HEC MONTRÉAL

From Rants to Riches: A Review-Reading Pipeline for Multilingual Sentiment Analysis Using Neural Networks and Transformers

Inès Ben Hamza 11287022
Inès Battah 11351814

December 2023

Contents

1 Introduction

Customer reviews offer valuable insights into user experiences and expectations, shaping critical business areas such as product development, customer service, and marketing strategies. However, analyzing these reviews becomes increasingly challenging when they are written in multiple languages and collected at scale. Consider a global retailer launching a new product that receives thousands or even millions of reviews in diverse languages within weeks. While some reviews praise the product's quality, others highlight issues like delayed deliveries or product defects. Extracting actionable business insights from such multilingual feedback is essential for timely improvements, yet the complexity and volume of data can overwhelm traditional analytical methods. Rule-based systems like VADER, while computationally efficient, lack the sophistication to manage large-scale, nuanced, and multilingual datasets. These limitations often result in delayed analysis, hindering businesses' ability to respond promptly to customer concerns, thereby risking customer trust and satisfaction.

This project aims to address these challenges by developing a robust multilingual review analysis pipeline that seamlessly integrates sentiment analysis and topic modeling. By leveraging state-of-the-art machine learning techniques, particularly deep learning approaches such as transformer-based models and custom-built neural networks, the pipeline overcomes the limitations of traditional methods in handling large-scale, multilingual datasets. It enables effective processing and extraction of actionable insights from diverse customer reviews. Specifically, we focus on analyzing user-generated Amazon review datasets written in three Romance languages: English, French, and Spanish. This linguistic selection offers a manageable yet diverse starting point within the academic scope of this project, laying the groundwork for future generalizations to additional languages.

1.1 Problem Statement

Analyzing large-scale, multilingual customer reviews is difficult due to the complexity of nuanced sentiment, diverse linguistic contexts, and unstructured data. This project intends to mitigate this pain point for many marketers by building a comprehensive review analysis pipeline that incorporates sentiment analysis, topic modeling, and neural networks, offering actionable insights from Amazon reviews in English, French, and Spanish.

1.2 Objectives

The objectives of this study are threefold:

1. Improve sentiment classification accuracy using transformer-based models that can handle complex, multilingual datasets.
2. Extract actionable themes from customer feedback through advanced topic modeling techniques.
3. Develop a custom-built neural network tailored to domain-specific multilingual data, capable of capturing nuanced patterns in customer reviews.

Our goal extends beyond simply classifying customer feedback as positive, neutral, or negative sentiments (understanding how customers feel). We also aim to uncover the specific factors driving these sentiments (understanding why they feel this way).

By combining pre-trained transformer models with a custom-built neural network, this dual approach offers a powerful framework for businesses to reinforce successful practices and address critical pain points, ultimately improving customer satisfaction and trust.

1.3 Related Works

Recent advancements in multilingual sentiment analysis have leveraged transformer models and neural networks to address the challenges of analyzing multilingual and nuanced sentiment patterns. Studies like **Balahur and Turchi (2014)** explored machine translation for multilingual sentiment analysis, highlighting the potential of converting text into a single language before analysis, a principle relevant to our handling of multilingual Amazon reviews.

Pires et al. (2019) demonstrated the zero-shot capabilities of Multilingual BERT in generalizing across languages, a feature particularly applicable to our use of bert base models for review-specific sentiment classification.

Additionally, **Zhang et al. (2018)** provided a comprehensive survey on deep learning approaches, including RNNs and transformer models, validating their effectiveness in capturing nuanced sentiment patterns.

Our project builds on these foundations by combining sentiment analysis with topic modeling techniques (LDA, NMF, and LSA) to extract actionable themes, integrating methodologies from both advanced transformers and traditional NLP approaches to address the complexity of multilingual datasets.

2 Sentiment Analysis

2.1 Preprocessing

The original dataset is a collection of 1200000 Amazon reviews in English, Japanese, German, French, Spanish and Chinese. In our analysis we decided to keep the French, English and Spanish reviews as these languages are more accessible for us to understand. The dataset was initially balanced, containing an equal number of 1 to 5-star reviews across 30 categories. Due to computational constraints, we sampled the data by retaining only 5,000 reviews per star level for each language, resulting in a balanced dataset of 25,000 reviews per language. We then categorized the reviews based on their star ratings: negative reviews were classified as those with ratings from 1 to 2 stars, neutral reviews as those with a rating of 3 stars, and positive reviews as those with ratings from 4 to 5 stars. This categorization aligned with our focus on analyzing negative reviews, as they are most indicative of customer dissatisfaction and areas for potential improvement.

It is worth noting, that the threshold for categorization might be considered subjective because it relies on a fixed numerical range (e.g., 1–2 stars for negative reviews) that may not always align with the true sentiment expressed in the review text. For example, a 3-star review could express strong dissatisfaction or nuanced positivity, depending on the context, and grouping it as "neutral" might oversimplify or misrepresent the actual sentiment. This subjectivity could also influence the evaluation of sentiment analysis models, as the predefined thresholds may not accurately reflect the true underlying sentiment in the review.

2.2 Methodology

2.2.1 VADER

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a traditional sentiment analysis tool that performs well on small English datasets. In our analysis, we used VADER specifically for **English reviews,** as it relies on a sentiment lexicon tailored to English contexts. It computes a compound sentiment score to classify text into positive, neutral, or negative categories: a compound score above 0.05 indicates positive sentiment, a score between -0.05 and 0.05 reflects neutral sentiment, and a score below -0.05 signifies negative sentiment.

We did not apply any preprocessing before using Vader on our reviews as VADER is designed to work directly with raw text, including punctuation, capitalization, emojis, and other textual elements that influence sentiment.

2.2.2 BERT Methods

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model that uses Transformer architecture to understand the context of words by processing text bidirectionally. It employs pre-trained tokenization techniques, such as WordPiece or Byte-Pair Encoding, to break words into subwords, allowing it to capture semantic and contextual relationships effectively.

We did not preprocess any of the reviews before using BERT-based models, as they are specifically designed to handle raw text. Their tokenization and embedding mechanisms can process raw input directly, making preprocessing steps like removing stopwords or punctuation unnecessary. By providing the raw reviews, we ensured that the BERT-based models could fully leverage their contextual understanding to deliver accurate sentiment classification.

BERT-Base-multilingual-uncased-sentiment

The bert-base-multilingual-uncased-sentiment model is a fine-tuned BERT model specifically designed for multilingual sentiment analysis of product reviews. It is capable of processing reviews written in six languages, including English, French, Spanish, German, Dutch, and Italian. This model uses a bidirectional transformer architecture to understand the full context of words in a sentence, reading both forward and backward.

The input text is first tokenized using WordPiece tokenization, which breaks text into smaller subword units, allowing the model to handle rare or unfamiliar words efficiently. The tokenized text is then converted into numerical embeddings, combining token, segment, and positional information. These embeddings are processed through the transformer layers, which capture relationships between words using attention mechanisms.

Fine-tuned on a dataset of product reviews with star ratings from 1 to 5 (mapped as: 1–2 for negative, 3 for neutral, and 4–5 for positive), the model outputs probabilities for each star rating, selecting the one with the highest probability as its prediction. We used this model because its multilingual design enables it to analyze sentiment across different languages, making it suitable for projects like ours.

Amazon Reviews Finetuned BERT-Based

LiYuan is another BERT-based model specifically fine-tuned on Amazon product reviews for sentiment analysis in six languages: English, Dutch, German, French, Spanish and Italian. It predicts the sentiment of the review as a number of stars (between 1 and 5). However, its training on Amazon-specific reviews gives it an edge in capturing nuances related to products, shipping, and customer experience on that platform.

XLM ROBERTA

We used the multilingual XLM-RoBERTa-base model, which was trained on approximately 198 million tweets and fine-tuned for sentiment analysis in 8 languages, as a general sentiment classifier. Unlike models specifically trained on product reviews, XLM-RoBERTa was chosen to provide insights from a broader linguistic context, allowing us to evaluate how well a model not explicitly tailored to reviews could generalize to this domain. It uses subword tokenization through Byte-Pair Encoding (BPE) to break text into smaller units, enabling it to handle rare words by splitting them into common subwords and to manage multilingual text by sharing subwords across languages. Additionally, XLM-RoBERTa leverages masked language modeling (MLM), a method where parts of the text are masked and the model is trained to predict the missing parts, allowing it to capture rich contextual information from multilingual text data.

2.3 Results and Comparative Analysis

For all methods, we evaluated their performance against the star ratings of the reviews, where, as explained in the preprocessing section, negative reviews were classified as those with ratings from 1 to 2 stars, neutral reviews as those with a rating of 3 stars, and positive reviews as those with ratings from 4 to 5 stars.

2.3.1 VADER

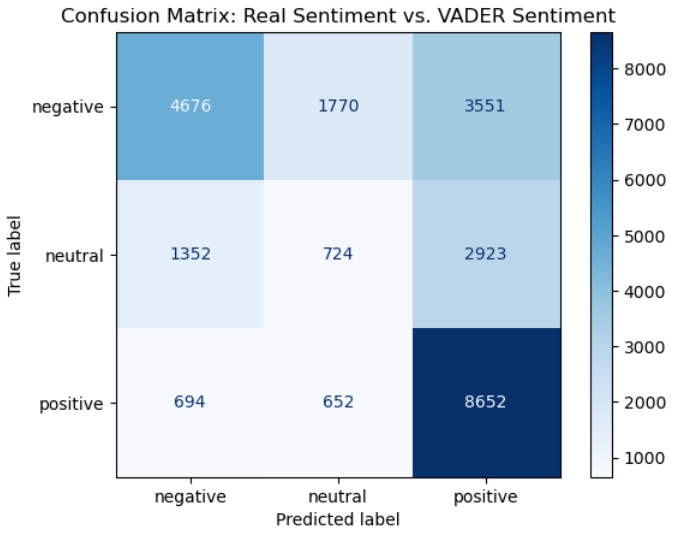


Figure 1: Vader Sentiment Classification Results : Confusion Matrix

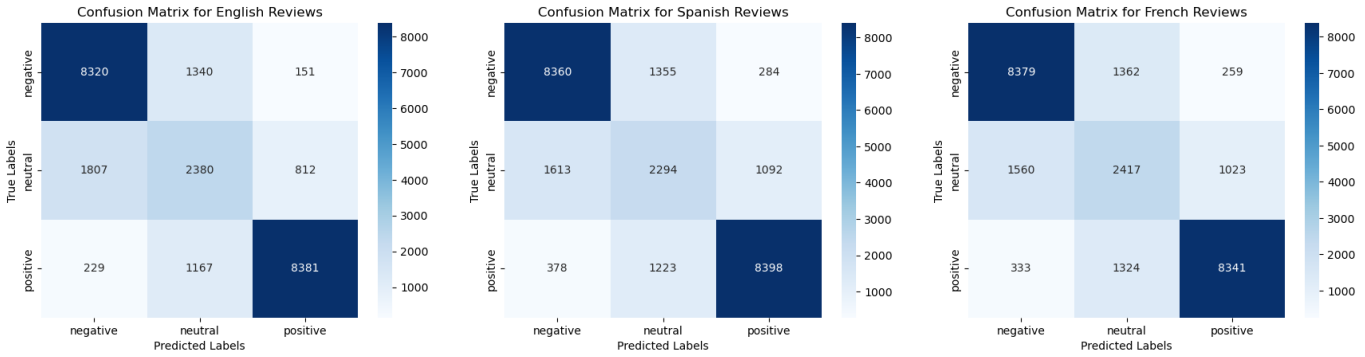
From the confusion matrix for VADER, it is evident that while it performs well in identifying positive sentiments, correctly classifying 8,652 positive reviews, it struggles significantly with negative and neutral sentiments. A large

portion of negative reviews is misclassified as neutral (1,770) or positive (3,551), reflecting its difficulty in capturing nuanced or subtle negative sentiments. VADER demonstrates a bias toward the positive class, as its lexicon-based approach tends to assign higher sentiment scores to reviews containing words with even slightly positive connotations. Similarly, VADER exhibits weak performance in neutral sentiment classification, misclassifying many neutral reviews as either positive (2,923) or negative (1,352), likely due to the inherent challenge of detecting neutral reviews that often contain a mix of positive and negative words.

While VADER effectively identifies strong positive sentiments, its limitations in handling nuanced and neutral sentiments make it insufficient for the complex demands of this project, which focuses on analyzing negative reviews. These shortcomings highlight the need for advanced machine learning models, such as transformer-based classifiers, which are better equipped to handle the intricacies of multilingual and sentiment-diverse data.

2.3.2 BERT-Base-uncased Sentiment Analysis Model

Confusion Matrix and Metrics for Multilingual Reviews BERT base



(a) English Reviews (b) Spanish Reviews (c) French Reviews

Figure 2: Confusion Matrices for Sentiment Analysis of English, Spanish, and French Reviews.

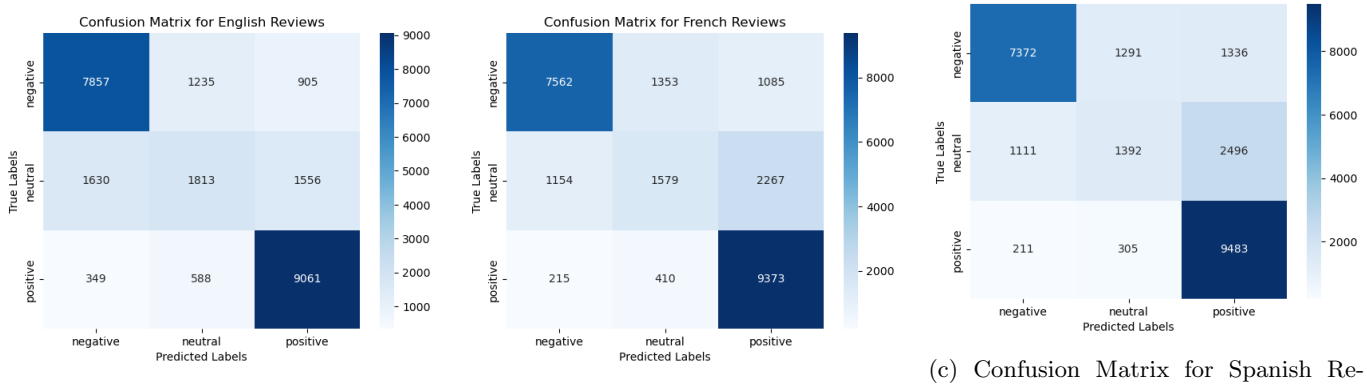
We got a F1-score for negative review of 0.81 for english, 0.82 for Spanish and 0.83 for french (See complete results in appendix). The confusion matrices for English, Spanish, and French reviews highlight the performance of the sentiment classification model across different languages. For all three languages, the model demonstrates strong performance in classifying positive and negative sentiments, as indicated by the high counts along the diagonal for these classes. However, it struggles significantly with neutral reviews. This is reflected in lower true positive counts for neutral labels and higher misclassifications where neutral reviews are often confused as either negative or positive. The lower score for neutral reviews can be attributed to several factors. First, there is an inherent imbalance in the dataset, with approximately only 5000 neutral reviews compared to 10000 positive and 10000 negative reviews. This imbalance naturally makes it more challenging for the model to perform well on the neutral category. Additionally, the way we chose to categorize neutral reviews, aligning them specifically with 3-star reviews, may introduce bias into the model's understanding of neutrality. Neutral sentiment often contains a mix of positive and negative words, making it inherently more complex for a model to accurately classify. However, since the primary focus of this project is on analyzing negative and positive reviews, neutral reviews are less critical to our objectives. These challenges and explanations apply to all the sentiment classifiers used in this project.

Despite these challenges, the model's performance on positive and negative reviews is robust, demonstrating its suitability for tasks focused on these sentiment classes.

2.3.3 Amazon Reviews Fine-tuned BERT-Based Sentiment Analysis Model

The LiYuan model demonstrates strong performance in classifying negative reviews, as reflected in F1-scores of 0.79 for English, 0.8 for French, and 0.79 for Spanish. The model accurately identifies a substantial number of negative reviews, achieving true negative counts of 7,857 for English, 7,562 for French, and 7,372 for Spanish reviews. This indicates a reliable capability in handling clear negative sentiments across all three languages. The recall for negative sentiment classification are consistent, with English achieving the highest scores at 0.79, followed by French at 0.76, and Spanish at 0.74. Nonetheless, the model demonstrates some **high degree of misclassification of negative reviews into positive categories** compared to the previous model (bert-based-uncased), indicating some difficulties with reviews that are less explicitly negative, **impacting its recall for negative reviews**.

Furthermore, as with the previous models, the classification of neutral reviews poses significant challenges.

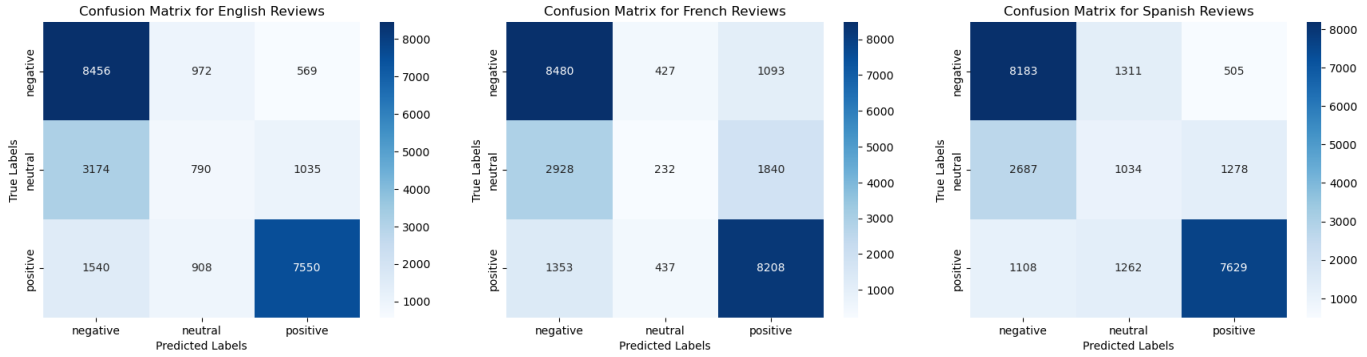


(a) Confusion matrix for English Reviews (b) Confusion Matrix for French Reviews (c) Confusion Matrix for Spanish Reviews

Figure 3: Confusion Matrices for Sentiment Analysis of English, French, and Spanish Reviews.

For positive reviews, the model performs exceptionally well, with high true positive counts of 9,061 for English, 9,373 for French, and 9,483 for Spanish reviews.

2.3.4 XLM-RoBERTa Sentiment Analysis Model



(a) Confusion matrix for English Reviews (b) Confusion Matrix for French Reviews (c) Confusion Matrix for Spanish reviews

Figure 4: Confusion Matrices for Sentiment Analysis of English, French, and Spanish Reviews.

The XLM-RoBERTa model achieved F1 scores of 0.73 for English, 0.75 for French, and 0.74 for Spanish in classifying negative reviews. While XLM-RoBERTa is a robust model for general multilingual sentiment classification, its lack of fine-tuning on product review datasets limits its ability to perform optimally in this specific domain. This limitation is reflected in the confusion matrix, where the model demonstrates strong performance in identifying true negative sentiments. However, it also **frequently misclassifies neutral and positive reviews as negative, lowering its precision for the negative class**. This highlights the challenges of using a more generalized sentiment classifier for domain-specific tasks like product reviews, where tailored models tend to perform better.

Based on these results, we decided to perform sentiment analysis on our negative reviews using the BERT base multilingual uncased model's predicted negative reviews. This model demonstrated a robust performance and appeared to handle the nuances of negative feedback in this domain effectively, making it a more suitable choice for our analysis.

3 Topic Modeling

Topic modeling is an essential part of Natural Language Processing (NLP) that helps identify themes within a collection of documents. In this project, our goal was to uncover the main topics discussed in low-rated reviews. Given that the BERT-Base-uncased model was the sentiment classifier that perform the best on our negative reviews, we used it as a reference point for performing topic modeling.

To ensure a comprehensive and reliable analysis of recurring issues in negative reviews, we employed three classical different topic modeling methods: LDA (Latent Dirichlet Allocation), NMF (Non-Negative Matrix Factorization), and LSA (Latent Semantic Analysis). By combining these approaches, we could cross-validate the results and capture diverse perspectives. Each method offers a unique strength: LDA's probabilistic modeling, NMF's focus on non-negative decompositions, and LSA's ability to detect latent structures in the data. Together, these techniques provided robust and insightful themes to better understand customer satisfaction/dissatisfaction. Additionally, we applied BERTopic, a state-of-the-art topic modeling approach that combines transformers and clustering techniques to capture highly contextualized and dynamic themes in the reviews.

3.1 Preprocessing

Row Deletion

Reviews were further scrutinized to eliminate inconsistencies between the expressed sentiment and the user's rating. Reviews rated below 3 but classified as positive, or above 3 but classified as negative, were discarded.

Additionally, any neutral reviews with extreme ratings of either 1 or 5 were also excluded. This selective filtering ensures that the analysis considers only reviews whose sentiments and ratings are congruent.

English reviews went from 24994 rows to 24587 rows, spanish reviews went from 24997 to 24636 rows and french reviews went from 24998 to 24692 rows.

Stopwords Removal

To prepare the reviews for topic modeling, we performed a series of preprocessing steps to clean and normalize the text data. First, all special characters, punctuation, and unnecessary symbols were removed, and the text was converted to lowercase to ensure uniformity. Next, we tokenized the reviews, breaking them down into individual words or tokens. We then removed stopwords, which are common words like "the," "is," and "and," that do not contribute meaningfully to the context, by combining stopword lists from both NLTK and SpaCy for better coverage. Additionally, only alphabetic tokens were retained, discarding numbers and other irrelevant characters.

Lemmatization

Lemmatization was performed before topic modeling to ensure that words are reduced to their root form, helping to group different variations of the same word. This process reduces noise in the data and ensures that similar terms contribute to the same topic. By normalizing words, lemmatization enhances the quality and interpretability of the resulting topics. For instance, reviews about shipping might contain variations like "delivered on time" or "delivering late," but lemmatization ensures they are grouped under the common concept of "deliver," enhancing the quality and interpretability of the resulting topics.

Topic Modeling Parameters

All topics were created using :

• "max_df=0.80": This parameter ensures that words appearing in more than 80% of the documents are excluded from the topic modeling process. Such words are considered too common across the corpus and provide little value for distinguishing between topics.

• "min_df=30": This parameter ensures that only words appearing in at least 30 documents are included in the topic modeling process. Words that occur less frequently are often too specific or rare, which may lead to noise rather than meaningful patterns.

Together, these parameters help focus on terms that are neither too rare nor too common, improving the quality of the topics generated by excluding irrelevant or overly generic words.

3.2 Methodology

In order to find the optimal number of topics we used a coherence score. For the english reviews the max Coherence Score was approximately 0.615 for 3 topics. To maintain consistency and simplify our analysis while providing a comprehensive overview of the topics, we decided to use 3 topics throughout the entire study across all methods (LDA, LSA, and NMF) and languages.

It is worth noting that the resulting optimal number of topics can vary depending on the method used (e.g., NMF, LSA) and the language of the reviews. Alternative approaches, such as minimizing reconstruction error for NMF or maximizing explained variance for LSA, might suggest a different number of topics, potentially leading to higher coherence or reduced variance across topics.

3.2.1 LDA

LDA (Latent Dirichlet Allocation) is a probabilistic machine learning method used to uncover hidden topics in a collection of documents. It models each document as a mixture of multiple topics and each topic as a probability distribution over words. By analyzing word patterns across all documents, LDA identifies these topics and determines the contribution of each topic to each document. For example, LDA might infer topics such as 'pricing issues,' 'delivery problems,' or 'product quality,' and quantify how likely each review is to relate to these topics.

3.2.2 NMF

NMF works by factorizing the document-term matrix into two smaller matrices: one that represents the topics and another that shows how strongly each topic contributes to each document. Unlike other methods, NMF uses only non-negative values, ensuring that topics are purely additive. This makes the resulting topics more interpretable, as they represent distinct and non-overlapping parts or features of the data. For instance, NMF can highlight specific problem areas in customer reviews by isolating key terms linked to different aspects of a product or service, such as 'delivery', 'pricing', or 'product quality'.

3.2.3 LSA

LSA uses Singular Value Decomposition (SVD) to reduce the dimensions of the document-term matrix, capturing the latent relationships between words and documents. By grouping terms that frequently co-occur, LSA reveals hidden structures in the data and extracts meaningful topics. However, LSA can produce both positive and negative contributions, which makes topics less interpretable compared to NMF.

3.2.4 BERTopic

The last model we chose to include in our topic modeling part of the pipeline is BERTopic and as the name highlights, this model uses Transformer embeddings to represent text in a high-dimensional semantic space. This is especially relevant to our context with reviews in our data set having an average sentence length of around 36.47 tokens (according to NLTK's tokenizer), meaning that as sentences get longer, it is crucial to use models that can capture contextual relationships and semantic nuances across the entire sequence, this is achieved through the self-attention mechanism introduced by Vaswani et al. (2017). This enables the model to process long-range dependencies more efficiently.

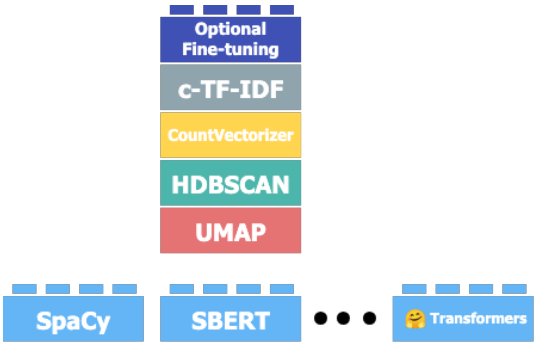


Figure 5: BERTopic Model Architecture, For details, see BERTopic Documentation.

After tokenization and embedding, dimensionality reduction methods such as UMAP help the model process high-dimensional semantic data sets by mapping them into a lower-dimensional space while preserving underlying structures and relationships.

Once these positional embeddings are used on reviews, the output vector is then passed through the model and BERTopic clusters reviews using HDBSCAN (Hierarchical Density-Based Spatial Clustering) which is another key aspect of this topic model since it allows for us to avoid predefining a number of topics desired, letting the model choose the optimal one during its processing of reviews. BERTopic then refines the cluster of reviews of similar topics by computing "importance" scores for the most recurring and significant words throughout the newly formed topics through TF-IDF.

This model is particularly suitable as it uses Transformer embeddings to generate semantically rich topics, which we later on incorporated into the Neural Network for improved predictions. However, we avoided direct comparisons with traditional models like LDA, LSA, or NMF to ensure fairness, as these models lack the contextual depth needed for training.

3.3 Results and Comparative Analysis

3.3.1 English Reviews

Table 1: Topics Generated by NMF, LDA, and LSA for Low-Rated Reviews

Model	Topics
NMF	Topic 1: like, break, buy, look, use, come, product, time, good, return
	Topic 2: work, stop, month, week, return, charge, great, light, buy, battery
	Topic 3: receive, order, item, product, send, return, seller, package, amazon, deliver
LDA	Topic 1: break, use, buy, phone, like, come, case, try, work, time
	Topic 2: work, order, product, receive, return, buy, item, send, box, time
	Topic 3: like, look, fit, small, quality, size, product, return, good, wear
LSA	Topic 1: work, product, buy, return, like, break, receive, month, order, time
	Topic 2: work, stop, month, week, charge, great, light, quit, battery, plug
	Topic 3: receive, order, item, product, work, send, seller, deliver, package, refund

The main topics identified from the English reviews highlight key areas of customer dissatisfaction. Product quality and durability emerge as significant concerns, with frequent mentions of items breaking, failing, or not meeting expectations, often leading to high return rates. Delivery and order fulfillment issues are also prominent, including complaints about delays and unreliable service. Additionally, size problems are emphasized in the LDA results, reflecting recurring dissatisfaction with fit. Finally, there are notable mentions of battery and charging performance in electronics, with customers reporting issues such as poor battery life, unreliable charging, and products stopping to work after a short period. These insights pinpoint critical areas where the business can focus its efforts to address customer pain points and improve overall satisfaction in anglophone markets.

3.3.2 French Results

Table 2: Topics Generated by NMF, LDA, and LSA for Low-Rated Reviews in French

Model	Topics
NMF	Topic 1: produit, qualité, cest, déçu, fonctionne, mauvais, bien, nest, mois, bon
	Topic 2: recevoir, jamais, nai, colis, article, command, vendeur, livrer, produit, commander
	Topic 3: trop, petit, beaucoup, taille, grand, fragile, vraiment, cher, cest, déçue
LDA	Topic 1: trop, produit, petit, qualité, déçue, déçu, vraiment, nest, bien, dommage
	Topic 2: recevoir, produit, nai, jamais, colis, vendeur, livraison, article, amazon, command
	Topic 3: produit, fonctionne, mois, bien, qualité, bout, acheter, mauvais, jour, téléphone
LSA	Topic 1: recevoir, produit, jai, jamais, nai, cest, qualité, trop, déçu, bien
	Topic 2: recevoir, jamais, nai, colis, command, article, vendeur, livrer, rembourser, remboursement
	Topic 3: trop, petit, beaucoup, recevoir, taille, grand, nai, jamais, cest, article

For French reviews, the topics highlight recurring issues, including product quality (e.g., products stop working after a short time) and delivery problems. In particular, complaints about delivery (e.g., terms like 'recevoir,' 'colis,' 'livraison,' 'vendeur,') suggest that improving the efficiency and reliability of our courier services could significantly enhance customer satisfaction. Additionally, topics related to sizing, which are recurrent across all three methods, indicate an area that should be prioritized for improvement in the French market.

3.3.3 Spanish Results

Table 3: Topics Generated by NMF, LDA, and LSA for Low-Rated Reviews in Spanish

Model	Topics
NMF	Topic 1: calidad, malo, romper, pequeño, comprar, recomeir, precio, quedar, venir, plástico Topic 2: llegar, producto, pedir, recibir, esperar, mes, devolver, dinero, vendedor, amazon Topic 3: funcionar, dejar, mes, devolver, año, carga, comprar, luz, semana, compre
LDA	Topic 1: producto, llegar, venir, funcionar, devolver, amazon, pedir, comprar, cajar, problema Topic 2: calidad, malo, esperar, recibir, pequeño, pantalla, producto, foto, precio, color Topic 3: romper, funcionar, comprar, quedar, mes, dejar, durar, servir, año, duro
LSA	Topic 1: producto, llegar, él, funcionar, calidad, malo, mes, esperar, devolver, comprar Topic 2: llegar, producto, pedir, recibir, esperar, vendedor, fecha, valorar, paquete, dinero Topic 3: funcionar, dejar, mes, llegar, devolver, carga, pedir, correctamente, él, luz

In Spanish reviews, common themes include product fragility and delivery delays. Similar to the other languages, there is some redundancy in the identified topics, particularly between quality and durability concerns.

3.3.4 Uses Cases

The primary purpose of performing topic modeling on negative reviews was not to compare the different methods. Instead, we aimed to identify robust topics, those that consistently appeared across multiple methods. By doing so, we could pinpoint recurring complaints and uncover areas for improvement in our business. For example, we could extract topics using any of the above methods for specific categories, such as "shoes," to better understand customer feedback :

Topic Number	Top Words
Topic 1	size, small, return, order, run, fit, shoe, need, right, tight

Table 4: An NMF Topic for negative Reviews in Shoes.

The NMF topic analysis for the "shoes" category highlights issues primarily related to sizing, as evident from the top words such as "size," "small," "fit," "tight," and "run". These words suggest that customers frequently face problems with shoe sizes being smaller or tighter than expected.

To address this, the business could include size recommendations on the product page, such as advising customers to select a larger size if the shoe tends to run small. Additionally, providing detailed size charts and encouraging user feedback on sizing can help reduce dissatisfaction and returns. This insight could lead to better customer satisfaction and fewer size-related complaints.

3.3.5 BERTopic

BERT tends to generate **more granular and category-specific topics** compared to the broader themes extracted by NMF, LDA, and LSA. It output recurrent topics and category in negatives reviews. This is achieved through the power of attention that is built into the model, which traditional methods may lack.

Table 5: Summary of the Top 3 Topics Identified by BERT for French and Spanish Reviews

Language	Topic 1	Topic 2	Topic 3
French	Colors (couleur, photo, bracelet)	Packaging Issues (plastique, emballage)	Phones (écran, téléphone, batterie)
Spanish	Colors and Material (color, luz)	Price and Quality (precio, calidad)	Electronics (batería, móvil, teléfono)

In the French reviews, BERT identifies more granular topics like packaging and phones issues, while NMF, LDA, and LSA highlight broader themes of delivery problems and quality dissatisfaction. In the Spanish reviews, price and quality concerns alongside electronics issues arise, which align closely with NMF and LDA outputs. The recurring dissatisfaction with colors and materials highlights a consistent theme in customer feedback.

For the English reviews, BERTopic identifies key themes such as delivery issues (e.g., "refund," "order," "return"), and product quality concerns (e.g., "leak," "broken," "assembled"). Specific mentions of battery life, which was also

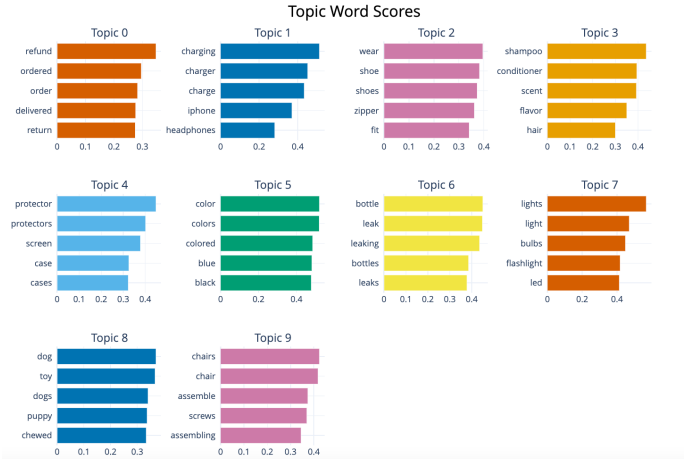


Figure 6: Example of Topics from English Reviews

observed in LSA and NMF, requiring improvement to address customer dissatisfaction. Additionally, we observe topics related to shampoo and scent in negative reviews (e.g., "shampoo," "conditioner," "scent") suggesting dissatisfaction with beauty or personal care product.

BERTopic acts as a powerful tool for us to include into our Recurrent Neural Network model, which will act as a "built-from-scratch" sentiment analysis classification model that harnesses the capabilities of attention mechanisms as well as fine-tuning.

4 Neural Network

In the final part of our project, we utilized Neural Networks to build a custom sentiment analysis model tailored for multilingual customer feedback. Neural networks were preferred because of their capacity to map out complicated patterns in data, notably through architectures like as Long Short-Term Memory (LSTM) networks, which are effective at capturing sequential dependencies in text. Unlike classic lexicon-based approaches such as VADER or fine-tuned transformers, our neural network was created to incorporate many sources of information, making it both flexible and domain specific. The fundamental goal of this strategy was to create a model that could process multilingual input while also including information like language and BERTopic-derived topics.

4.1 Design and Implementation

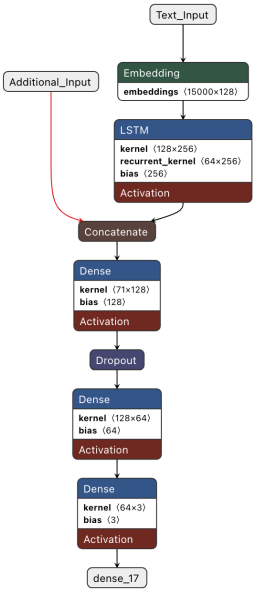


Figure 7: Final RNN Model Architecture

To create a custom sentiment analysis model for multilingual and domain-specific data, we used a Neural Network (NN) architecture based on Long Short-Term Memory (LSTM) layers. LSTMs were chosen for their ability to efficiently process sequential data while maintaining context over long text sequences, making them ideal for tasks such as sentiment analysis. For example, in a sentence like *"The product was good, but the delivery was terrible,"* the LSTM can detect the progressive sentiment shift from positive to negative for example, which simpler methods like bag-of-words frequently miss.

Our neural network was built to be multimodal, incorporating both text-based features and structured metadata to improve interpretability and flexibility. To capture sequential patterns, the text input was tokenized and embedded into dense vectors with an embedding dimension of 128, before being fed via LSTM layers. Metadata, such as language and BERTopics, were encoded and sent into the network alongside the text output, allowing the architecture to adapt to the subtleties of different datasets.

The final model architecture consists of one LSTM layer and two fully linked hidden layers with 128 and 64 neurons, triggered by ReLU functions, and achieved the highest test accuracy of 71.32% with a competitive test loss of 0.69. To prevent overfitting, the dropout rate was set to 0.2. Using a softmax activation function, outputs were classified into three categories (positive, neutral, and negative). Adam was used as an optimizer, with a learning rate

of 0.0005, to minimize the sparse categorical cross-entropy loss function, the full description of the model is found in Table 13.

4.1.1 Hyperparameter Search

To reach maximum performance, some preprocessing of the training set allowed it to rebalance after random sampling, in order to oversample "neutral" rated reviews and ensure the model was exposed to sufficient examples of these more ambiguous instances, thereby improving its ability to accurately classify neutral sentiments alongside positive and negative categories. Additionally, hyperparameters were fine-tuned and several configurations were tested. With a training set of around 60,000 reviews, methods such as Cross-Validation grid search were deemed too computationally expensive due to resource limitations, but different model architectures were still explored nonetheless.

Activation Function / Model	Test Accuracy (%)	Test Loss	Observation
Activation Functions with 3 Hidden Layers			
ReLU	71.32	0.69	Best model overall, achieving highest accuracy.
Tanh	67.80	0.76	Slightly lower accuracy, lowest loss.
Leaky ReLU	67.41	0.77	Similar performance to ReLU, slightly higher loss.
ELU	68.35	0.76	Competitive accuracy with balanced performance.
Custom Weights and Preprocessing			
Custom Weights	65.00	0.74	Penalized misclassification, prioritizing specific errors.
Lemmatized Reviews	48.19	0.97	Significant drop in performance, indicating preprocessing impact.
Hidden Layers			
3 Hidden Layers	71.32	0.69	Improvement in accuracy and reduced loss.
4 Hidden Layers	67.00	0.77	Slight decrease in accuracy.
5 Hidden Layers	66.32	0.78	Decreased accuracy and increased loss, suggesting overfitting or gradient issues.

Table 6: Comparison of Activation Functions, Custom Weights, Preprocessing, and Network Depth in Neural Network Performance

This table shows the results of testing variations in the architecture and the optimization of our RNN. For example, different activation functions (Tanh, ELU) were used, as well as trying to implement custom weights for penalizing errors in misclassification of positive reviews into negatives, which was critical with our marketing context as misclassifying positive reviews into negatives can lead to erroneous business insights. For example, a loyal or satisfied customer could be incorrectly flagged as a detractor, potentially leading to unnecessary and misdirected engagement efforts. We also considered text preprocessing such as lemmatization, but it proved counterproductive in this RNN since models that use LSTMs rely on semantic context to process reviews, and lemmatization strips words of their contextual and semantic richness, thereby reducing the ability to capture nuanced meanings in reviews.

Additionally, tests with varying architecture depth (three to five hidden layers) showed that overfitting occurred when the depth was increased above four layers, with a larger test loss. With its ability to use both textual and structured inputs, this multimodal NN is a significant improvement over lexicon and compound score based techniques such as VADER, especially for neutral and nuanced sentiments.

4.2 Results and Comparative Analysis

The results of our R-Neural Network demonstrate its capacity to classify sentiment into three categories (positive, neutral, and negative) with notable improvements over previous approaches like VADER.

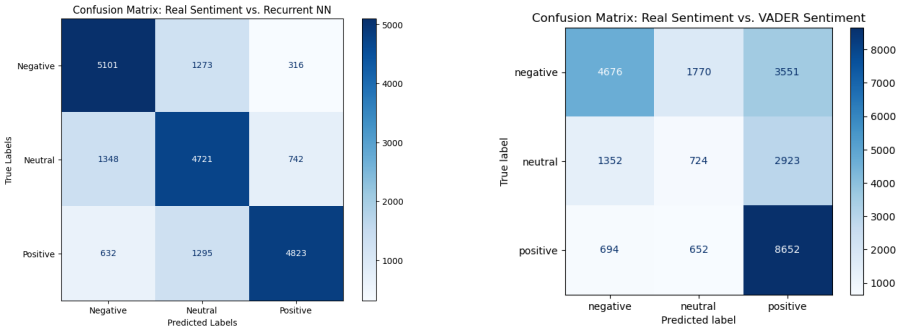


Figure 8: Comparison of Confusion Matrices for Recurrent Neural Network and VADER Sentiment Classifications.

The final model has a test accuracy of 71.32% and a test loss of 0.6884, indicating superior overall performance. This performance gain (compared to the previously reported accuracy of 69% during the presentation) was due to addressing a preprocessing issue in which the model was trained on a 'clean' review variable that eliminated stop words. Stop words are vital for an RNN-based architecture because they assist capture subtle emotions, resulting in a small increase in accuracy from 69% to 71 percent.

Model	**Sentiment**	**F1-Score**
Baseline (Vader, English only)	Negative	0.56
	Neutral	0.18
	Positive	0.69
Neural Network (Multimodal & Multilingual)	Negative	0.76
	Neutral	0.69
	Positive	0.71
Overall Accuracy: Baseline = 0.56, Neural Network = 0.71		

Table 7: Comparison of Sentiment Classification Results Between Baseline and Improved Model

The results clearly show that the multimodal and multilingual Neural Network (NN) outperformed the more 'generic' VADER model in sentiment classification across all categories. The NN got an F1 score of 0.76 for negative feelings, which was much higher than VADER's 0.56. Similarly, the NN outperformed VADER on neutral sentiments, with an F1 score of 0.69, a significant improvement given the inherent ambiguity and difficulty of categorizing neutral assessments. For positive reviews, the NN had an F1 score of 0.71, a little higher than VADER's 0.69, which proves that a lexicon based approach is still performant in identifying positive sentiment.

A key advantage of our model is its multimodal and multilingual approach obtained through metadata, like language and BERTopic topics. Firstly, the model's ability to take language as an explicit input to account for differing grammatical structures and word orders across English, French, and Spanish. This allows the model to better adapt to each language's unique syntactic patterns, resulting in more accurate sentiment predictions. Secondly, including topics from BERTopic improves the model's understanding of context. For example, knowing if a review is about "delivery issues" or "product quality" provides useful additional information that allows the NN to focus on important areas of the text, enhancing its ability to classify nuanced sentiments.

Lastly, one of the most apparent improvements is the NN's handling of neutral feelings, as oversampling during preprocessing and multimodal input integration helped mitigate class imbalance and catch subtleties that VADER struggled with. This is critical in real-world applications, since neutral reviews frequently contain significant information that might help with customer service or product improvement efforts and make all the difference between a negative and bad client experience.

Overall, the NN's higher performance illustrates the relevance of combining deep learning and multimodal approaches for nuanced, large-scale multilingual sentiment analysis, resulting in a more resilient and adaptable solution for a variety of datasets.

However, it's important to underline that our model could be improved, since its integration of different languages make it differ in terms of performance from one language to another. Though careful sampling made sure that the model was trained on equal instances of each language, it could benefit from an oversampling of French reviews, which is were it seems less confident in prediction in. An example to illustrate would be these three sample sentences that we tested the model on, these are direct translations of a made-up review in English Spanish and French.

Sentence	Predicted Sentiment	Confidence Score
This product is amazing and exceeded expectations, I'm really happy with it!	Positive	0.84
Ce produit est incroyable et a dépassé mes attentes, je suis vraiment content de l'avoir!	Positive	0.40
Este producto es increíble y superó mis expectativas, estoy muy feliz con él!	Positive	0.90

Table 8: Sample Sentences with Predicted Sentiments and Confidence Scores

It's important to highlight that we use the term "confidence" to refer to the greatest softmax output/predicted probability, which assesses the model's association between the input and its anticipated classification. It indicates the model's degree of certainty in its predictions.

For English reviews, the model performed well, predicting "Positive" sentiment with a confidence score of 0.84, showing strong alignment with the true sentiment. For Spanish reviews, it achieved similar high performance, with a confidence score of 0.90 for "Positive" sentiment. However, for French reviews, the confidence score dropped significantly to 0.40 for "Positive" sentiment, suggesting the model struggled more with French reviews.

This small caveat is a future area of improvement for further refinement of our fine-tuned sentiment classifier.

5 Conclusion

Throughout the scope of our project, we've used powerful methods and applied them to a tangible marketing setting. This project aimed to overcome the challenges of analyzing large-scale, multilingual customer reviews by developing a comprehensive pipeline that combined sentiment analysis, topic modeling, and neural networks. The continuous development of this pipeline demonstrates an intentional effort to improve sentiment classification accuracy, find meaningful insights, and adapt powerful algorithms to multilingual data. Starting with classic techniques like VADER for baseline sentiment analysis and gradually introducing transformer-based models, we established a solid basis while addressing limitations in more nuanced subjective neutral reviews.

Our method transitioned seamlessly from traditional to advanced transformer-based models, culminating in a customized recurrent neural network. This pipeline allowed us to learn from each method and integrate our findings in the reflection that led to build our very own sentiment classifier while improving accuracy and demonstrating adaptability across several languages: English, French, and Spanish. The integration of BERTopics with the neural network was a turning point, allowing for greater understanding of context and more robust sentiment classification. We made significant gains by combining cutting-edge models with domain-specific tuning, particularly in detecting complicated patterns in neutral and negative reviews—critical pain areas in marketing analytics.

The topic modeling phase complemented sentiment analysis by identifying recurring themes in low-rated (negative) reviews. Traditional methods such as LDA, NMF, and LSA revealed overlapping themes, including delivery issues, product quality concerns, and dissatisfaction with sizing or aesthetics. BERTopic further enhanced this analysis by leveraging transformer embeddings and clustering techniques to extract highly contextualized and dynamic topics. such as product quality, delivery issues or battery performance in english reviews while in French reviews, topics centered on packaging, delivery problems, and sizing issues. Finally Spanish customers complained about price and quality alongside electronics performance.

This pipeline's added value comes from its multimodal and multilingual aspect, which would allow marketers to properly understand not just "what" customers are saying but also "why" they feel the way they do. This combined focus on emotion and context guarantees that firms can successfully address particular issues while also using positive feedback to support successful practices. Furthermore, the flexibility of our neural network architecture enables scalability and adaptability for future modifications, such as adding other languages or fine-tuning product category specific scenarios.

Ultimately, this study emphasizes the need of combining cutting-edge machine learning algorithms with careful pipeline design to address real-world problems. The end result is a marketing tool that bridges the gap between raw unstructured data and strategic decision-making, allowing firms to improve customer satisfaction, create trust, and achieve a competitive advantage in global markets. However, as the comparison results show, there is still potential for improvement, particularly in terms of maintaining consistent performance across languages, revealing promising areas for future study and innovation.

References

[1] Balahur, A., & Turchi, M. (2014). Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. *Computer Speech & Language*, *28*(1), 56–75.

[2] Grootendorst, M. (2024). *BERTopic Dimensionality Reduction Documentation*. Retrieved from GitHub.

[3] Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4996–5001. Florence, Italy. https://aclanthology.org/P19-1493

[4] Zhang, L., Wang, S., & Liu, B. (2018). Deep Learning for Sentiment Analysis: A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253. https://doi.org/10.1002/widm.1253

[5] Hugging Face. (n.d.). *Transformers Documentation: BERT Model*. Retrieved from https://huggingface.co/docs/transformers/v4.46.3/en/model_doc/bert#transformers.BertModel

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved from https://arxiv.org/abs/1910.03771

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Retrieved from https://arxiv.org/abs/1706.03762.

6 Appendix

6.1 Sentiment Analysis

Model	English Sentiment (Score)	French Sentiment (Score)	Spanish Sentiment (Score)
Multilingual BERT-base	1 star (0.93)	1 star (0.72)	1 star (0.92)
XLM-RoBERTa-base	Negative (0.96)	Negative (0.96)	Negative (0.95)
Fine-Tuned Multilingual BERT-base	1 star (0.95)	5 stars (0.58)	1 star (0.96)

Table 9: Sentiment Predictions for the Same Sample Sentence Across Languages

Language	Metric	Negative	Neutral	Positive	Macro Avg	Weighted Avg
English	Precision	0.79	0.49	0.88	0.72	0.76
	Recall	0.83	0.48	0.84	0.72	0.76
	F1-Score	0.81	0.48	0.86	0.72	0.76
	Accuracy	0.76				
Spanish	Precision	0.81	0.47	0.86	0.71	0.76
	Recall	0.84	0.46	0.84	0.71	0.76
	F1-Score	0.82	0.46	0.85	0.71	0.76
	Accuracy	0.76				
French	Precision	0.82	0.47	0.87	0.72	0.77
	Recall	0.84	0.48	0.83	0.72	0.77
	F1-Score	0.83	0.48	0.85	0.72	0.77
	Accuracy	0.77				

Table 10: Confusion matrix metrics for English, Spanish, and French reviews of BERT BASE. Metrics include precision, recall, F1-score, macro average, weighted average, and overall accuracy.

Language	Metric	Negative	Neutral	Positive	Macro Avg	Weighted Avg
English	Precision	0.80	0.50	0.79	0.69	0.73
	Recall	0.79	0.36	0.91	0.68	0.75
	F1-Score	0.79	0.42	0.84	0.68	0.74
	Accuracy	0.75				
French	Precision	0.85	0.47	0.74	0.69	0.73
	Recall	0.76	0.32	0.94	0.67	0.74
	F1-Score	0.80	0.38	0.82	0.67	0.73
	Accuracy	0.74				
Spanish	Precision	0.85	0.47	0.71	0.68	0.72
	Recall	0.74	0.28	0.95	0.65	0.73
	F1-Score	0.79	0.35	0.81	0.65	0.71
	Accuracy	0.73				

Table 11: Confusion matrix of LIYUAN metrics for English, French, and Spanish reviews. Metrics include precision, recall, F1-score, macro average, weighted average, and overall accuracy.

Language	Metric	Negative	Neutral	Positive	Macro Avg	Weighted Avg
English	Precision	0.64	0.30	0.82	0.59	0.65
	Recall	0.85	0.16	0.76	0.59	0.67
	F1-Score	0.73	0.21	0.79	0.57	0.65
	Accuracy	0.67				
French	Precision	0.66	0.21	0.74	0.54	0.60
	Recall	0.85	0.05	0.82	0.57	0.68
	F1-Score	0.75	0.08	0.78	0.53	0.62
	Accuracy	0.68				
Spanish	Precision	0.68	0.29	0.81	0.59	0.65
	Recall	0.82	0.21	0.76	0.60	0.67
	F1-Score	0.74	0.24	0.79	0.59	0.66
	Accuracy	0.67				

Table 12: Confusion matrix metrics for XLM-RoBERTa sentiment analysis across English, French, and Spanish reviews. Metrics include precision, recall, F1-score, macro average, weighted average, and overall accuracy.

6.2 Topic Modeling

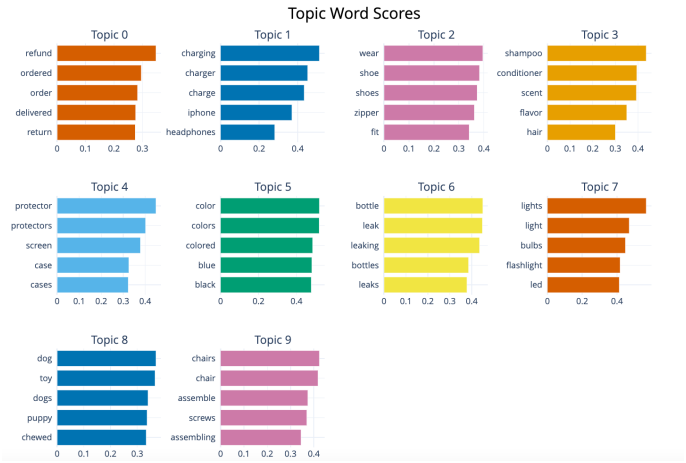


Figure 9: BERTopic for English Reviews

Figure 10: BERTopic for French Reviews

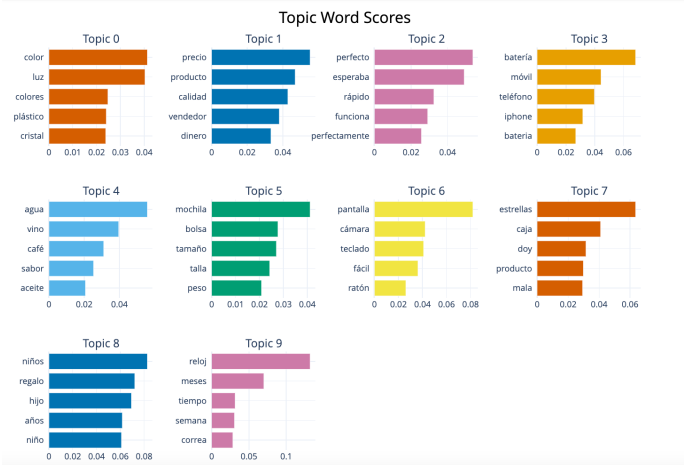


Figure 11: BERTopic for Spanish Reviews

6.3 Neural Network

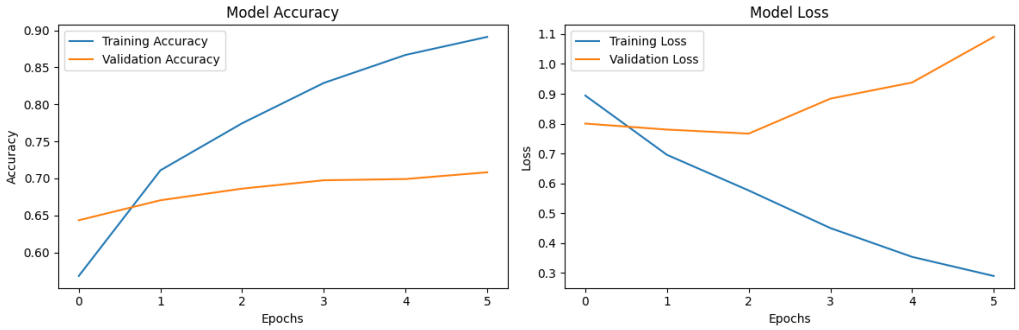


Figure 12: Training and Test Scores

Hyperparameter	**Value**
Embedding Dimension	128
Sequence Length	512
Hidden Layers	3 (LSTM, 128, 64 neurons)
Dropout Rate	0.2
Output Classes	3 (positive, neutral, negative)
Activation Function	ReLU (hidden), Softmax (output)
Optimizer	Adam
Learning Rate	0.0005
Loss Function	Sparse Categorical Crossentropy
Batch Size	32
Epochs	Up to 50 (Early Stopping after 3 epochs)

Table 13: Hyperparameters of the Best Neural Network Model