Sarcasm Detection on Reddit: Comparing Pre-trained Models for Sarcasm Recognition and Sentiment Analysis

Matteo Pasotti[1], Alexandre Crivellari[1], Andrea Muscio[1]
[1]MSc Data Science, Università degli Studi Milano-Bicocca

Abstract

This study examines the effectiveness of transformer-based models in detecting sarcasm and analyzing sentiment in Reddit comments. We compare four pre-trained models—BERT, RoBERTa, DistilBERT, and ALBERT—fine-tuned on a balanced dataset of labeled comments. Additionally, we explore the interplay between sarcasm and sentiment using both machine learning and rule-based approaches.

Our analysis shows that all four models perform well in sarcasm detection, with accuracies between 0.77 and 0.79. BERT achieved the highest performance (accuracy: 0.79, F1 score: 0.79), followed closely by DistilBERT. RoBERTa and ALBERT demonstrated comparable precision but lower recall. For sentiment analysis, we employed a RoBERTa-based model and VADER, revealing significant differences in their classification patterns, particularly in handling sarcastic content.

The study highlights the capabilities of transformer architectures in capturing nuanced contextual cues indicative of sarcasm. However, it also underscores persistent challenges in automated sarcasm detection and sentiment analysis, especially in contexts requiring broader cultural knowledge or subtle interpretation.

Summary

1. Introduction

In the digital age, social media platforms like Reddit, Twitter, and Facebook have emerged as key arenas for public discourse, providing users with spaces to express opinions, share experiences, and engage in discussions. Among these platforms, Reddit stands out for its unique structure, characterized by a vast array of subreddits dedicated to specific topics. This diversity in content and community makes Reddit an invaluable resource for analyzing public sentiment and discourse.

However, the richness and informality of social media communication also present significant challenges for natural language processing (NLP) systems, particularly when it comes to detecting sarcasm—a prevalent and nuanced form of expression online. Sarcasm, a type of verbal irony where the intended meaning contrasts with the literal one, is often used for humor, criticism, or emphasis. Its interpretation is highly context-dependent and typically relies on shared cultural knowledge, making it particularly challenging for automated systems to decode accurately.

1.1 The Challenge of Sarcasm in NLP

The rise of large language models (LLMs) has significantly advanced the ability of machines to understand and generate human-like text. Unlike straightforward text classification, sarcasm detection requires a deep understanding of context, cultural references, and the specific dynamics of online communities. For instance, a seemingly positive comment could be a sarcastic critique, and misinterpreting it could lead to substantial errors in sentiment analysis and intent recognition.

On platforms like Reddit, where communication is often brief and informal—featuring slang, abbreviations, and emojis—sarcasm can be particularly difficult for NLP systems to detect. The challenge is compounded by the fact that sarcastic remarks often depend on a shared understanding between the speaker and the audience, something that LLMs, despite being trained on vast datasets, may struggle to grasp without extensive contextual knowledge. This highlights the need for more refined models that can navigate the nuances of human language, particularly in the ever-evolving landscape of social media.

1.2 Objectives of the Study

In light of the challenges posed by sarcasm in NLP, this study is guided by three primary objectives:

Sarcasm Detection: The foremost goal of this study is to develop a model capable of accurately detecting sarcasm in Reddit posts. The study will focus on identifying the subtle cues that signal sarcasm, considering the unique linguistic and cultural contexts of Reddit.

Sentiment Analysis: Beyond sarcasm detection, this study will conduct a detailed sentiment analysis of Reddit comments. This analysis involves categorizing comments as positive, negative, or neutral and examining how these sentiments manifest in both sarcastic and non-sarcastic contexts. The objective is to understand the impact of sarcasm on the perception of sentiment and to explore the challenges it introduces in accurately interpreting user intent.

Models Comparison: To achieve a comprehensive evaluation, the study will compare the performance of different pre-trained models in the tasks of sarcasm detection and sentiment analysis. This comparison will involve assessing the models' accuracy, robustness, and ability to generalize across different subreddits and topics. By analyzing these aspects, the study seeks to identify the strengths and weaknesses of each model, providing insights into their effectiveness and potential areas for enhancement.

2. Domain of Interest

The domain of interest for this study centers around the intricate and evolving dynamics of online communication, particularly on Reddit, where the interplay of language, culture, and community creates a fertile ground for nuanced expressions like sarcasm. This study focuses on Reddit as it offers a unique blend of content and interaction styles across its myriad of subreddits, making it an ideal platform to explore the complexities of sarcasm within various contexts and communities. The ultimate goal is to enhance the ability of NLP models to accurately interpret sarcasm, thereby improving sentiment analysis and understanding of online discourse in this rich and diverse domain.

2.1 The Importance of Context

Context plays a crucial role in understanding sarcasm. On social media, context can be provided by previous interactions in a thread, the relationship between users, or the broader cultural or political environment. Sarcasm detection models must therefore go beyond analyzing isolated statements and instead consider the broader context in which a comment is made. This need for contextual understanding distinguishes sarcasm detection from more straightforward text classification tasks.

For example, a sarcastic comment on Reddit might be easily recognized by users familiar with the ongoing discussion or the subreddit’s culture, but the same comment could be misinterpreted by an NLP model lacking that contextual knowledge. An important piece of Reddit’s culture is, for example, the use of the string “/s”, which is used as a cultural expression for expliciting a (even obvious) sarcastic comment in a conversation.

2.2 Reddit as a Platform for Analysis

Reddit is one of the largest and most popular social media platforms, often described as "the front page of the internet." Founded in 2005, it functions as a vast collection of forums, known as "subreddits," where users can post content, share ideas, and engage in discussions on a wide range of topics. These subreddits are organized around specific interests, ranging from technology and science to hobbies, entertainment, and niche communities. With millions of active users and thousands of subreddits, Reddit provides a diverse and dynamic environment for studying human communication and behavior.

Reddit's structure encourages a unique style of interaction that is different from other social media platforms. Unlike platforms that emphasize short, real-time updates (like Twitter) or visual content (like Instagram), Reddit is primarily text-based and discussion-driven. Users engage in detailed conversations, often posting long-form content and participating in extensive comment threads.

Reddit’s unique environment makes it particularly well-suited for studying sarcasm and other forms of nuanced communication. The platform’s emphasis on text-based interaction, combined with the diverse and often highly engaged user base, results in a significant amount of sarcastic content. Moreover, the varied nature of discussions across subreddits provides a broad context for analyzing how sarcasm is used in different communities and situations.

3. Dataset

Sarcasm detection requires a substantial amount of labeled data to train and evaluate models accurately. The dataset used in this study contains a rich collection of Reddit comments that are labeled to indicate whether they are sarcastic or non-sarcastic. This section details the source of the dataset, its key characteristics, and the preprocessing steps applied to ensure its suitability for model training and analysis.

3.1 Data Source and Collection

The dataset used for this study is sourced from Kaggle, a well-known platform for data science and machine learning competitions. The specific dataset titled "Sarcasm on Reddit" is designed to aid in the detection of sarcasm within Reddit comments. This dataset is composed of comments from various Reddit threads, labeled as either sarcastic or non-sarcastic, making it a valuable resource for training and evaluating models aimed at sarcasm detection.

The dataset was downloaded directly from Kaggle and contains 1.3 million labeled comments. Each entry in the dataset includes the original comment text and a binary label indicating whether the comment is sarcastic (1) or non-sarcastic (0).

Notably, the labeling process utilized a common Reddit convention for sarcasm detection: comments containing the marker "/s"—a textual indicator often used to signal sarcasm—were initially labeled as sarcastic (1). This marker was subsequently removed from the comments to ensure that the dataset reflects natural, unaltered text. Each entry in the dataset includes the original comment text and a binary label indicating sarcasm.

3.2 Dataset Characteristics

The "Sarcasm on Reddit" dataset consists of a wide range of comments extracted from different subreddits, ensuring a diverse representation of topics and tones. This diversity is crucial for training robust sarcasm detection models, as it exposes the model to various linguistic patterns and community norms that may influence the use of sarcasm.

Key characteristics of the dataset include:

- **Volume**: The dataset contains over a million comments.
- **Label Distribution**: The labels are binary, with comments either being classified as sarcastic or non-sarcastic. The distribution of these labels is balanced.
- **Diversity**: Comments come from a broad spectrum of subreddits, covering a wide range of topics, from politics and technology to humor and entertainment. This variety helps the model generalize better across different contexts.

The dataset attributes are explained as follows:

- **label**: This is an integer value (0 or 1) indicating whether the comment is sarcastic (1) or not (0).
- **comment**: A textual field containing the actual comment made by the user.
- **author**: The username of the person who made the comment.
- **subreddit**: The specific subreddit or community where the comment was posted.

- **score**: An integer representing the score of the comment, which is a net vote count (upvotes minus downvotes).
- **ups**: The number of upvotes the comment received.
- **downs**: The number of downvotes the comment received.
- **date**: The date (in YYYY-MM format) when the comment was posted.
- **created_utc**: The exact timestamp (in UTC format) of when the comment was created.
- **parent_comment**: The text of the parent comment to which the current comment is a reply.

3.3 Data Preprocessing and Cleaning

To prepare the dataset for effective model training, a series of preprocessing steps were undertaken to ensure that the data was both clean and relevant.

The preprocessing began with thorough text cleaning. First, all URLs were removed from the comments using regular expressions. URLs were deemed extraneous since they do not contribute to the sentiment or sarcastic tone of the text. Similarly, user mentions and hashtags were stripped away. These elements, while common in social media discourse, can introduce noise that might detract from the core content of the comments.

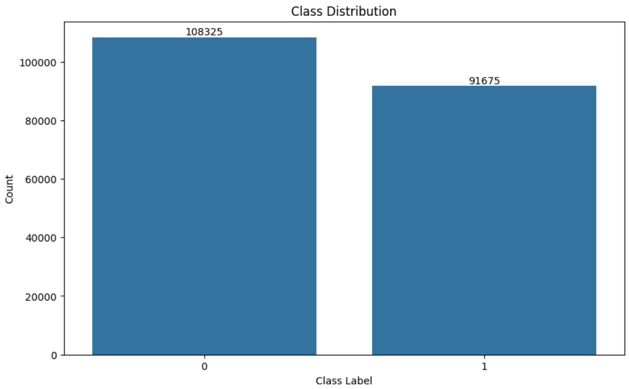
Next, numeric characters and punctuation marks were also removed. The primary aim was to simplify the text, ensuring that the model's focus remained on the words themselves rather than on ancillary symbols or numbers. To further standardize the data, all text was converted to lowercase, and any leading or trailing whitespace was trimmed. This step helped maintain consistency across the dataset, making it easier for the model to process.

Following the initial text cleaning, the dataset underwent a validation process. Emojis were removed, as they can either support or contradict the textual content, complicating the accurate detection of sarcasm. To address potential irregularities in text length and spacing, extra spaces were condensed to a single space. Comments were then evaluated based on their length and complexity. Only comments with at least three words and more than ten characters were retained. Additionally, comments that contained any word repeated more than five times were discarded, as such repetition often indicates low-quality or spam content.

Finally, after these cleaning and validation procedures, the dataset was filtered to ensure only valid entries were included. To strike a balance between computational efficiency and data richness, a sample of 200,000 comments was selected for the study. This sample size was deemed sufficient to capture the nuances of sarcasm while remaining manageable for processing and analysis.

3.3.1 Class Distribution

Unlike many datasets in natural language processing tasks, particularly in sarcasm detection, our dataset exhibits a relatively balanced distribution between sarcastic and non-sarcastic comments. This balance is important as it helps in training models that are equally adept at recognizing both classes, reducing the risk of bias towards the majority class.



As evident from the graph, the difference between the two classes is minimal, with 54.16% of comments being sarcastic and 45.84% being non-sarcastic (an imbalance-ratio of 1.18). This near-equal distribution eliminates the need for advanced balancing techniques such as oversampling, undersampling, or synthetic data generation, which are often necessary in highly imbalanced datasets. The balanced nature of our dataset contributes to the reliability of our model performance metrics, as they are less likely to be skewed by class imbalance. It also allows for a more straightforward interpretation of accuracy scores, as they are equally representative of the model's performance on both classes.

4. NLP Tasks Addressed

4.1 Sarcasm Detection

Sarcasm detection stands as one of the more intricate tasks in the field of natural language processing. This complexity arises from the nuanced and highly context-dependent nature of sarcastic expressions.

4.1.1 Task Definition

At its core, sarcasm detection can be framed as a binary classification problem. The primary objective is to ascertain whether a given piece of text contains sarcasm or not. The process involves taking a text string as input, subjecting it to analysis through various natural language processing techniques and machine learning models, and ultimately producing a binary classification - sarcastic or non-sarcastic - often accompanied by a probability or confidence score.

This study employs four cutting-edge transformer-based models: BERT, RoBERTa, DistilBERT, and ALBERT, to tackle this challenging task.

4.1.2 Challenges in Sarcasm Recognition

The recognition of sarcasm presents a unique set of challenges that collectively make it one of the more formidable tasks in natural language processing. The heavy reliance of sarcasm on context poses a significant hurdle. This context can encompass cultural references, current events, or shared knowledge between the speaker and audience. For machine learning models to accurately detect sarcasm, they must capture and interpret this broader context, a task that proves particularly challenging.

Sarcastic statements often employ subtle linguistic cues such as hyperbole, understatement, or incongruity between literal and intended meanings. These nuances, while often easily grasped by humans, can be exceptionally difficult for machines to detect and interpret correctly. The challenge is further compounded in text-based communication like Reddit comments, where the absence of vocal tone, facial expressions, and gestures - elements that often signal sarcasm in face-to-face interactions - makes detection even more complex.

The variability in sarcastic expression presents another significant challenge. Sarcasm can manifest in countless ways, ranging from obvious exaggeration to deadpan delivery. This wide spectrum of expression makes it difficult to establish comprehensive rules or patterns for sarcasm detection. Moreover, the subjective nature of sarcasm can lead to disagreements even among human annotators, creating challenges in the development of high-quality labeled datasets for training and evaluating models.

Sarcasm often exhibits domain specificity, with its expression varying across different online platforms, communities, or topics. Consequently, models trained on one domain may struggle to perform well on others without additional fine-tuning. The frequent use of figurative language in sarcastic expressions, including metaphors and idioms, adds another layer of complexity, as these elements can be particularly challenging for models to interpret correctly.

A practical challenge in sarcasm detection lies in the nature of real-world data. Typically, sarcastic comments are less prevalent than non-sarcastic ones, leading to imbalanced datasets. This imbalance introduces difficulties in model training and evaluation, requiring careful consideration of appropriate metrics and techniques to address class imbalance.

Addressing these challenges necessitates the application of sophisticated natural language processing techniques, the use of large and diverse training datasets, and the employment of models capable of capturing subtle linguistic patterns and contextual information. The approach adopted in this study, leveraging advanced transformer-based models, aims to tackle these challenges head-on. These models' ability to capture long-range dependencies and contextual nuances in text makes them particularly well-suited to the task of sarcasm detection, offering promise in overcoming the hurdles that have long made this task so challenging in the field of natural language processing.

4.2 Sentiment Analysis

Sentiment analysis, a crucial task in natural language processing, involves determining the sentiment expressed in a piece of text. This task is particularly complex when sarcasm is involved, as sarcasm can dramatically alter the perceived sentiment of a statement.

4.2.1 Task Definition

Sentiment analysis aims to classify text into categories such as positive, negative, or neutral based on the expressed sentiment. The primary objective is to understand the underlying emotion or attitude conveyed in the text. In the context of Reddit comments, sentiment analysis involves evaluating comments to determine whether they reflect a positive, negative, or neutral sentiment.

The complexity of sentiment analysis increases when dealing with sarcastic comments. Sarcasm can mask or invert the actual sentiment being expressed. For instance, a comment that superficially appears positive might be intended as a criticism when interpreted sarcastically. Therefore, the sentiment analysis model must be adept at distinguishing between literal and implied sentiments to provide accurate results.

4.2.2 Literal Sentiment in Sarcastic Contexts

In sarcastic contexts, the literal sentiment of a comment often contrasts sharply with its intended meaning. Sarcasm involves saying the opposite of what one means, which can lead to misleading interpretations if the literal sentiment is taken at face value. For example, a comment like "Oh great, another meeting—just what I needed!" might appear positive due to the use of words like "great" and "needed," but the sarcastic tone indicates that the actual sentiment is negative.

5. Methodology

The methodology is designed to compare the performance of different transformer-based architectures in the task of sarcasm detection, while also incorporating sentiment analysis to explore the relationship between sarcasm and sentiment in online discourse. Given the balanced nature of our dataset, as discussed in Section 3.3.1, we did not need to employ any specific techniques to address class imbalance. This allowed us to focus on optimizing our models for overall performance without the additional complexity of balancing strategies.

5.1 Pre-trained Models for Sarcasm Detection

Our approach utilizes four distinct pre-trained models, each representing a different architecture or variation of transformer-based models. These models are fine-tuned on our dataset of Reddit comments, which has been carefully preprocessed to ensure quality and relevance. The fine-tuning process involves adjusting the model parameters to optimize performance on the sarcasm detection task while retaining the deep linguistic knowledge acquired during pre-training.

5.1.1 BERT-based model

The first model employed in this study is based on BERT (Bidirectional Encoder Representations from Transformers), specifically the 'bert-base-uncased' variant. BERT, introduced by Devlin et al. (2018), has revolutionized natural language processing with its bidirectional training approach and attention mechanism. In our implementation, we utilize the BertForSequenceClassification architecture, which adds a classification layer on top of the BERT base model. The model is initialized with pre-trained weights and then fine-tuned on our sarcasm dataset. We set the maximum sequence length to 256 tokens, balancing between capturing sufficient context and computational efficiency. The model is trained using the AdamW optimizer with a learning rate of 2e-5 and epsilon of 1e-8, over 5 epochs with a batch size of 32. To handle variable-length inputs, we employ attention masking, ensuring the model focuses only on non-padded tokens.

5.1.2 RoBERTa-based model

The second model in our study is based on RoBERTa (Robustly Optimized BERT Pretraining Approach), specifically the 'roberta-base' variant. RoBERTa, introduced by Liu et al. (2019), builds upon BERT's architecture with modified key hyperparameters and training approaches. For our implementation, we use the RobertaForSequenceClassification architecture. Similar to the BERT model, we fine-tune RoBERTa on our sarcasm dataset, but with some key differences in tokenization and preprocessing. RoBERTa uses a byte-level Byte-Pair Encoding (BPE) tokenizer, which allows for a more flexible vocabulary. We set the maximum sequence length to 128 tokens, slightly shorter than BERT, based on empirical observations of performance and efficiency trade-offs. The model is trained using the same AdamW optimizer settings as BERT, maintaining consistency in the training process. RoBERTa's training modifications, including dynamic masking and larger batch sizes during pre-training, potentially make it more robust for tasks like sarcasm detection, where nuanced understanding of language is crucial. By including RoBERTa, we aim to evaluate whether these optimizations translate to improved performance in detecting the subtle linguistic cues associated with sarcasm in online comments.

5.1.3 ALBERT-based model

The third model employed in our study is based on ALBERT (A Lite BERT), specifically the 'albert-base-v2' variant. ALBERT, introduced by Lan et al. (2019), is designed to address the computational and memory constraints of BERT while maintaining or even improving performance. In our implementation, we utilize the AlbertForSequenceClassification architecture. ALBERT incorporates two key innovations: factorized embedding parameterization and cross-layer parameter sharing. These modifications significantly reduce the number of parameters without substantially affecting model performance. For our sarcasm detection task, we maintain consistency with previous models by setting the maximum sequence length to 256 tokens. The model is fine-tuned using the AdamW optimizer with a learning rate of 2e-5 and epsilon of 1e-8, over 5 epochs with a batch size of 32. ALBERT's more efficient architecture allows for faster training and inference times, which is particularly beneficial when dealing with large datasets or deployment in resource-constrained environments. By including ALBERT in our study, we aim to evaluate whether its parameter-efficient design can effectively capture the nuanced linguistic patterns necessary for accurate sarcasm detection.

5.1.4 DistilBERT-based model

The fourth and final model in our sarcasm detection lineup is based on DistilBERT, specifically the 'distilbert-base-uncased' variant. DistilBERT, developed by Sanh et al. (2019), is a distilled version of BERT that retains 97% of its language understanding capabilities while being 40% smaller and 60% faster. We employ the DistilBertForSequenceClassification architecture for our task. DistilBERT achieves its efficiency through knowledge distillation during the pre-training phase, where it learns to mimic a larger BERT model. In our implementation, we set the maximum sequence length to 128 tokens, balancing between capturing sufficient context and leveraging DistilBERT's speed advantages. The model is fine-tuned using the AdamW optimizer with the same hyperparameters as the previous models: a learning rate of 2e-5 and epsilon of 1e-8, over 5 epochs with a batch size of 32. By including DistilBERT, we aim to assess whether a more compact model can effectively detect sarcasm, potentially offering insights into the trade-off between model size and performance in this specific NLP task.

5.2 Sentiment Analysis Models

In addition to sarcasm detection, our study incorporates sentiment analysis to explore the effectiveness of different approaches in classifying sentiment in online discourse. For this task, we tested two distinct methods: a machine learning-based approach using the *cardiffnlp/twitter-roberta-base-sentiment* model and a rule-based approach with VADER (Valence Aware Dictionary and sEntiment Reasoner).

The *cardiffnlp/twitter-roberta-base-sentiment* model, a variant of RoBERTa fine-tuned specifically for sentiment analysis on Twitter data, was chosen for its strong performance in understanding the nuances of social media text. We employed the AutoModelForSequenceClassification architecture, initializing the model with pre-trained weights and fine-tuning it on our dataset. To ensure the model captured the extended context within comments, we set a maximum sequence length of 512 tokens. The sentiment analysis performed by this model classifies comments into three categories: Positive, Neutral, and Negative, offering a more nuanced sentiment classification compared to binary approaches.

In parallel, we used VADER, a rule-based sentiment analysis tool that is well-suited for processing social media content and short text. VADER applies a lexicon and predefined rules to assess sentiment polarity and intensity, making it a quick and interpretable tool for sentiment analysis without the need for training.

Recognizing that both models struggle with interpreting sarcasm accurately, we implemented a strategy where sentiment was inverted when sarcasm was detected in the text. This adjustment was designed to correct the common misclassification of sarcastic content by these models.

Our analysis primarily focused on comparing the performance of these two methods, with particular attention to how sarcasm was distributed across different sentiment categories. Following this, we conducted an analysis of the most frequent bigrams and trigrams within each sentiment category, as identified by both the RoBERTa model and VADER. This allowed us to uncover common phrases and linguistic patterns characteristic of each sentiment, providing deeper insights into how sentiment is expressed in the presence of sarcasm and how each model interprets these patterns.

5.3 Evaluation Metrics

To comprehensively assess the performance of our sarcasm detection and sentiment analysis models, we employ a diverse set of evaluation metrics. For the sarcasm detection task, we primarily focus on accuracy, precision, recall, and F1-score. Accuracy provides an overall measure of correct predictions, while precision and recall offer insights into the model's performance on positive (sarcastic) instances. The F1-score, being the harmonic mean of precision and recall, provides a balanced measure of the model's performance. Additionally, we generate confusion matrices for each model to visualize the distribution of true positives, true negatives, false positives, and false negatives. This allows for a more detailed analysis of where each model excels or struggles in classification.

To further evaluate the models' discriminative capabilities, we employ Receiver Operating Characteristic (ROC) curves and calculate the Area Under the ROC Curve (AUC). The ROC curve illustrates the trade-off between true positive rate and false positive rate at various classification thresholds, while the AUC provides a single scalar value representing the model's overall ability to distinguish between classes. For sentiment analysis, we use similar metrics, adapting them for the multi-class scenario. About the sentiment, we try to understand how the two models interpret sarcasm and sentiment in text. We explore where the models agree or disagree, examine common words and phrases associated with different sentiments, and highlight the challenges in accurately detecting sarcasm and sentiment. The approach is largely qualitative, focusing on the nuances of language interpretation rather than strict quantitative metrics.

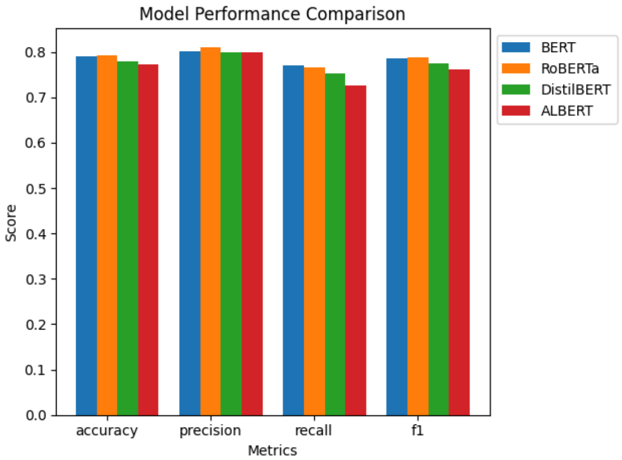
6. Results and Analysis

The evaluation of multiple models for sarcasm detection provides valuable insights into their comparative performance. This section presents a detailed analysis of the results obtained from four state-of-the-art transformer-based models: BERT, RoBERTa, DistilBERT, and ALBERT. These models were fine-tuned on a sarcasm detection task and evaluated using a common test dataset. The balanced distribution of classes in our dataset, as highlighted earlier, contributes to the reliability of these performance metrics. The similar proportions of sarcastic and non-sarcastic comments ensure that our models' high performance is indicative of their ability to accurately detect sarcasm across both classes, rather than excelling only on a majority class.

6.1 Performance Comparison of Sarcasm Detection Models

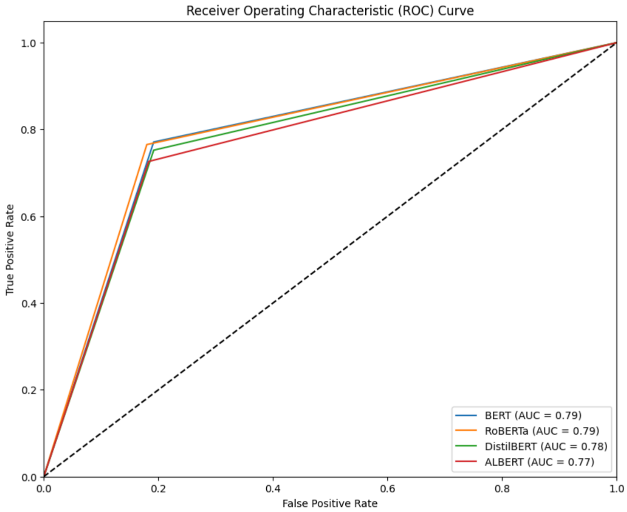
To facilitate a comprehensive understanding of the models' performance, we present the results through both visual representations and numerical metrics. The bar plot and ROC curves offer intuitive comparisons, while the specific metric values provide precise quantitative insights.

Pre-trained Model	Accuracy	Precision	Recall	F1 Score
BERT	0.79	0.80	0.77	0.79
RoBERTa	0.77	0.79	0.73	0.76
DistilBERT	0.78	0.80	0.75	0.77
ALBERT	0.77	0.80	0.73	0.76



As summarized in the table, the models were evaluated based on four key metrics: accuracy, precision, recall, and F1 score. BERT emerged as the top performer with an accuracy of 0.79, precision of 0.80, recall of 0.77, and an F1 score of 0.79, demonstrating its robustness in detecting sarcasm. DistilBERT followed closely with an accuracy of 0.78 and an F1 score of 0.77, highlighting its effectiveness despite being a more lightweight model. RoBERTa and ALBERT both showed slightly lower performance, with accuracies of 0.77 and corresponding F1 scores of 0.76. Interestingly, while ALBERT matched the precision of BERT, it lagged in recall, suggesting that it might be more conservative in identifying sarcasm, leading to fewer false positives but also potentially missing some instances. These results underscore the competitive performance of these models in sarcasm detection, with BERT and RoBERTa generally leading in most metrics, yet all models demonstrating considerable capability in handling this complex NLP task

To complement the bar plot, we also analyze the models' performance using Receiver Operating Characteristic (ROC) curves:

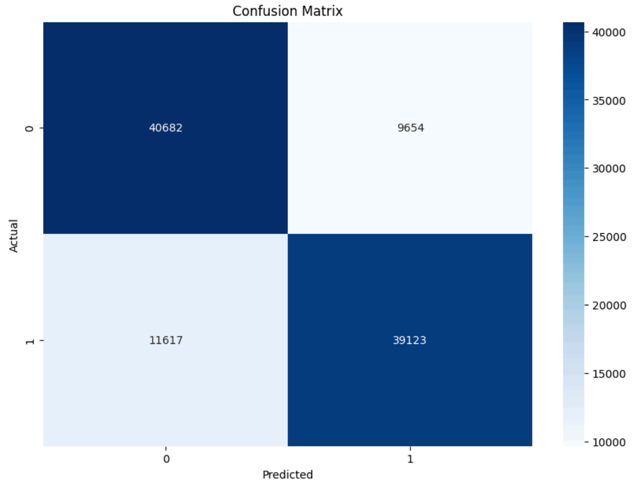


The ROC curves provide insight into the models' performance across various classification thresholds. The Area Under the Curve (AUC) serves as a single scalar value to summarize the overall performance of each model.

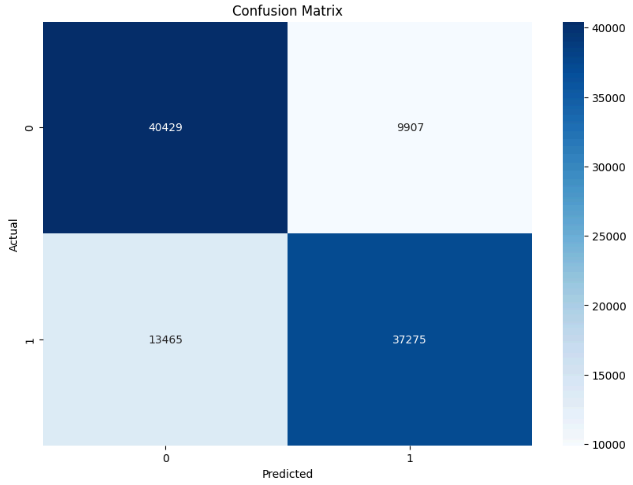
As evident from the graph, all four models perform significantly better than random chance (represented by the diagonal dashed line). The curves for BERT and RoBERTa are nearly identical and slightly outperform those of DistilBERT and ALBERT. This is reflected in their AUC scores, with BERT and RoBERTa both achieving an AUC of 0.79, DistilBERT closely following at 0.78, and ALBERT at 0.77.

The proximity of the curves, particularly between BERT and RoBERTa, indicates that these models have very similar discriminative power in distinguishing between sarcastic and non-sarcastic text. The slight divergence of DistilBERT and ALBERT curves suggests a marginal decrease in performance, but they still maintain strong overall performance.

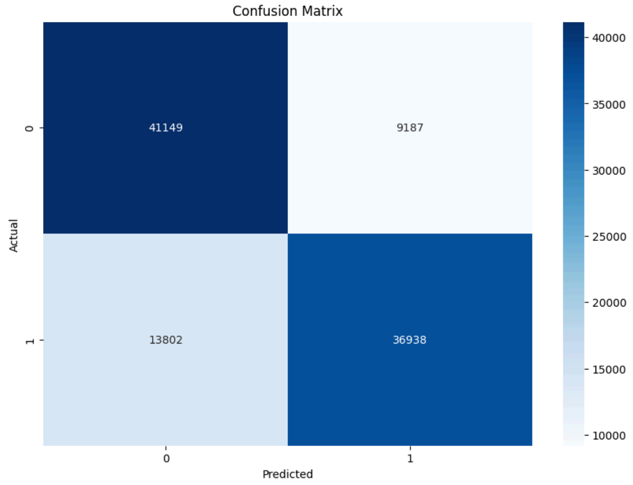
Finally, the Confusion Matrices for all four models give a visual overview of the ability of the models to discriminate between true and false positives and negatives.



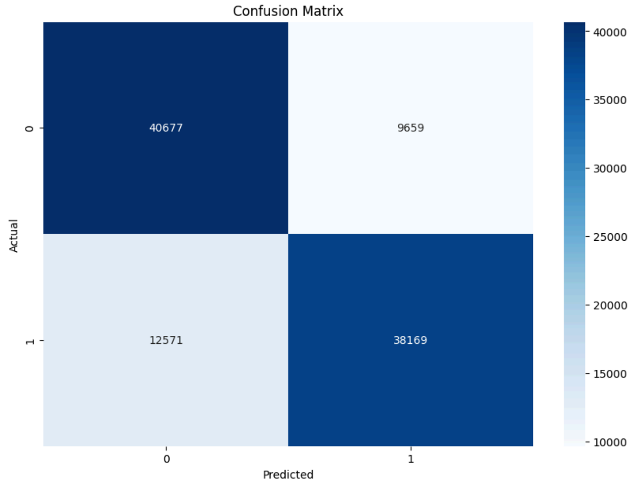
BERT Pre-trained Model: Confusion Matrix



RoBERTa Pre-trained Model: Confusion Matrix



ALBERT Pre-trained Model: Confusion Matrix



DIstilBERT Pre-trained Model: Confusion Matrix

It is important to note that while these models show promising results, there is still room for improvement in sarcasm detection. The complexity and contextual nature of sarcasm present ongoing challenges that may require further advancements in natural language understanding and possibly the incorporation of additional contextual information beyond the text itself.

6.2 Sarcasm Detection model experiments

6.2.1 BERT Results

Text: *I absolutely love waiting in long lines at the DMV.*
Prediction: Not sarcastic
Probability: 0.94

Text: *The weather is beautiful today.*
Prediction: Not sarcastic
Probability: 0.56

Text: *Wow, getting a root canal is so much fun!*
Prediction: Sarcastic
Probability: 0.96

Text: *I'm excited about the new movie coming out.*
Prediction: Not sarcastic
Probability: 0.73

6.2.2 RoBERTa Results

Text: I absolutely love waiting in long lines at the DMV.
Prediction: Sarcastic
Probability: 0.99

Text: The weather is beautiful today.
Prediction: Not sarcastic
Probability: 0.77

Text: Wow, getting a root canal is my favorite way to spend an afternoon!
Prediction: Sarcastic
Probability: 0.91

Text: I'm excited about the new movie coming out next week.
Prediction: Sarcastic
Probability: 0.98

6.2.3 ALBERT Results

Text: I absolutely love waiting in long lines at the DMV.
Prediction: Not sarcastic
Probability: 0.81

Text: The weather is beautiful today.
Prediction: Not sarcastic
Probability: 0.99

Text: Wow, getting a root canal is so much fun!
Prediction: Sarcastic
Probability: 1.00

Text: I'm excited about the new movie coming out.
Prediction: Not sarcastic
Probability: 0.79

6.2.4 DistilBERT Results

Text: I absolutely love waiting in long lines at the DMV.
Prediction: Not sarcastic
Probability: 0.93

Text: The weather is beautiful today.
Prediction: Not sarcastic
Probability: 0.74

Text: Wow, getting a root canal is so much fun!
Prediction: Sarcastic
Probability: 0.99

Text: I'm excited about the new movie coming out.
Prediction: Not sarcastic
Probability: 0.72

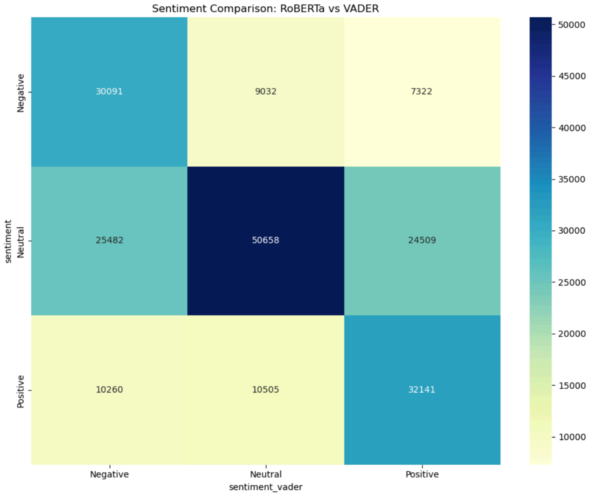
6.2.5 Experiments discussion

Comparing the models' predictions on these newly created example sentences reveals interesting insights into their strengths and limitations. All models correctly identified the sarcasm in 'Wow, getting a root canal is so much fun!' with high confidence. However, they showed varying performance on more subtle cases. For instance, 'I absolutely love waiting in long lines at the DMV' was correctly identified as sarcastic by RoBERTa with high confidence, while the other models classified it as non-sarcastic. This suggests that RoBERTa might be more sensitive to contextual cues of sarcasm. Interestingly, RoBERTa also classified 'I'm excited about the new movie coming out' as sarcastic, which might indicate a tendency to over-identify sarcasm in positive statements. BERT, ALBERT, and DistilBERT were more conservative in their sarcasm predictions, which could result in fewer false positives but might miss some instances of subtle sarcasm. The sentence 'This is the best day ever! My car broke down and I lost my wallet' presents a challenging case of juxtaposition between a positive statement and negative events. DistilBERT was the only model to correctly identify this as sarcastic, albeit with low confidence, highlighting the difficulty in detecting sarcasm that relies on broader context."

6.3 Sentiment Analysis model experiments

In our sentiment analysis, we aimed to explore whether sarcastic comments tend to be more positive or negative and to identify the most frequently used words within these comments. By analyzing sarcasm in conjunction with sentiment, we sought to understand the linguistic nuances that characterize sarcastic language and its impact on sentiment classification.

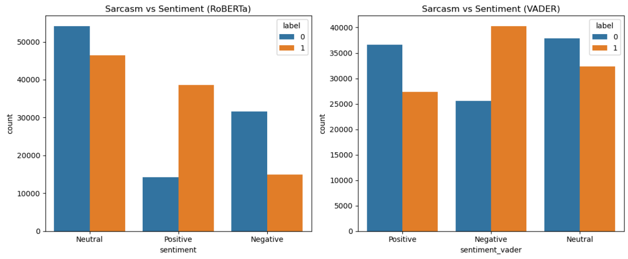
To achieve this, we employed two distinct methods: a machine learning-based approach using the *cardiffnlp/twitter-roberta-base-sentiment* model, and a rule-based approach with VADER. Our analysis revealed significant differences in the results produced by these two models.



The first thing is that the models often agree, especially when it comes to identifying neutral sentiments. In fact, in a whopping 50,658 cases, both said, "This text is neutral." That's quite a number and suggests that, at least for neutral texts, the two systems work similarly. However, things get interesting when we look at the disagreements. RoBERTa seems to be a bit more cautious in judging texts as positive or negative. In many cases where VADER saw a negative sentiment (25,482 times) or a positive one (24,509 times), RoBERTa preferred to set it neutral. There are also cases where they completely disagree. For instance, 7,322 times RoBERTa said "negative" when VADER said "positive," and 10,260 times it was the other way around.

This comparison really highlights how sentiment analysis can be: there's a lot of gray area. Different models can interpret the same text differently, much like humans do.

6.4 Linking Sarcasm with Sentiment



The disagreements between the models, particularly when it comes to positive and negative sentiments, highlight the nuanced nature of language and emotion

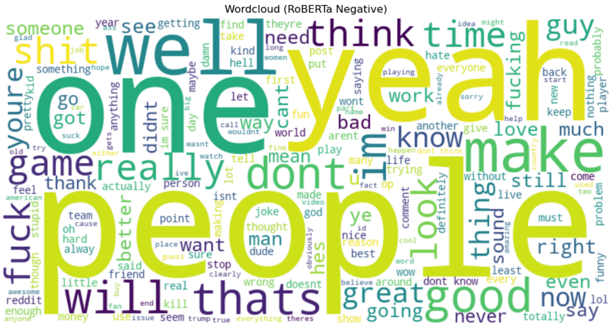
RoBERTa's tendency to classify more texts as neutral compared to VADER might indicate that it has a higher threshold for assigning positive or negative sentiment. This suggests that RoBERTa is more conservative in its assessments, which could help reduce false positives but might also lead to missing subtle emotional cues.

On the other hand, VADER's inclination to assign positive or negative sentiments more readily might make it more sensitive to emotional nuances in text. However, this sensitivity could also result in overinterpretation in some cases, detecting emotions where there might be none.

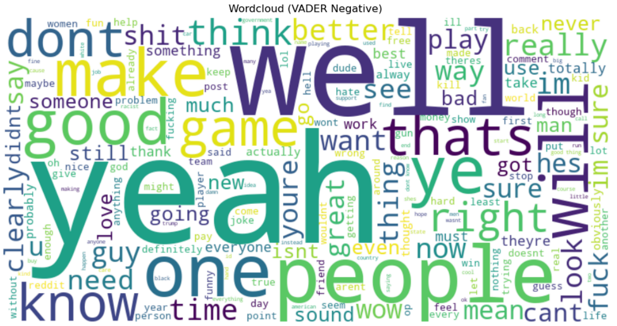
Instances where the models completely disagree, such as one labeling a text as positive while the other labels it as negative, create doubt about the interpretation of the sentence. These cases likely represent complex or ambiguous texts that pose a challenge even for advanced NLP models.

In this analysis, the models do account for the more common scenario where sarcasm inverts the sentiment—for instance, when a positive statement is meant to convey a negative sentiment. However, they fall short when it comes to handling cases where sarcasm doesn't invert the sentiment but instead emphasizes it. For example, a sarcastic comment intended to exaggerate negativity or positivity might amplify the sentiment rather than reverse it. These cases, where sarcasm intensifies the underlying sentiment rather than flipping it, are particularly complex and were not managed by the models in this evaluation.

Taking the negative sentiment as an example, we can observe interesting differences in how RoBERTa and VADER classify and emphasize words in their respective negative sentiment categories.



RoBERTa's negative sentiment word cloud prominently features words like "one," "think," "year," and "time" alongside more explicitly negative terms such as "shit" and "fuck." This suggests that RoBERTa may be picking up on contextual negativity or associating certain neutral words with negative sentiments based on their frequent occurrence in negative contexts.

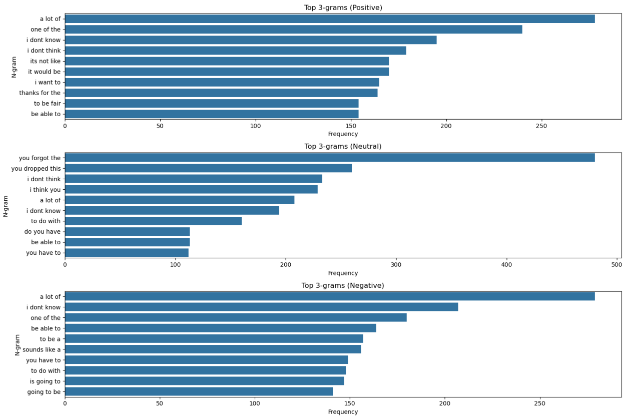


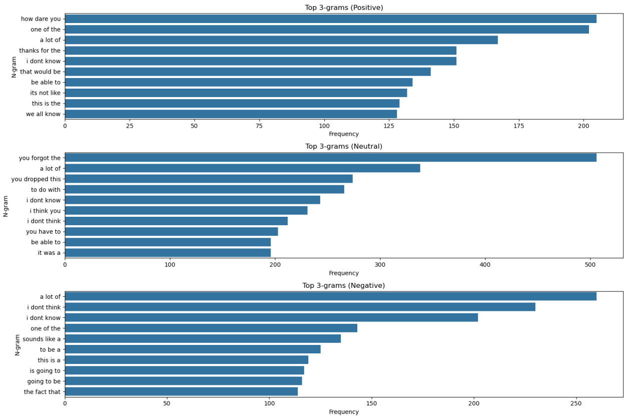
In contrast, VADER's negative sentiment word cloud gives more prominence to words like "don't," "well," "yeah," and "want." Interestingly, VADER seems to place greater emphasis on negations and words that might indicate disagreement or dissatisfaction. The presence of seemingly positive words like "good" and "love" in large font sizes is notable, possibly indicating VADER's recognition of these terms being used sarcastically or in negative contexts.

Both models include profanities and explicitly negative words, but they differ in the relative prominence given to these terms. This variance could reflect differences in the training data or the underlying algorithms used by each model to classify sentiment.

6.4.1 N-Grams analysis for Sentiment Analysis

Examining the top 3-grams for RoBERTa and VADER across positive, neutral, and negative sentiments provides further insight into how these models interpret and classify text, complementing our earlier observations from the word clouds.





Examining the top 3-grams identified by RoBERTa and VADER across various sentiments provides a more nuanced understanding of how these models interpret and classify text, complementing the observations derived from earlier word cloud analyses.

In the context of positive sentiment, both models frequently identify phrases such as "a lot of" and "one of the," which are generally indicative of favorable opinions or experiences. Notably, RoBERTa's top positive 3-gram is "how dare you," a phrase that might initially appear counterintuitive in a positive context. This could suggest that RoBERTa is capable of detecting the playful or affectionate use of this phrase, which can indeed convey positive sentiment depending on the surrounding context. In contrast, VADER highlights more straightforwardly positive phrases such as "thanks for the" and "to be fair," which are clearly associated with expressions of gratitude or fairness.

Regarding neutral sentiment, both RoBERTa and VADER exhibit similar patterns. Phrases like "you forgot the" and "you dropped this" are frequently identified, indicating that both models recognize these statements as factual or emotionally neutral. Additionally, the presence of phrases such as "i don't know" and "i think you" in the neutral category for both models corresponds with the prominence of the words "think" and "know" observed in earlier word cloud analyses. This suggests that these phrases are commonly used in contexts that are neither strongly positive nor negative but rather neutral or speculative.

In the analysis of negative sentiment, the phrase "a lot of" appears prominently for both models, which is particularly interesting given its appearance in the positive sentiment category as well. This underscores the critical importance of context in sentiment analysis, as the same phrase can convey different sentiments depending on its usage. Both models also frequently classify phrases such as "i don't think" and "i don't know" as negative, aligning with the earlier word cloud findings where these expressions were linked to negative sentiment. RoBERTa appears to give greater weight to phrases like "sounds like a" and "is going to" in negative contexts, possibly reflecting its sensitivity to implied negativity or sarcasm. On the other hand, VADER includes phrases such as "be able to" and "you have to" among its top negative 3-grams, which may indicate a different approach to identifying implicit negativity.

7. Conclusions

7.1 Final remarks on Sarcasm Detection

Our comprehensive analysis of sarcasm detection using four state-of-the-art transformer-based models—BERT, RoBERTa, DistilBERT, and ALBERT—has yielded valuable insights into the capabilities and limitations of these models in tackling this complex natural language processing task.

The performance metrics reveal that all four models demonstrated strong capabilities in detecting sarcasm, with accuracies ranging from 0.77 to 0.79. BERT emerged as the top performer, achieving the highest accuracy (0.79), precision (0.80), recall (0.77), and F1 score (0.79). This suggests that BERT's bidirectional training approach and attention mechanism are particularly effective in capturing the nuanced contextual cues that often characterize sarcastic statements.

DistilBERT, despite being a more lightweight model, closely followed BERT's performance with an accuracy of 0.78 and an F1 score of 0.77. This indicates that the knowledge distillation techniques employed in DistilBERT's design have been successful in retaining much of BERT's sarcasm detection capabilities while reducing computational requirements. This finding is particularly relevant for applications where computational resources may be limited but high performance is still necessary.

RoBERTa and ALBERT, while slightly behind in overall accuracy (both at 0.77), showed comparable precision to BERT (0.80). However, their lower recall scores suggest they may be more conservative in identifying sarcasm, potentially missing some instances but also reducing false positives. This trade-off could be advantageous in scenarios where high precision is more critical than high recall.

The ROC curve analysis further corroborated these findings, with all models significantly outperforming random chance. The close proximity of the curves, particularly between BERT and RoBERTa, underscores the competitive performance of these models in distinguishing between sarcastic and non-sarcastic text.

However, our qualitative analysis of model predictions on example sentences revealed some limitations. While all models excelled at identifying obvious sarcasm, they showed varying performance on more subtle cases. RoBERTa demonstrated higher sensitivity to contextual cues of sarcasm but also showed a tendency to over-identify sarcasm in some positive statements. BERT, ALBERT, and DistilBERT were more conservative in their sarcasm predictions, which could result in fewer false positives but might miss instances of subtle sarcasm.

These findings highlight the ongoing challenges in sarcasm detection, particularly in cases where broader context or cultural knowledge is required. The juxtaposition of positive and negative elements, a common feature in sarcastic statements, proved especially challenging for the models to interpret consistently.

In conclusion, while these transformer-based models have shown impressive capabilities in sarcasm detection, there remains room for improvement. Future work could focus on incorporating broader contextual information, perhaps by considering the entire thread of a conversation or the specific subreddit context. Additionally, fine-tuning these models on more diverse datasets that include various types of sarcasm could potentially enhance their ability to detect subtle or culturally-specific instances of sarcasm.

7.2 Final remarks on Sentiment Analysis

Our analysis of sentiment using the RoBERTa and VADER models yielded intriguing findings. Both models performed well in identifying neutral sentiments, but their handling of positive and negative sentiments varied. RoBERTa tended to classify many statements as neutral, whereas VADER was more decisive in categorizing sentiments.

The challenge of detecting sarcasm proved particularly difficult for both models, underscoring the complexity of this linguistic feature. The models struggled with sarcasm, especially when it served to emphasize rather than invert the sentiment.

Additionally, context played a crucial role in interpreting sentiment. The same phrase could convey different meanings depending on its usage, much like deciphering sincerity versus sarcasm in casual conversation.

Overall, this study highlights the advancements AI has made in sentiment analysis while also revealing ongoing challenges.

8. References

Regarding the choice of models:

- LinkedIn Update on Model Choice: https://www.linkedin.com/feed/update/urn:li:activity:7231282597962620929/

Link to our GitHub repository:

- GitHub Repository: https://github.com/alexcri90/NLP_SarcasmDetector_Bicocca

Bibliography:

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- 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 preprint arXiv:1907.11692.
- 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 preprint arXiv:1909.11942.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
- Hutto, C.J., & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- Tepperman, J., Traum, D., & Narayanan, S. (2014). "Yeah right": Sarcasm recognition for spoken dialogue systems. In Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 97-104).
- Barbieri, F., Saggion, H., & Ronzano, F. (2014). Modelling Sarcasm in Twitter, a Novel Approach. In Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 50-58).

- González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: a closer look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 581-586).
- Ptáček, T., Habernal, I., & Hong, J. (2014). Sarcasm detection on Czech and English twitter. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 213-223).
- Ghosh, A., & Veale, T. (2016). Fracking sarcasm using neural network. In Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 161-169).