Analyzing Sentiment Across Languages: Comparing ML Models on French and English Amazon Reviews

Shamim Mahmoudzadeh Vaziri
Department of Mathematics and Industrial Engineering
Polytechnique Montreal, Canada
shamim.mahmoudzadeh-vaziri@polymtl.ca
11353631

Maria Camila Gomez Lopez
Département d'analytique, opérations et technologies de l'information
Université du Québec à Montréal, Canada
gomez_lopez.maria_camila@courriel.uqam.ca
11360206

Louis Fillon
Département des Sciences de la décision, Technologies de l'information et marketing
HEC Montréal, Canada
louis.fillon@hec.ca
11253508

Mahmoud Moubarak
Department of Mathematics and Industrial Engineering
Polytechnique Montreal, Canada
mahmoud.moubarak@polymtl.ca
11364954

Abstract

In today's globalized digital economy, understanding customer sentiment across multiple languages is crucial for personalized and culturally aware services. This study explores multilingual sentiment analysis by comparing English and French Amazon product reviews. We preprocess a subset of the Multilingual Amazon Reviews Corpus (MARC) and map star ratings to negative, neutral, and positive categories. Consistent vectorization techniques—Bag-of-Words, TF-IDF, and Word2Vec—are applied to both languages. Models include Naive Bayes, SVM, Random Forests, LSTM, and transformer-based models like multilingual BERT. Transformer-based models outperform traditional ones in accuracy, but performance varies with linguistic features and vectorization. This study highlights the promise and limitations of current approaches and supports future research in cross-lingual sentiment detection.

1 Introduction

Nowadays, people are constantly sharing their opinions in many different languages on websites, social media, and product reviews. For companies that work in different countries, understanding customer satisfaction in multiple languages is important in order to improve their services, create better products, increase sales, and protect their brand reputation. However, reading and analyzing all these reviews manually is not practical due to the high volume of information. This has led to the use of Machine Learning (ML) models to perform sentiment analysis (SA) and help classify opinions quickly and efficiently.

Moreover, evaluating sentiment across languages contributes not only to improved customer experience but also supports inclusive AI development. Ensuring that sentiment analysis tools perform equitably across languages helps mitigate bias in global applications, allowing businesses and researchers to make fairer and more informed decisions. This becomes especially critical in an era where AI systems are expected to respect linguistic and cultural diversity Levy et al. [2023] .

Sentiment analysis has become an essential tool in natural language processing (NLP), which focuses on automatically identifying and categorizing opinions expressed in text, typically into three classes: positive, neutral, and negative. While sentiment analysis has achieved high accuracy in monolingual

Preprint. Under review.

Rating	French_Review	English_Translation
5	C'est exactement ce que je voulais.	This is exactly what I wanted.
1	Très mauvais produit. Il ne fonctionne pas du tout.	Very bad product. It doesn't work at all.
3	Produit correct, mais la livraison a été lente.	Decent product, but delivery was slow.
4	Fonctionne bien, mais la notice est uniquement en chinois.	Works fine, but the manual is only in Chinese.
2	Pas satisfait, la qualité n'est pas au rendez-vous.	Not satisfied, the quality is not up to expectations.

Figure 1: Sample of the Dataset

contexts, particularly in English, the extension to multilingual settings introduces new challenges. Linguistic structures, cultural differences, and variations in sentiment expression across languages can significantly impact the performance of sentiment analysis modelsNazir et al. [2025].

Several strategies have been proposed in the literature for this task. Traditional machine learning models like Naïve Bayes and Support Vector Machines (SVM) have been adapted for multilingual data Pang et al. [2002], Fang and Zhan [2015b], while deep learning approaches, such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), have shown good results due to their ability to capture syntactic and semantic structures from raw text Singh and Singla [2023], Ouyang et al. [2015]. Transformer-based models such as BERT and its derivatives leverage pre-training on large multilingual data to learn common representations across different languages, enabling them to perform sentiment classification without the need for explicit translation Conneau et al. [2020], Devlin et al. [2019]. Despite these advances, some key challenges remain in this multilingual sentiment analysis. As highlighted by Mao et al. [2024], accurately detecting implicit sentiment, sarcasm, and culturally-specific expressions remains difficult.

In this project, we aim to compare the performance of multiple machine learning models in capturing sentiment across different languages. By analyzing a multilingual dataset of Amazon product reviews in English and French, we explore how linguistic and cultural differences impact model performance. By systematically comparing model performance on both English and French datasets, this project provides valuable insights into the adaptability of different machine learning approaches in multilingual environments and lays the groundwork for scalable, language-aware sentiment classification solutions. Section 2 describes the dataset and preprocessing steps, Section 3 outlines the Machine Learning methods we proposed for this project, Section 4 presents the results, and Section 5 concludes the report with insights and possible future work.

2 Dataset and Preprocessing

In this project, we use a dataset from Kaggle titled "French reviews on Amazon items and EN translation" Kaggle User: dargolex [2025]. This dataset is part of the Multilingual Amazon Reviews Corpus (MARC) Amazon Science [2025], a large collection of Amazon product reviews in several languages, including English, Japanese, German, French, Spanish, and Chinese. The reviews were collected between 2015 and 2019.

The dataset contains 200,000 reviews in total, with 40,000 reviews for each rating from 1 to 5. Fig. 1 shows a sample from the dataset to illustrate its format and structure of the French reviews, their English translations, and rating labels. It is also important to note that the translations were generated using the Google Translate API.

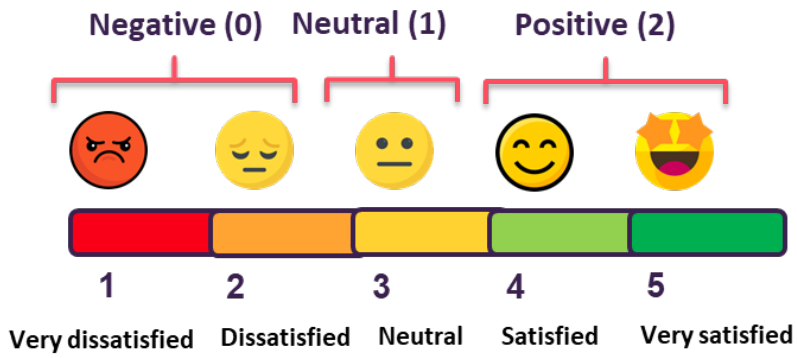

Figure 2: Label Transformation

The dataset used for ML models is prepared through several preprocessing steps before applying vectorization methods, which are covered in the next section. The steps are as follows:

- Data Reduction: Because training machine learning models on a large dataset can be time-consuming and need powerful GPUs, we reduce the number of reviews in each rating class. The resulting dataset contains 5,000 reviews each for ratings 1, 2, 4, and 5, and 10,000 reviews for rating 3.
- Label Transformation: Since discovering the difference between two adjacent rating scores can be difficult for models, we convert 5-star ratings into three sentiment categories: negative (0), neutral (1), and positive (2). This simplification helps the models perform better. Fig. 2 illustrates this transformation.
- Dataset Splitting: The dataset is divided into three parts — 70% for training, 15% for validation, and 15% for testing.

Then, we apply several text preprocessing steps to clean and standardize the reviews as follows:

- Removing stop words (e.g., the, is, in, et, le, la))
- Lemmatizing (convert words to their base form)
- Tokenization (split text into words or phrases)
- lowercasing

3 Methodology

In this section, we describe the vectorization techniques, machine learning models, and evaluation metrics employed in our study. The goal was to evaluate the performance of various algorithms for multilingual sentiment classification using both traditional and deep learning approaches, applied to preprocessed French and English Amazon product reviews.

3.1 Vectorization Techniques

In this step, we apply three techniques to convert the text reviews into numerical representations, which are more understandable for ML models:

- Bag-of-Words (BoW): Counts occurrences of words in texts
- Term Frequency-Inverse Document Frequency (TF-IDF): Assigns weights to words based on their importance in a text relative to the entire collection of texts.
- Word2Vec: A neural network-based approach, capturing semantic relationships between words by placing them in a vector space. Kaggle Notebook by suvroo [2025]

English and French have different vocabularies. For example, the English word "computer" becomes "ordinateur" in French, and French also includes accents like é and à. As a result, we apply these techniques separately to the English and French reviews. Additionally, we use Word2Vec only for the English reviews, as training Word2Vec on French text is very time-consuming.

3.2 Machine Learning Models

3.2.1 Baseline Model

We define the K-Nearest Neighbors (KNN) model with TF-IDF vectorization as our baseline. In this model, we set the number of neighbors k = 5. The baseline accuracy is 0.497 for English reviews and 0.522 for French reviews.

3.2.2 Naive Bayes

For this method, we implement both Complement Naive Bayes and Multinomial Naive Bayes models. These models are well-suited for our task because, after vectorization, the data is represented using discrete features. Naive Bayes models are commonly used as basic machine learning models for text classification tasks like sentiment analysis Analytics Vidhya [2025].

3.2.3 Support Vector Machine (SVM)

(SVM) model was implemented using linear and RBF kernels across the vectorization techniques: Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and Word2Vec embeddings. Initially, hyperparameter tuning for the regularization parameter C and kernel type was performed using GridSearchCV. However, due to high computational costs, we switched to a manual grid search by iteratively testing predefined values. Its robustness in handling high-dimensional and sparse text data has made it a widely adopted model in sentiment analysis, as supported by previous research Guia et al. [2019] Gope et al. [2022] Hawlader et al. [2021]

3.2.4 Random Forest (RF)

We applied Random Forest on both English and French reviews using the previous vectorization techniques, and Word2Vec embeddings. As an ensemble learning method, Random Forest builds multiple decision trees and combines their outputs using majority voting for classification. Its use of bagging and feature randomness promotes diversity among trees and helps reduce overfitting Gope et al. [2022] Hawlader et al. [2021] Karthika et al. [2019].The hyperparameter tuning was conducted using a manual grid search approach, adjusting key parameters such as n-estimators, max-depth, and min-samples-split. This decision was driven by the computational cost of using GridSearchCV on our large multilingual dataset.

3.2.5 Long Short-Term Memory Neural network (LSTM)

An LSTM was evaluated using Word2Vec embeddings for vectorization. These embeddings transformed text into semantic numerical representations within the LSTM's embedding layer. We also buffed it with a bidirectional component. Bidirectional LSTMs processed input sequences in both forward and backward directions, capturing contextual dependencies. Final sentiment classification was determined through majority voting across model instances. Hyperparameters such as embedding size, dense layers, dropout rate, activation functions, and optimizer were tuned manually with a grid search for performance optimization and regularization. (TensorFlow [2024], Scikit-learn [n.d.])

3.2.6 BERT

We used the `nlptown/bert-base-multilingual-uncased-sentiment` model (nlptown [2020]), a pre-trained model, fine-tuned specifically for sentiment analysis tasks on product reviews. This model supports multiple languages, including English, French, Spanish, German, and Italian, making it well-suited for our multilingual sentiment analysis project.

The model is based on BERT (Bidirectional Encoder Representations from Transformers), a transformer-based architecture introduced by Devlin et al. [2019]. BERT differs from other models such as RNNs and LSTMs, because it uses a deep bidirectional transformer encoder. This allows the

model to consider the full context of a word based on both its left and right information simultaneously, instead of processing text in one direction. This helps the model to understand the meaning of each word based on its full context, which improves its ability to capture relationships between words.

BERT is trained using a technique called Masked Language Modeling (MLM). In this method, some words in a sentence are randomly replaced with a MASK token, and the model learns to predict the missing words. This helps BERT learn deep representations of language in a self-supervised way, without needing manually labeled data.

The `nlptown` model we used classifies input texts into five sentiment categories: 1 to 5 stars. For consistency with our experimental setup, we map these star ratings to three sentiment classes: negative (1–2 stars), neutral (3 stars), and positive (4–5 stars). For the final results, we used the model's tokenizer without additional preprocessing like removing punctuation or lowercasing, because BERT is trained on raw text and modifying the input text actually lead to a slightly worse performance.

Finally, we evaluated the model on both English and French reviews. Since the model is multilingual, we used the same architecture for both languages without retraining. This allowed us to directly compare performance across languages in a consistent way without developing separate models.

3.3 Evaluation Metrics

To evaluate model performance, we used four standard classification metrics: accuracy, precision, recall, and F1-score. Given the presence of class imbalance across sentiment categories (especially the neutral class), we computed macro-averaged scores to ensure fair comparison across classes. In addition to numerical evaluation, we analyzed confusion matrices to understand class-wise misclassifications and language-specific sentiment shifts. These metrics and visualizations allowed us to draw meaningful comparisons and assess the impact of language and vectorization technique on model performance.

4 Results

This section presents the performance results of the machine learning models evaluated on both English and French datasets. We report standard classification metrics, including accuracy, precision, recall, and F1 score, to compare the effectiveness of each approach in the multilingual sentiment analysis task.

Table 1 and Table 2 summarize the performance of all evaluated models on the English and French datasets, respectively. The results allow us to compare the effectiveness of traditional machine learning models, deep learning models, and transformer-based models across both languages.

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes (BoW)	0.5900	0.5900	0.5900	0.5900
SVM	0.5369	0.5462	0.5369	0.5393
Random Forest	0.5373	0.5366	0.5373	0.5369
LSTM	0.6850	0.4570	0.5700	0.5070
BERT	**0.6970**	**0.6900**	**0.7000**	**0.6800**

Table 1: Performance of all models on the English dataset.

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes (BoW)	0.6167	0.6100	0.6200	0.6200
SVM	0.6427	0.6451	0.6427	0.6436
Random Forest	0.6216	0.6157	0.6216	0.6155
LSTM	0.7000	0.6660	0.5830	0.5210
BERT	**0.7260**	**0.7200**	**0.7300**	**0.7200**

Table 2: Performance of all models on the French dataset.

From the graph, it is immediately noticeable that across all models, sentiment prediction accuracy is consistently higher for the original French dataset compared to the English dataset, which was

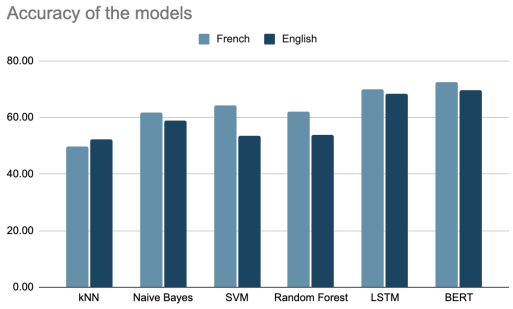

Figure 3: Graph of all models' accuracies in English and French.

generated through translation. Specifically, the highest accuracy is achieved by the BERT model, closely followed by the LSTM model, highlighting the strengths of advanced deep learning methods and transformer-based architectures in multilingual sentiment analysis. Traditional machine learning methods (Naïve Bayes, SVM, and Random Forest) generally perform less effectively compared to the deep learning models. Among these, SVM and Random Forest show a more pronounced difference in accuracy between French and English, suggesting they are more sensitive to linguistic nuances lost during translation. Naïve Bayes shows slightly lower accuracy overall, though still higher for French data, underscoring the challenges traditional statistical models face in handling multilingual and translated datasets. The smaller performance gap between French and English in the BERT and LSTM models indicates that these deep learning techniques are more robust in capturing context, semantics, and subtleties in translated text. However, despite their advanced capabilities, even BERT shows a discernible accuracy advantage for the French language data, emphasizing the importance of original context and vocabulary richness.

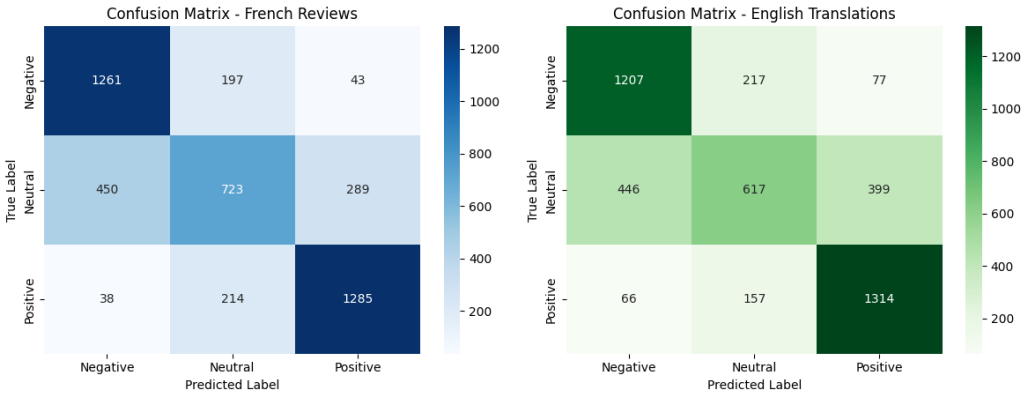

Figure 4: Confusion matrices for the BERT model: French (left) and English (right). Most misclassifications occur in the neutral class.

Among all the models evaluated, BERT outperforms the others in both English and French across all evaluation metrics. Figure 4 presents the confusion matrices for the BERT model on English and French, which show a strong performance in predicting positive and negative sentiment, but some difficulty in correctly identifying neutral sentiment (class 1). This pattern is also reflected in the class-wise precision, recall, and F1-score values shown in tables 3 and 4, respectively.

Class	Precision	Recall	F1-Score
Negative (0)	0.7022	0.8041	0.7497
Neutral (1)	0.6226	0.4220	0.5031
Positive (2)	0.7341	0.8549	0.7899

Table 3: BERT class-wise performance metrics on the English dataset.

Class	Precision	Recall	F1-Score
Negative (0)	0.7210	0.8401	0.7760
Neutral (1)	0.6376	0.4945	0.5570
Positive (2)	0.7947	0.8360	0.8148

Table 4: BERT class-wise performance metrics on the French dataset.

In both English and French, BERT achieved a good performance for negative and positive reviews (class 0 and 2). However, the model struggled more on neutral reviews (class 1), with a recall of 0.422 and F1-score of 0.503 for English, and a recall of 0.495 and F1-score of 0.557 for French, indicating that many neutral reviews were misclassified as either positive or negative. These results are consistent with common challenges in sentiment analysis, where neutral sentiment often lacks strong lexical signals and is more context-dependent.

To better understand the model's difficulties with the neutral sentiment, we conducted a brief lexical analysis of the reviews that were incorrectly classified as neutral but actually belonged to either the positive or negative class. We extracted the top 10 most frequent words in these misclassified reviews for both English and French.

As shown in Table 5, the most frequent words in reviews misclassified as neutral are mostly very common grammatical words such as "the", "to", and "is" in English, or "de", "le", and "la" in French. These words are widely used in all types of sentences and generally do not have emotional meanings. However, some more meaningful terms like "not" and "but" in English, or "pas" and "mais" in French, also appear among the top 10. These words can show contrast or negation in a sentence, potentially leading to confusion for the model. For example, the word "but" often signals a contradiction or shift in sentiment, which may confuse the model when determining the overall sentiment of the review. This suggests that many of these misclassified reviews contain subtle or conflicting opinions, making them harder to categorize.

Rank	English Word	French Word
1	the (768)	de (406)
2	to (314)	le (347)
3	a (309)	la (303)
4	is (265)	un (265)
5	not (248)	pour (258)
6	but (218)	pas (257)
7	and (199)	mais (251)
8	it (199)	et (232)
9	of (184)	à (201)
10	i (182)	les (194)

Table 5: Top 10 most frequent words in reviews misclassified as neutral in English and French.

We also examined the inverse case: reviews whose true sentiment was neutral but were incorrectly predicted as either positive or negative. We found that terms like "but", "not", and "very" in English, as well as "mais", "pas", and "très" in French, frequently appeared in these misclassified samples. Although these reviews were labeled as neutral, the presence of those words could have introduced ambiguity, leading the model to assign a positive or negative sentiment.

5 Conclusion

5.1 Selection of the Model

The BERT (Bidirectional Encoder Representations from Transformers) model demonstrated superior performance on the multilingual sentiment analysis dataset utilized in this project. Among the various machine learning approaches evaluated, including traditional models and other deep learning techniques, BERT's accuracy and consistency were notably higher. This good performance can be attributed to several intrinsic features and strengths of the BERT model.

A critical factor behind BERT's effectiveness is its use of the self-attention mechanism, which allows the model to weigh the importance of each word relative to every other word in a given sentence. This mechanism provides contextual understanding by capturing the relationships between words irrespective of their positions. Unlike traditional sequential processing methods, self-attention enables BERT to comprehend intricate linguistic patterns and long-range dependencies, significantly enhancing its ability to accurately interpret the sentiment expressed in complex sentences.

Furthermore, BERT benefits immensely from extensive pre-training. It undergoes training on large corpora of text from diverse sources, enabling it to acquire general language understanding before being fine-tuned on specific tasks such as sentiment analysis. This extensive pre-training equips BERT with a robust foundation in language semantics and syntax, ensuring it can generalize well even when applied to smaller, task-specific datasets.

Another distinguishing capability of BERT is its nuanced specification of linguistic subtleties such as idiomatic expressions and sarcasm. Multilingual sentiment analysis often encounters challenges arising from cultural and linguistic differences, particularly in capturing nuances that drastically alter sentiment interpretation. BERT's deep contextual embedding allows it to identify subtle linguistic cues effectively, thus accurately capturing expressions and sarcastic remarks that could otherwise be misinterpreted by more straightforward models.

The underlying Transformer architecture also significantly contributes to BERT's superior performance. Unlike recurrent neural networks (RNNs), transformers do not suffer from vanishing or exploding gradient issues and can process input data more efficiently through parallelization. The architecture's ability to analyze data bidirectionally ensures comprehensive contextual understanding, further enhancing the accuracy and reliability of sentiment analysis across languages. (Goodfellow et al. [2016], Fang and Zhan [2015a])

5.2 Conclusion to our research question

Based on our experiments, sentiment prediction accuracy is higher for French data compared to English. This observed difference in performance can be primarily attributed to the original nature of the dataset, which was initially collected in French. The English data, in contrast, was obtained through direct translations from French using the Google Translate API. This translation process introduces a layer of ambiguity, often altering the nuance and context of the original sentiments.

To answer our initial question, we conclude that Sentiment predictions indeed vary significantly depending on the language, influenced heavily by the richness and complexity inherent to each language. French, for instance, possesses a rich vocabulary with numerous synonyms and nuanced expressions, providing clearer indications of positive or negative sentiment. Conversely, English vocabulary can be ambiguous due to words having multiple meanings, potentially shifting the sentiment contextually from positive to negative.

Language-specific features such as sarcasm, slang, and idiomatic expressions further complicate multilingual sentiment analysis. These linguistic nuances require deep contextual understanding, posing considerable challenges to sentiment models that may lack comprehensive training data or methodologies capable of capturing such subtleties.

Technological limitations also impede multilingual sentiment analysis. Detection of negation is a notably complex task, as negations can drastically alter sentiment but are represented differently across languages. Additionally, comprehensive multilingual datasets capturing slang, sarcasm, emojis, special characters, and various punctuation forms are limited, hindering the model's ability to generalize effectively across languages. Furthermore, preprocessing such diverse linguistic data is computationally intensive and resource-demanding, particularly when dealing with extensive multilingual corpora.

Ultimately, current models, including BERT, still lack sufficient training data and specialized methods required to be fully versatile across diverse languages. Addressing these limitations by expanding datasets, refining preprocessing methods, and improving negation detection will be essential for enhancing model performance in multilingual sentiment analysis tasks.

6 Future Work

To achieve higher accuracy in multilingual sentiment analysis, future work can focus on improving current methods in several key areas.

Fine-tuning existing models such as BERT remains a crucial step toward enhancing cross-lingual performance. By systematically adjusting hyperparameters—such as learning rates, regularization strengths, kernel types, embedding dimensions, and dropout rates—models can be better optimized for specific linguistic characteristics and dataset nuances.

Equally, exploring new machine learning and deep learning architectures also presents an opportunity for innovation. Techniques like Transformer variants, attention-based RNNs, and hybrid models that combine statistical and neural methods may offer better handling of syntactic and semantic complexities inherent in multilingual sentiment tasks.

One of the most challenging aspects of sentiment analysis that has to be developed is expanding dataset scale and diversity. Larger volumes of labeled data across various languages, dialects, and cultural contexts help reduce bias and enhance adaptability to real-world linguistic variation. Notably, the SEMEVAL (Semantic Evaluation) series has contributed significantly through its sentiment-related shared tasks, offering high-quality, multilingual datasets focused on Twitter sentiment, aspect-based sentiment, and informal language, which have become benchmarks for evaluating model performance on noisy, real-world data. Similarly, the Amazon Multilingual Reviews Corpus is one of the most comprehensive sentiment datasets to date, comprising millions of product reviews in multiple languages with rich metadata and domain diversity. However, despite their value, these datasets still lack sufficient representation of low-resource languages, code-switching, non-Western cultural contexts, and multimodal sentiment cues (e.g., combining text with images or voice). They also often rely on star ratings as sentiment proxies, which may not fully capture nuanced emotional tone or user intent. Addressing these gaps is crucial for developing sentiment analysis models that are not only scalable but also fair, inclusive, and better aligned with global real-world communication. Amazon Science [2025], sem [2024]

Enhancing translation quality and improving the model's language understanding capabilities is another critical direction. Translated data often introduces errors or loses subtle sentiment indicators like idioms, cultural references, or emotional tone. Incorporating more accurate neural machine translation (NMT) tools and language-specific preprocessing techniques can help retain sentiment fidelity. Furthermore, training models to handle informal expressions such as slang, abbreviations, emojis, and code-switching can significantly boost performance in user-generated content like social media posts or customer reviews.

Lastly, Multimodal sentiment analysis, which combines textual data with additional modalities such as images, audio, or video, can provide richer context. For example, pairing written product reviews with associated product photos or video testimonials can help disambiguate unclear sentiments and offer deeper emotional cues that text alone may not convey. Mao et al. [2024]

References

Semeval-2024: Semantic evaluation. https://semeval.github.io/SemEval2024/, 2024. Accessed: 2025-04-18.

Amazon Science. The multilingual amazon reviews corpus. https://www.amazon.science/publications/the-multilingual-amazon-reviews-corpus, 2025. Accessed: 2025-04-18.

Analytics Vidhya. Naive bayes algorithms: A complete guide for beginners. https://www.analyticsvidhya.com/blog/2023/01/naive-bayes-algorithms-a-complete-guide-for-beginners/, 2025. Accessed: 2025-04-18.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, 2020. doi: 10.18653/v1/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):5, 2015a. doi: 10.1186/s40537-015-0015-2. URL https://doi.org/10.1186/s40537-015-0015-2.

Xueke Fang and Jun Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(5), 2015b. doi: 10.1186/s40537-015-0015-2. URL https://doi.org/10.1186/s40537-015-0015-2.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org. Chapter 10 consulted.

Joy Chandra Gope, Tanjim Tabassum, Mir Md Mabrur, Keping Yu, and Mohammad Arifuzzaman. Sentiment analysis of amazon product reviews using machine learning and deep learning models. In *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pages 1–6. IEEE, 2022.

Márcio Guia, Rodrigo Rocha Silva, and Jorge Bernardino. Comparison of naïve bayes, support vector machine, decision trees and random forest on sentiment analysis. *KDIR*, 1:525–531, 2019.

Mohibullah Hawlader, Arjan Ghosh, Zaoyad Khan Raad, Wali Ahad Chowdhury, Md Sazzad Hossain Shehan, and Faisal Bin Ashraf. Amazon product reviews: Sentiment analysis using supervised learning algorithms. In *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, pages 1–6. IEEE, 2021.

Kaggle Notebook by suvroo. Complete nlp pipeline. https://www.kaggle.com/code/suvroo/complete-nlp-pipeline#notebook-container, 2025. Accessed: 2025-04-18.

Kaggle User: dargolex. French reviews on amazon items and english translation. https://www.kaggle.com/datasets/dargolex/french-reviews-on-amazon-items-and-en-translation, 2025. Accessed: 2025-04-18.

P Karthika, R Murugeswari, and R Manoranjithem. Sentiment analysis of social media network using random forest algorithm. In *2019 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)*, pages 1–5. IEEE, 2019.

Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. Comparing biases and the impact of multilingual training across multiple languages, 2023. URL https://arxiv.org/abs/2305.11242.

Yanying Mao, Qun Liu, and Yu Zhang. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, 36(4):102048, 2024. ISSN 1319-1578. doi: 10.1016/j.jksuci.2024.102048. URL https://doi.org/10.1016/j.jksuci.2024.102048. Retrieved February 12, 2025.

Muhammad Kashif Nazir, CM Nadeem Faisal, Muhammad Asif Habib, and Haseeb Ahmad. Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages. *IEEE Access*, 2025.

nlptown. nlptown/bert-base-multilingual-uncased-sentiment. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment, 2020. Accessed: 2025-04-18.

Xi Ouyang, Pan Zhou, Cheng Hua Li, and Lijun Liu. Sentiment analysis using convolutional neural network. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 2359–2364, 2015. doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.349.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86. Association for Computational Linguistics, 2002.

Scikit-learn. scikit-learn: machine learning in python — scikit-learn 1.6.1 documentation, n.d. URL https://scikit-learn.org/stable/index.html. Retrieved March 10, 2025.

Shubham Singh and Neetu Singla. Sentiment analysis on imdb review dataset. *Journal of Computers, Mechanical and Management*, 2:18–29, 2023. doi: 10.57159/gadl.jcmm.2.6.230108.

TensorFlow. Api documentation, September 30 2024. URL https://www.tensorflow.org/api_docs. Retrieved March 10, 2025.

A Appendix

A.1 BERT Model

As part of our experiments, we also evaluated the `cardiffnlp/twitter-xlm-roberta-base-sentiment` model, a RoBERTa-based model pre-trained and fine-tuned on sentiment analysis for Twitter data. This model was evaluated on the same dataset as the other models to explore if we could get better results.

However, as shown in Tables 6 and 7, this model underperformed when compared to the `nlptown/bert-base-multilingual-uncased-sentiment` model. In the English dataset, the CardiffNLP model achieved an accuracy of 57.02% and a macro-averaged F1-score of 0.5362. Even though it performed well on the positive class (F1-score: 0.7130), it struggled with the neutral class, achieving only 0.2560 in F1-score and a low recall of 0.1929.

Class	Precision	Recall	F1-Score	Support
Negative (0)	0.5295	0.8075	0.6396	1501
Neutral (1)	0.3806	0.1929	0.2560	1462
Positive (2)	0.7293	0.6975	0.7130	1537
Accuracy		0.5702		4500
Macro Avg	0.5464	0.5659	0.5362	4500
Weighted Avg	0.5493	0.5702	0.5400	4500

Table 6: Classification report for the CardiffNLP RoBERTa model on the English dataset.

Similarly, for French we got an accuracy of 58.38% and a macro F1-score of 0.4951. Also, the model showed strong performance for the positive class (F1-score: 0.7383), but its F1-score and recall for the neutral class were 0.0875 and 0.0499, respectively.

Class	Precision	Recall	F1-Score	Support
Negative (0)	0.5389	0.8501	0.6596	1501
Neutral (1)	0.3527	0.0499	0.0875	1462
Positive (2)	0.6639	0.8315	0.7383	1537
Accuracy		0.5838		4500
Macro Avg	0.5185	0.5772	0.4951	4500
Weighted Avg	0.5211	0.5838	0.5006	4500

Table 7: Classification report for the CardiffNLP RoBERTa model on the French dataset.

These results suggest that although the CardiffNLP model benefits from the RoBERTa architecture, its training on Twitter data does not generalize well to the structure and vocabulary of the product reviews in our dataset.

Additionally, we explored several text preprocessing techniques before passing the inputs into BERT, such as removing punctuation, special characters, and lowercasing text. These modifications resulted in a slightly worse performance across all classes. This is consistent with findings in the literature, which suggest that pre-trained transformer models like BERT are optimized to work with raw, unaltered text.

As a result, we decided to use the `nlptown/bert-base-multilingual-uncased-sentiment` model with no additional preprocessing, given its superior performance and better generalization across both languages and sentiment classes.