.ipm.2017.02.004

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences, 63*(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm detection on Twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)* (pp. 97–106). https://doi.org/10.1145/2684822.2685316

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv*. https://doi.org/10.48550/arXiv.1910.01108

Shu, X. (2024). BERT and RoBERTa for sarcasm detection: Optimizing performance through advanced fine-tuning. *Applied and Computational Engineering, 97*(1), 1–11. https://doi.org/10.54254/2755-2721/97/20241354

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(86), 2579–2605.

Wang, Z., Wu, Z., Wang, R., & Ren, Y. (2015). Twitter sarcasm detection exploiting a context-based model. In *Web Information Systems Engineering – WISE 2015* (pp. 77–91). Springer. https://doi.org/10.1007/978-3-319-26190-4_6

Wilson, D. (2006). The pragmatics of verbal irony: Echo or pretence? *Lingua, 116*(10), 1722–1743. https://doi.org/10.1016/j.lingua.2006.05.001

Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics* (pp. 1218–1227). Chinese Information Processing Society of China. https://aclanthology.org/2021.ccl-1.108/