Souad Khalifeh 12132429

Section b

Sentiment Analysis

Topics in Computer Science

**Table of Contents**

[**Abstract:** 2](#_Toc167397574)

[**Introduction:** 2](#_Toc167397575)

[**Background:** 3](#_Toc167397576)

[**Related Works:** 3](#_Toc167397577)

[Convolutional Neural Networks(CNNs): 3](#_Toc167397578)

[Recurrent Neural Networks (RNNs): 4](#_Toc167397579)

[Deep Neural Networks (DNNs): 4](#_Toc167397580)

[Naïve Bayes: 4](#_Toc167397581)

[Support Vector Machine (SVM): 4](#_Toc167397582)

[**Motivation:** 5](#_Toc167397583)

[**Methodology:** 5](#_Toc167397584)

[Model Selection: 5](#_Toc167397585)

[Twitter RoBERTa Base: 5](#_Toc167397586)

[DistilBERT-base-uncased emotion: 5](#_Toc167397587)

[BERT-base-multilingual-uncased-sentiment: 6](#_Toc167397588)

[Models Workflow: 6](#_Toc167397589)

[Experimentation Protocol: 8](#_Toc167397590)

[Hadrware and Software Specifications: 9](#_Toc167397591)

[Hardware: 9](#_Toc167397592)

[Software: 9](#_Toc167397593)

[**Experiments and Results:** 9](#_Toc167397594)

[Experiments: 10](#_Toc167397595)

[Evaluation of Sentiment Classification Accuracy using English Dataset: 11](#_Toc167397596)

[Evaluation of Sentiment Classification Accuracy using French Dataset: 12](#_Toc167397597)

[**Discussion:** 13](#_Toc167397598)

[Model Performance in English Sentiment Analysis: 13](#_Toc167397599)

[Model Performance in French Sentiment Analysis: 14](#_Toc167397600)

[Comparative Analysis: 15](#_Toc167397601)

[**Conclusion:** 16](#_Toc167397602)

[**References** 18](#_Toc167397603)

Sentiment Analysis

## **Abstract:**

Sentiment analysis involves the process of extracting and analyzing sentiments from textual data. This research delves into evaluating various sentiment analysis models to identify their strengths and weaknesses in handling text classification tasks. The primary objective is to compare the performance of three-state-of-the-art models: Twitter RoBERTa Base, DistilBERT-base-uncased emotion, and BERT-base-multilingual-uncased-sentiment. The models were assessed using a consistent methodology across English and French datasets, covering multiple domains such as social media posts, product and movie reviews. Our experiments demonstrate that Twitter RoBERTa Base excels in social media English Text, and BERT-base-multilingual-uncased-sentiment shows robustness across multiple languages. The results underscore the importance of model selection based on the specific requirements of sentiment analysis tasks. Future work will focus on enhancing model accuracy through retraining, fine-tuning and exploring diverse datasets.

## **Introduction:**

Sentiment analysis, also called Opinion analysis or Opinion mining, is the process of gathering and analyzing people’s opinions, thoughts, and impressions regarding various topics, products, subjects, and services [1]. It involves determining the information’s polarity such as positive, negative, or neutral [2]. Beyond its application in understanding consumer perceptions, sentiment analysis is pivotal in diverse fields such as marketing, customer service, and social media monitoring.

In marketing, sentiment analysis helps companies gain insights into consumer sentiments, assess brand reputation, and evaluate marketing campaigns’ effectiveness. Similarly, in customer service, sentiment analysis enables organizations to prioritize and address customer feedback promptly, leading to improved customer satisfaction and loyalty. Moreover, sentiment analysis is valuable in social media monitoring, where it facilitates real-time tracking and analysis of public opinions, trends, and sentiments on social media platforms [3].

As sentiment analysis continues to evolve in response to the challenges posed by natural language intricacies, this paper aims to delve deeper into the methodologies employed, evaluate their effectiveness, and propose avenues for further enhancement.

## **Background:**

Sentiment analysis is a Natural Language Processing (NLP) task that deals with the detection and classification of sentiments in texts [4] . As a subfield of cognitive science, artificial intelligence, and linguistics, NLP aims to process and analyze natural language data [3]. NLP relies heavily on linguistic theories and principles to develop algorithms and models that can understand and interpret human language.

However, sentiment analysis encounters unique challenges due to the complexity of the human language in textual communication. The presence of slang, irony, sarcasm, abbreviations, spelling errors, and emoticons complicates sentiment analysis [5]. For instance, sarcasm and even simple negations can completely reverse the predicted sentiment from the actual opinion [6]. These challenges impact the accuracy of sentiment analysis algorithms, since expressions of sentiment may vary in different contexts which leads to inaccurate classifications.

Linguistics plays a pivotal role in understanding the intricacies of language structure and semantics, which are essential for developing robust sentiment analysis algorithms. Linguistics is the science of language which includes Phonology which refers to sound, Morphology word formation, Syntax sentence structure, Semantics syntax, and Pragmatics which refers to understanding [7]. By incorporating linguistic insights into NLP techniques, researchers can address the complexities of sentiment analysis more effectively and improve the accuracy of sentiment classification in diverse textual data.

## **Related Works:**

Over the past few years, significant advancements have been made in the realm of sentiment analysis, with a focus on improving opinion mining methods. The use of deep learning and machine learning algorithms for prediction, modeling, training, and imitating human-like behavior has increased [8]. Among these approaches, deep learning algorithms have gained considerable attention. Deep learning, a subfield of machine learning, employs neural networks such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Deep Neural Networks (DNNs) inspired by the structure of the human brain [9]. These neural networks are adept at learning complex patterns and relationships from data, making them particularly well-suited for sentiment analysis tasks.

Alongside these advanced deep learning techniques, traditional machine learning algorithms such as Naïve Bayes and Support Vector Machine (SVM) have also proven effective in sentiment analysis, offering alternative approaches to text classification.

## Convolutional Neural Networks(CNNs):

Convolutional Neural Networks (CNNs) are a sub-class of neural networks that take advantage of the spatial structure of the inputs [10], inspired by the human visual system [11]. They consist of several layers of convolutions and activation functions like ReLU (Rectified Linear Unit, an activation function commonly used in neural networks to introduce non-linearity by setting all negative values in the input to zero [12]), enabling them to model non-linear relationships effectively. Originally developed for tasks such as image classification, object detection, and semantic segmentation [13]. However, CNNs have been successfully adapted to process sequential data, including text. Their ability to capture local patterns and hierarchical representations in text sequences makes them particularly suitable for sentiment analysis tasks, where contextual understanding is crucial [11].

## Recurrent Neural Networks (RNNs):

Recurrent Neural Networks (RNNs) are a class of artificial neural networks designed to handle sequential data by maintaining internal memory. They have connections that form directed cycles, allowing them to capture temporal dependencies in data sequences [14]. They have been widely used in various natural language processing tasks, including sentiment analysis, language modeling, and machine translation, RNNs are particularly well-suited for tasks where context plays a crucial role as they can process input sequences of variable length and maintain a state that summarizes information from previous steps [15]. Despite their effectiveness, RNNs suffer from the vanishing gradient problem, which limits their ability to capture long-term dependencies in data sequences [16]. Recent advancements such as Long Short-Term Memory (LTSM) and Gated Recurrent Unit (GRU) architectures have helped mitigate this issue, making RNNs a popular choice for sequential data processing tasks in both academia and industry [17].

## Deep Neural Networks (DNNs):

Deep Neural Networks (DNNs) are a class of artificial neural networks with multiple layers between the input and the output layers excelling at learning hierarchical representations of data, extracting abstract features from raw inputs, and performing complex transformations [18]. They have been widely used in various machine-learning tasks, including image recognition, natural language processing, and speech recognition. Their ability to automatically learn intricate patterns and relationships from data has led to significant advancements in these fields [19]. However, training deep neural networks can be computationally expensive and requires large amounts of labeled data. Recent research has focused on techniques such as transfer learning, regularization, and optimization algorithms to improve the efficiency and performance of DNNs [20].

## Naïve Bayes:

Naïve Bayes is a probabilistic machine learning approach based on Bayes’ theorem. It assumes independence between features and calculates the probability of a given input belonging to a particular class, making it suitable for classification tasks such as sentiment analysis. [21]

## Support Vector Machine (SVM):

Support Vector Machine (SVM), specifically the Support Vector Classifier (SVC) variant, are robust supervised learning models that are effective in both linear and non-linear classification challenges. The strength of SVMs lies in their ability to identify the most informative data points, known as support vectors, which are critical in constructing an optimal decision boundary [22].

## **Motivation:**

The rapid growth of sentiment analysis approaches is of critical importance in comprehending and evaluating textual content across diverse fields [1]. Despite advancements in deep learning and traditional machine learning, achieving high accuracy remains a challenge [2]. To find opportunities for innovation and development, this study aims to compare and evaluate existing sentiment analysis models, identifying strengths and weaknesses [4].

The need to improve sentiment analysis accuracy is clear from its growing use of this tool in crucial decision-making processes, including brand management, product management, and consumer engagement strategies [3]. Inaccurate results can lead to poorly informed decisions and negative consequences for businesses and organizations.

The study highlights how important it is to thoroughly assess and contrast sentiment analysis models to ensure their effectiveness and reliability in real-world applications.

## **Methodology:**

This section delves into the theoretical foundations of sentiment analysis models by examining relevant research papers. Additionally, we explore the technical details of each model, including their training objectives, tokenization methods, and architectural differences. A deep understanding of these aspects is crucial for conducting meaningful comparisons and assessing the accuracy of these models.

## Model Selection:

We introduce three popular sentiment analysis models for evaluation:

### Twitter RoBERTa Base:

Twitter RoBERTa Base [23] is specifically trained on a large corpus of Twitter data, approximately 60 million tweets [24], making it well-suited for sentiment analysis tasks involving social media text using the TweetEvalBenchmark [25]. The unique linguistic characteristics of Twitter, such as slang, abbreviations, and emoticons, pose challenges that require specialized models for accurate analysis.

RoBERTa is a variant of BERT (Bidirectional Encoder Representations from Transformers) that employs a masked language model objective and dynamic masking during training. It captures contextual information effectively and has demonstrated superior performance on a wide range of NLP tasks: Emotion Recognition, Emoji Prediction, Irony Detection, Hate Speech Detection, Offensive Language Identification, Stance Detection and Sentiment Analysis.

### DistilBERT-base-uncased emotion:

DistilBERT- base-uncased emotion is a variant of the DistilBERT architecture, specifically fine-tuned for emotion-related tasks.lt prioritize efficiency and effectiveness. As a distilled version of BERT, it offers comparable performance to BERT while being smaller and faster. It retains over 97% of BERT’s performance while having 40% fewer parameters and computational resource. It was pretrained on same corpus in a self-supervised manner (using raw texts without human labeling) using the BERT base model as a teacher. This corpus comprises 11,038 books and English Wikipedia articles excluding lists, tables and headers. This makes it suitable for deployment in resource-constrained environments or applications where speed is crucial.

DistilBERT-base-uncased emotion does not differentiate between uppercase and lowercase letters during tokenization. This makes the model more flexible and suitable for tasks where case sensitivity is not crucial. Also, the model is specially fine-tuned for emotion classification tasks. [26]

### BERT-base-multilingual-uncased-sentiment:

The BERT- base-multilingual-uncased-sentiment model [27] represent a variant of the BERT architecture, tailored specifically for sentiment analysis tasks to meet the demand for multilingual sentiment analysis, particularly in the realm of product reviews.

This model extends the capabilities of BERT to handle text in six languages: English, Dutch, German, French, Spanish, and Italian. It predicts the sentiment of product reviews as a number of stars (ranging from 1 to 5). [28]

Now, that we have introduced the selected sentiment analysis models and their respective characteristics, we will proceed to outline the step-by-step workflow for sentiment analysis using each model, highlighting their similarities and differences.

## Models Workflow:

The sentiment analysis of our three selected models involves the following steps: text preprocessing, feature extraction, sentiment classification, model architecture, and evaluation metrics

1. ***Text Preprocessing:***

In the realm of text preprocessing, the Twitter RoBERTa Base model undertakes the task of cleaning input text of Twitter data. It replaces usernames (beginning with ‘@’) and URLs with placeholders and remove hashtags, mentions, punctuations, numbers, stop words, whitespace and duplicate tweets.[25]

Conversely, the DistilBERT-base-uncased emotion sentiment model follows a similar path, aiming remove any noise and irrelevant information. It includes handling special characters, punctuation, and HTML tags, as well as converting the text to lowercase for consistency. Additionally, DistilBERT employs tokenization using WordPiece and a vocabulary size of 30,000. It may also apply token masking in certain cases, which indicates to the model that those tokens should be ignored during training or inference [26].

Similarly, the BERT-base-multilingual-uncased-sentiment model operates on tokenization input text into subword tokens using WordPiece Tokenizer. It too addresses special characters, punctuation and HTML while ensuring text uniformity through to lowercase conversion. [28]

1. ***Feature Extraction:***  
   Moving to feature extraction, the Twitter RoBERTa Base model encode tokenized tezt into numerical representation or embeddings. This process converts the data into a format that the model can understand and process effectively. It employs tokenization where the input text is split into individual tokens or words and then each token is assigned to a unique numerical value based on its position in a predefined vocabulary. [25]

On the other hand, DistilBERT-base-uncased emotion takes a different approach by randomly making 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words. This method allows the model to learn a bidirectional representation of the sentence. [26]

Similarly, BERT-base-multilingual-uncased-sentiment converts tokenized input text into numerical embeddings. BERT employs a transformer architecture, which captures contextual information effectively by processing the entire input text simultaneously rather than sequentially [27].

1. ***Sentiment Classification:***

In the domain of sentiment classification, the Twitter RoBERTa Base model predicts sentiment scores for positive, neutral and negative labels [23]. It then generates probability distributions over the possible sentiment labels. Finally, the sentiment label with the highest probability is assigned to the input text as the predicted sentiment.

Conversely. The DistilBERT-base-uncased emotion model assigns one or more emotion labels to the input text (joy, sadness, love, fear, anger, surprise). It returns a list of these labels along with their corresponding confidence scores (probabilities).

Meanwhile, the BERT-base-multilingual-uncased-sentiment model predicts the sentiment of a review as a number of stars ranging from 1 to 5. Employing a softmax classifier, it assigns a probability distribution over the possible sentiment labels and selects the sentiment label with the highest probability for the input text [29].

1. ***Evaluation Metrics:***

Lastly, all three models are evaluated using metrics such as accuracy, precision, recall, F1-score and area under the receiver operating characteristic curve (AUC-ROC) for assessing its performance. These metrics provide insights into the efficacy and reliability of each model’s sentiment analysis capabilities, allowing for informed comparisons and conclusions.

## Experimentation Protocol:

Our evaluation prioritizes assessing the performance and accuracy of these models without retraining. Although retraining models can provide optimized results, our methodology aims to assess the out-of-the-box- functionality and capabilities of these pre-trained models.

We examine each model’s unique characteristics, functionalities, and underlying architectures to discern their strengths and weaknesses. This comparative analysis will be based on a series of test scenarios designed to evaluate their performance in identifying accurately sentiment polarity (positive, negative, or neutral).

To ensure a comprehensive evaluation, we employ a diverse set of evaluation metrics and test datasets tailored to assess the models’ proficiency in handling varied textual inputs. This approach allows us to provide insights into the real-world applicability of these models across different languages and domains.

By systematically testing these pre-trained models, we aim to offer a nuanced understanding of their operational efficiency and suitability for various sentiment analysis tasks, without retraining.

For our experimentation, we will utilize publicly available sentiment analysis datasets in multiple languages, including English, Dutch, German, French, Spanish and Italian . These datasets cover a wide range of domains such as product reviews, social media posts, news articles and movie reviews, ensuring a robust evaluation of the models’ capabilities across different contexts.

While all models will be evaluated on their performance in English sentiment analysis, we will specifically test the BERT-base-multilingual-uncased-sentiment-model on multiple languages. This model uniquely supports multilingual capabilities, which enable us to assess its performance in sentiment analysis across diverse linguistic landscapes.

We will be evaluating each model’s performance on the test datasets using the predefined evaluation metrics and conduct a comparative analysis of the model’s performance across different languages and domains. Based on this analysis, we will identify the strengths and weaknesses of each model. Finally, we will be discussing the implications of the findings and provide insights into the suitability of each model for various sentiment analysis tasks.

## Hadrware and Software Specifications:

In this subsection, we provide details about the hardware and software environment used for conducting the experiment.

### Hardware:

The experiments were conducted on a desktop workstation with the following specifications:

* Processor (CPU):
  + Google Colab provides access to virtual CPUs (vCPUs). The specific type and performance can vary, but typically it includes high-performance Intel Xeon processors.
* Memory (RAM): Approximately 12.72 GB RAM
* Graphics Card (GPU): Google Colab offers access to GPUs which can significantly speed up computation for deep learning tasks:
  + NVIDIA Tesla K80

### Software:

The experiments were conducted using the following software environment:

* Operating System: Windows 11Pro Version 23H2
* Python Environment: Python 3.8.5
* Integrated Development Environment (IDE): Google Collab
  + Google Collab was utilized for its cloud-based Python environment, facilitating collaboration and access to GPU resources.
* Libraries:
  + PyTorch
  + Transformers
  + NumPy
  + Collections

## **Experiments and Results:**

In this section, we present the experimental results of evaluating three states of Natural Language Processing (NLP) models for sentiment analysis: Twitter RoBERTa Base, DistilBERT- base-uncased emotion, BERT- base-multilingual-uncased-sentiment. Each model was tested using a consistent methodology across multiple datasets.

To ensure a fair comparison, we collected diverse datasets consisting of 1600 sentences for each sentiment label (positive, neutral and negative). These datasets were shuffled to prevent bias. For each experiment, a random sample of 100 sentences was selected from the shuffled dataset, and the models’ predictions were recorded. Additionally, to assess language versatility, a separate French dataset was collected,as both Twitter RoBERTa and DistilBERT lack native support for French language processing, while BERT Multilingual offers multilingual capabilities.

In total, five experiments were conducted: three for the English dataset and two for the French dataset. There results were analyzed to provide insights into the performance of each model in classifying sentiment across different languages and datasets. This systematic approach allows for a comprehensive assessment of the models’ capabilities and facilitates meaningful comparisons to identify the most effective model for sentiment analysis tasks.

For each experiment, three bars were plotted, each representing one of the three models and indicates the overall accuracy for all labels achieved by the respective model.

Besides, three confusion matrices were generated, one for each model, combining the results of all experiments. Each confusion matrix displays the counts of true positives, false positives, true negatives, and false negatives, as well as true neutrals and false neutrals, providing a comprehensive overview of the models’ performance across all experiments.

## Experiments:

Three experiments were conducted for each of the three models using the English dataset to determine and compare their overall accuracy

Two experiments were conducted for each of the three models using the French dataset to determine and compare their overall accuracy

## Evaluation of Sentiment Classification Accuracy using English Dataset:

#### **Twitter RoBERTa model:**

|  |  |  |
| --- | --- | --- |
|  | True | False |
| Positive | 73 | 29 |
| Negative | 75 | 28 |
| Neutral | 80 | 15 |

The results presented in the table above illustrates the performance of Twitter RoBERTa model across the first three experiments conducted on the English Dataset, each involving 100 sentences, totaling 300 sentences. True positives, false positives, true negatives, false negatives, true neutrals and false neutrals providing insights into the model’s ability to accurately classify sentiments.

#### **DistilBERT model:**

|  |  |  |
| --- | --- | --- |
|  | True | False |
| Positive | 16 | 115 |
| Negative | 8 | 79 |
| Neutral | 71 | 11 |

The results presented in the table above illustrates the performance of DistilBERT model across the first three experiments conducted on the English Dataset, each involving 100 sentences, totaling 300 sentences. True positives, false positives, true negatives, false negatives, true neutrals and false neutrals providing insights into the model’s ability to accurately classify sentiments.

#### **BERT multilingual model:**

|  |  |  |
| --- | --- | --- |
|  | True | False |
| Positive | 77 | 25 |
| Negative | 50 | 39 |
| Neutral | 24 | 85 |

The results presented in the table above illustrates the performance of BERT multilingual model across the first three experiments conducted on the English Dataset, each involving 100 sentences, totaling 300 sentences. True positives, false positives, true negatives, false negatives, true neutrals and false neutrals providing insights into the model’s ability to accurately classify sentiments.

## Evaluation of Sentiment Classification Accuracy using French Dataset:

#### **Twitter RoBERTa model:**

|  |  |  |
| --- | --- | --- |
|  | True | False |
| Positive | 18 | 72 |
| Negative | 6 | 104 |

The results presented in the table above illustrates the performance of Twitter RoBERTa model across the first two experiments conducted on the French Dataset, each involving 100 sentences, totaling 200 sentences. True positives, false positives, true negatives and false negatives providing insights into the model’s ability to accurately classify sentiments.

#### **DistilBERT model:**

|  |  |  |
| --- | --- | --- |
|  | True | False |
| Positive | 3 | 60 |
| Negative | 16 | 47 |

The results presented in the table above illustrates the performance of DistilBERT model across the first two experiments conducted on the French Dataset, each involving 100 sentences, totaling 200 sentences. True positives, false positives, true negatives and false negatives providing insights into the model’s ability to accurately classify sentiments.

#### **BERT multilingual model:**

|  |  |  |
| --- | --- | --- |
|  | True | False |
| Positive | 55 | 4 |
| Negative | 40 | 35 |

The results presented in the table above illustrates the performance of BERT multilingual model across the first two experiments conducted on the French Dataset, each involving 100 sentences, totaling 200 sentences. True positives, false positives, true negatives and false negatives providing insights into the model’s ability to accurately classify sentiments.

## **Discussion:**

The results of our experiments reveal valuable insights into the performance and suitability of the three selected sentiment analysis model Twitter RoBERTa Base, DistilBERT-base-uncased emotion, and BERT-base-multilingual-uncased-sentiment across different datasets and two languages. Our findings highlight the strengths and limitations of each model, contributing to a comprehensive understanding of their capabilities in various sentiment analysis tasks.

### Model Performance in English Sentiment Analysis:

1. **Twitter RoBERTa Base:**

* ***Strengths:***

 This model demonstrated a high degree of accuracy in classifying sentiments in English tweets. Its training on a large corpus of Twitter data enables it to effectively handle the unique linguistic features of social media text, such as slang, abbreviations, and emoticons.

It achieved the highest number of true sentiments across positive, negative and neutral sentiments, indicating its robust performance in identifying the correct sentiment.

* ***Limitations:***

Despite its strengths in processing Twitter data, the model's performance may decline when applied to text from other domains due to its specialized training. This limits its generalizability to other types of text outside social media platforms.

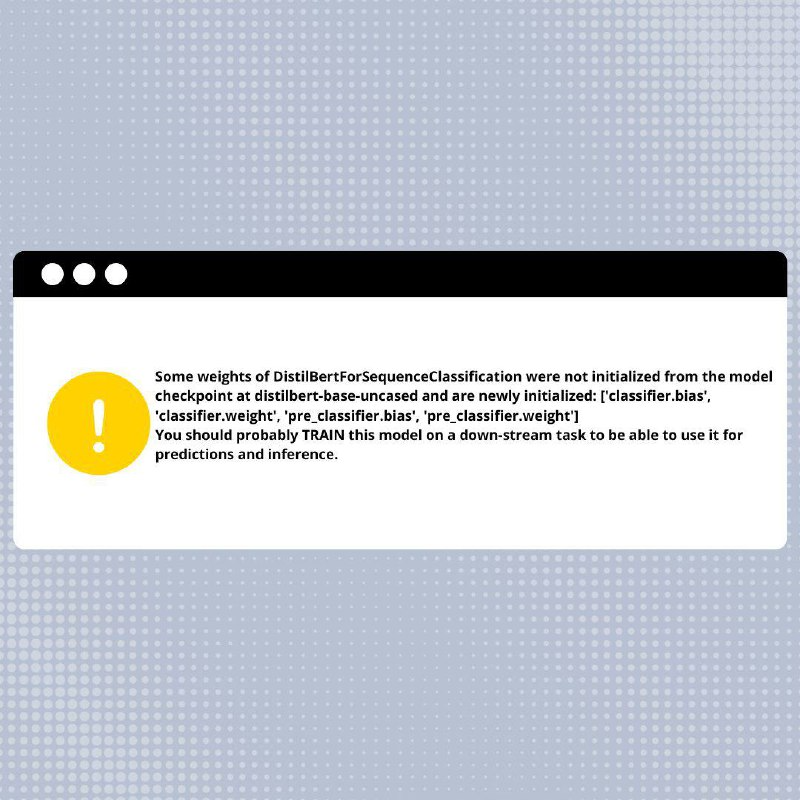
1. DistilBERT-base-uncased emotion:

* ***Strengths:***

This model is designed to detect a wide range of emotions, which can provide more granular sentiment analysis beyond simple positive, negative, or neutral classifications.  
It showed a moderate performance in identifying neutral sentiments, which suggests its potential utility in tasks requiring fine-grained emotion detection.

* ***Limitations:***

The model struggled with a high number of false positives and false negatives, particularly in positive sentiment detection, indicating a lower overall accuracy compared to the other models.  
Its general performance was less impressive compared to Twitter RoBERTa and BERT multilingual, which suggests it may need further fine-tuning for specific sentiment analysis tasks.

During experimentation, an important observation was noted: the following message was encountered, indicating some weights were not initialized from the model checkpoints: 

This message suggests that the model requires further training on a downstream task to optimize its weights for accurate predictions. This initial lack of fine-tuning could explain the higher error rates observed and underscores the necessity of additional training steps for enhanced performances.

1. BERT-base-multilingual-uncased-sentiment:

* ***Strengths:***  
  This model showed strong performance in classifying positive and negative sentiments, with a notable number of true positives.  
  Its multilingual capabilities allow it to be applied to sentiment analysis tasks across different languages without the need for retraining, making it a versatile option for global applications.  
  ***Limitations:***  
  The model had a higher number of false sentiments in neutral sentiment classification, indicating room for improvement in this area.  
  The complexity of handling multiple languages simultaneously might slightly affect its performance compared to models specialized in a single language.

### Model Performance in French Sentiment Analysis:

1. Twitter RoBERTa Base:

* ***Strengths:***

The Twitter RoBERTa model, although primarily trained on English Twitter data, managed to capture some sentiments in the French dataset. This can be attributed to the similarity in social media language structures across different languages, such as the use of emoticons and common abbreviations.

* ***Limitations:***

The model showed limitations in accurately classifying sentiments due to its lack of training on French text. The results indicated a lower accuracy, especially in identifying negative sentiments, which suggests the model's specialization in English text might hinder its performance in other languages.

1. DistilBERT-base-uncased emotion:

* ***Strengths:***

The Twitter RoBERTa model, although primarily trained on English Twitter data, managed to capture some sentiments in the French dataset. This can be attributed to the similarity in social media language structures across different languages, such as the use of emoticons and common abbreviations.

* ***Limitations:***

Similar to Twitter RoBERTa, DistilBERT struggled with the French dataset due to its training on English text. The model's lower accuracy in correctly identifying positive and negative sentiments highlights its need for fine-tuning on French-specific data.

1. BERT-base-multilingual-uncased-sentiment:

* ***Strengths:***  
  The model's multilingual capabilities shine in this context, demonstrating reliable sentiment classification in French text. Its ability to handle multiple languages without retraining is a significant advantage.
* ***Limitations:***  
  Although effective, the model may still face challenges with idiomatic expressions and linguistic nuances specific to each language.

### Comparative Analysis:

Our experiments yielded average accuracy values for each sentiment analysis model across both English and French datasets. Comparing these results with the findings reported in the literature provides valuable insights into the performance variations and potential limitations of the models.

* ***English Dataset:***  
  For the English dataset, our experiments revealed an average accuracy of 76% for the Twitter RoBERTa model, 31.67% for DistilBERT, and 50.33% for the BERT multilingual model.

Notably, our experiment's average accuracy with the BERT multilingual model was approximately 17.50% lower than the accuracy reported in the literature.

These differences may stem from variations in data preprocessing, model fine-tuning, or other experimental factors, warranting further investigation.

* ***French Dataset:***  
  Similarly, for the French dataset, our experiments demonstrated an average accuracy of 12% for the Twitter RoBERTa model, 15.075% for DistilBERT, and 70.395% for the BERT multilingual model.

In this case, our experiment's average accuracy with the BERT multilingual model was approximately 25.84% lower than the accuracy reported in the literature.

It's important to note that both the Twitter RoBERTa and DistilBERT models were primarily trained on English text and do not have the capability to work effectively on languages other than English. Therefore, their performance on the French dataset might was expected to be bordering on 0%.

By understanding these models' strengths and limitations, practitioners can make informed choices based on the specific requirements of their sentiment analysis tasks, leading to more accurate and reliable outcomes in real-world applications.

## **Conclusion:**

This research presents a comprehensive evaluation of three sentiment analysis models: Twitter RobERTa Base, DistilBERT-base-uncased emotion, and BERT-base-multilingual-uncased-sentiment. Through experimentation across varied datasets involving English and French languages, we have uncovered key insights into the performance, strengths, weaknesses and limitations of each model.

The Twitter RoBERTa Base model demonstrated superior performance in domain-specific contexts, particularly with social media text, where its training on extensive Twitter data proved advantageous. However, its domain specificity also limits its generalizability to other text types.

DistilBERT-base-uncased emotion model capable of detecting a wide range of emotions, providing granular sentiment analysis beyond simple classifications. High false positive and false negative rates in both datasets. Further fine-tuning is required, as evidenced by the message during experimentation about uninitialized weights.

BERT-base-multilingual-uncased-sentiment excelled in its multilingual capabilities, providing consistent performance across both English and French datasets. This model's strength lies in its ability to generalize across languages, making it a powerful tool for sentiment analysis in diverse linguistic contexts. However, it exhibited slightly lower accuracy compared to models tailored to specific languages or domains.  
  
The comparative analysis underscores the importance of model selection based on specific application requirements. For domain-specific tasks, such as analyzing social media sentiment, Twitter RoBERTa Base is the preferred choice. For applications demanding quick processing and lower resource use, DistilBERT-base-uncased emotion is recommended. For multilingual and cross-linguistic sentiment analysis, BERT-base-multilingual-uncased-sentiment stands out as the most versatile option.  
  
Future research should focus on enhancing model adaptability through hybrid approaches, combining the strengths of domain-specific and multilingual models. Additionally, fine-tuning models on more diverse and representative datasets will likely improve their robustness and accuracy. Exploring model will also be crucial in building trust and reliability in sentiment analysis systems.

In conclusion, this study not only highlights the current capabilities and limitations of state-of-the-art sentiment analysis models but also sets the stage for future advancements. By aligning model choice with application-specific needs and continuing to innovate, the field of sentiment analysis can achieve greater accuracy, efficiency, and applicability across various domains and languages

## **References**

|  |  |
| --- | --- |
| [1] | A. C. S. R. &. C. K. Mayur Wankhade, "A survey on sentiment analysis methods, applications, and challenges," 07 February 2022. [Online]. Available: https://doi.org/10.1007/s10462-022-10144-1. |
| [2] | S. R. Goniwada, " Sentiment Analysis," Apress, Berkeley, CA, 28 June 2023. [Online]. Available: https://doi.org/10.1007/978-1-4842-9496-3\_6. |
| [3] | G. Regkas, "Leveraging on NLP to gain insights in Social Media, News & Broadcasting," 7 April 2020. [Online]. Available: https://towardsdatascience.com/leveraging-on-nlp-to-gain-insights-in-social-media-news-broadcasting-ca89752ef638. |
| [4] | A. Balahur, "Sentiment analysis in social media texts," June 2013. [Online]. Available: https://aclanthology.org/W13-1617.pdf. |
| [5] | K. A. S. P. E. C. A. H. Ankita Gandhi, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," Information Fusion, March 2023. [Online]. Available: https://doi.org/10.1016/j.inffus.2022.09.025. |
| [6] | W. H. Brian Dickinson, "Sentiment Analysis of Investor Opinions on Twitter," Social Networking, 2015. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=57841. |
| [7] | A. K. K. K. &. S. S. Diksha Khurana, "Natural language processing: state of the art, current trends and challenges," *Springer Nature,* no. https://doi.org/10.1007/s11042-022-13428-4, 2023. |
| [8] | R. &. S. P. &. K. A. &. T. A. &. M. A. Khan, "Social media analysis with AI: Sentiment Analysis Techniques for rhe Analysis of Twitter Covid-19 Data," August 2020. |
| [9] | A. De, "Deep Learning: Subfield of Machine Learning". |
| [10] | S. A. a. A. Mahmoud, "A Framework for Designing the Architectures of Deep Convolutional Neural Networks," MDPI, 24 May 2017. [Online]. Available: https://doi.org/10.3390/e19060242. |
| [11] | C. M. Bishop, "Chapter 9: Neural Networks," in *Pattern Recognition and Machine Learning. Springer.*, 2006, p. 758. |
| [12] | J. Brownlee, "A Gentle Introduction to the Rectified Linear Unit (ReLU)," *Machine Learning Mastery,* no. https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/, 2020. |
| [13] | K. Rita, "CNN Sentiment Analysis," *Towards Data Science,* 2019. |
| [14] | Y. B. a. A. C. Ian Goodfellow, "Chapter 10: Sequence Modeling: Recurrent and Recursive Nets," in *Deep Learning*, MIT Press, 2016, pp. 377-403. |
| [15] | D. J. a. J. H.Martin, " Chapter 9: "Recurrent Neural Networks"," in *Speech and Language Processing*, Pearson, 2019, pp. 371-392. |
| [16] | J. W. R. F. e. a. Sainbayar Sukhbaatar, "A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015. |
| [17] | A. H. J. A. S.-C. L. Yuhuang Hu, "Overcoming the vanishing gradient problem in plain recurrent networks," *ArXiv,* no. https://api.semanticscholar.org/CorpusID:31839991, 18 January 2018. |
| [18] | Y. B. a. A. C. Ian Goodfellow, Deep Learning, MIT Press, 2016. |
| [19] | Y. B. a. G. H. Yann LeCun, "Deep Learning," *Nature,* vol. 521, no. 7556, pp. 436-444, 2015. |
| [20] | S. B. M. H. B. R. a. O. V. Authors: C. Zhang, "Understanding deep learning requires rethinking generalization.," in *5th International Conference on Learning Representations (ICLR)*, 2017. |
| [21] | S. H. R. H. X. Z. Hong Chen, "Improved naive Bayes classification algorithm for traffic risk management," *EURASIP Journal on Advances in Signal Processing,* no. https://doi.org/10.1186/s13634-021-00742-6, 2021. |
| [22] | A. Yadav, "Support Vector Machines(SVM)," *Towards Data Science,* no. https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589, 2018. |
| [23] | "Twitter RoBERTa-base for sentiment Analysis," CardiffNLP, [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment. |
| [24] | J. C.-C. L. E. A. a. L. N. Francesso Barbieri, Findings of the Association for Computational Linguistics: EMNLP 2020, vol. Findings of the Association for Computational Linguistics: EMNLP 2020, Y. H. Y. L. Trevor Cohn, Ed., Association for Computational Linguistics, 2020, pp. 1644--1650. |
| [25] | J. C.-C. L. E.-A. L. N. Francesco Barbieri, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". |
| [26] | "distilbert-base-uncased," Hugging Face Model Hub, [Online]. Available: https://huggingface.co/distilbert/distilbert-base-uncased. |
| [27] | Nlptown, "Bert-base-multilingual-uncased-sentiment: A multilingual BERT-based model for sentiment analysis.," [Online]. Available: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment. |
| [28] | M.-W. C. K. L. K. T. Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for," 2019. |
| [29] | S. NLP, "BERT Sequence Classifier for Multilingual," 2021. |
| [30] | C. M. Bishop, in *Pattern Recognition and Machine Learning*, Springer, 2006, p. 758. |
| [31] | B. Savani, "A DistilBERT-based model for emotion classification," Hugging Face Model Hub, [Online]. Available: https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion?text=I+feel+a+bit+let+down. |