Analyzing Sentiment Across Languages:

Comparing ML Models on French and English Amazon Reviews

Final Project - MATH60629A

Professor: Dr. Firoozi

Presenters: Shamim Mahmoudzadeh Vaziri
Mahmoud Moubarak
Maria Camila Gomez Lopez
Louis Fillon

Outline

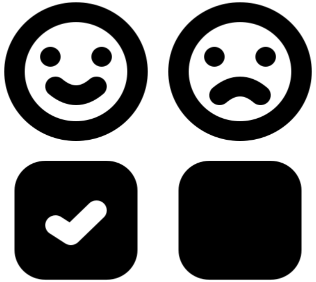
- Introduction & Background
- Problem Statement
- Machine Learning Models
- Conclusion & Results

Introduction & Background

Context

- Digital world is becoming more multilingual, with users expressing opinions in various languages across platforms.

- Companies need to understand customer satisfaction across different Region



- Manual analysis is impractical due to large review volumes.

- SA utilizing ML models = automated classification of opinions (positive, neutral, negative).
- Widely used in customer feedback, brand monitoring, and product development.

Literature Review

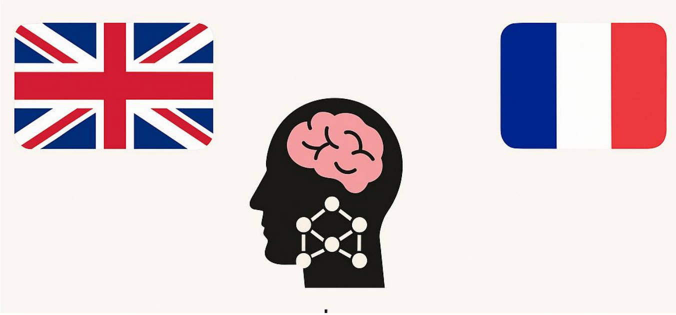
AUTHORS	FINDINGS	METHODS
Fang, X., Zhan, J. (2015)	Future work to extract implicit sentiment from texts. In general, the Random Forest model performs the best.	Naïve Bayes, Random Forest, SVM
Singh, Shubham & Singla, Neetu. (2023)	Deep learning models, particularly LSTM, offer significant advantages in processing and analyzing natural language data	Naïve Bayes, SVM, LSTM, Decision tree
Conneau, Alexis & Khandelwal, Kartikay & Goyal, Naman & Chaudhary, Vishrav & Wenzek, Guillaume & Guzman, Francisco & Grave, Edouard & Ott, Myle & Zettlemoyer, Luke & Stoyanov, Veselin. (2020).	Multilingual modeling without sacrificing perlanguage performance; XLM-RoBERTa	BERT, XLM-RoBERTa
Yanying Mao, Qun Liu, and Yu Zhang. (2024)	Challenges in Sentiment analysis include: Identifying sarcasm, understanding slang and abbreviations, implicit sentiments.	Review of multiple methods (traditional, deep learning, transformers)

Problem Statement

Problem Statement

- ML Models for SA often perform well on English data but struggle with multilingual datasets.
- Linguistic and cultural differences make it challenging to detect sentiment accurately across languages.
- Limited multilingual datasets restrict comprehensive cross-lingual research.

Task and Objective



Research Question:

How does sentiment vary across different languages, and can a machine learning model be trained to perform accurate sentiment analysis across multiple languages while addressing linguistic and cultural differences?

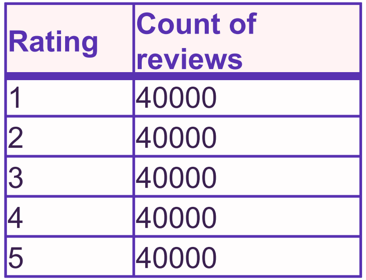
Objectives:

- Compare sentiment analysis performance across English and French.
- Benchmark traditional, deep learning, and transformer-based models.
- Identify model limitations and suggest improvements for multilingual settings.

Machine Learning Models

Dataset

- □ Dataset from Kaggle: “French reviews on amazon items and EN translation”
- □ Part of the Multilingual Amazon Reviews Corpus (MARC), a large-scale collection of Amazon reviews in English, Japanese, German, French, Spanish, and Chinese, between 2015 and 2019
- □ Dataset size: 200k reviews
- □ Balanced dataset
- □ French reviews are translated with the API of Google Traduction.



Rating	Count of reviews
1	40000
2	40000
3	40000
4	40000
5	40000

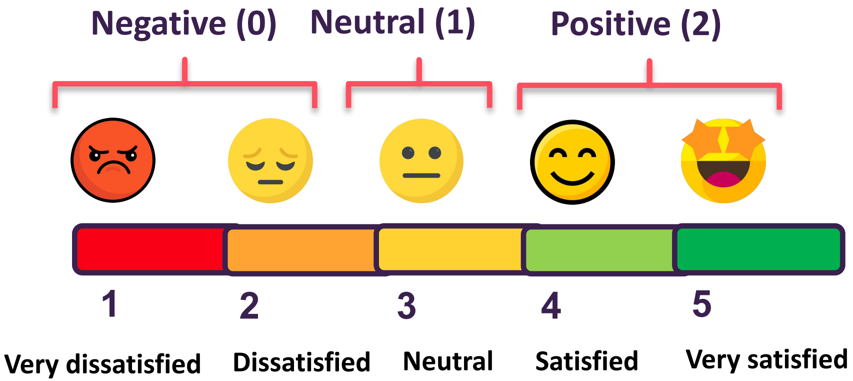
Dataset

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.

https://www.kaggle.com/datasets/dargolex/french-reviews-on-amazon-items-and-en-translation

Preprocessing

1) **Data Reduction:** sample specific class sizes
{1: 5000, 2: 5000, 3: 10000, 4: 5000, 5: 5000}
2) **Label Transformation:**
convert 5-point ratings to 3-class sentiment
3) **Dataset Splitting:**
- ○ Train set (0.7)
- ○ Validation set (0.15)
- ○ Test set (0.15)



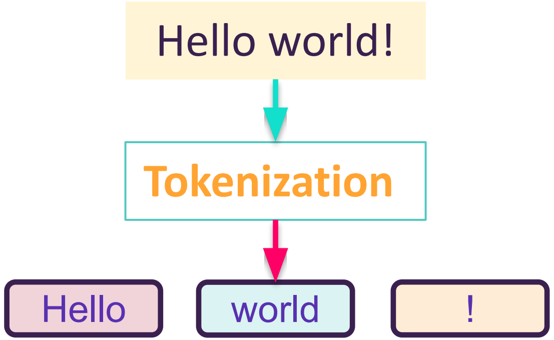
Preprocessing

4) Text Preprocessing:

- ○ Removing stop words (e.g., the, is, in, et, le, la))
- ○ Removing punctuations and special characters (to eliminate noise)
- ○ Lemmatizing (convert words to their base form)
- ○ Tokenization (split text into words or phrases)
- ○ lowercasing (standardize text)

! : -

& @ /



Word	Lemmatizing
feet	foot
computers	computer
changing	change
was	(to) be
better	good

Vectorization Methods

- **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 [5]

Method	**English**	**French**
BoW	✅	✅
TF-IDF	✅	✅
Word2Vec	✅	❌

Evaluation Metrics

- ❑ Accuracy
- ❑ Precision
- ❑ Recall
- ❑ F1-Score [4]

	Predicted	
Actual	True Positives TP	False Negatives FN
	False Positives FP	True Negatives TN

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ML Models

- ❑ Baseline Model
- ❑ Naive Bayes
- ❑ Support Vector Machines
- ❑ Random Forest
- ❑ Long Short-Term Memory (LSTM)
- ❑ Bert

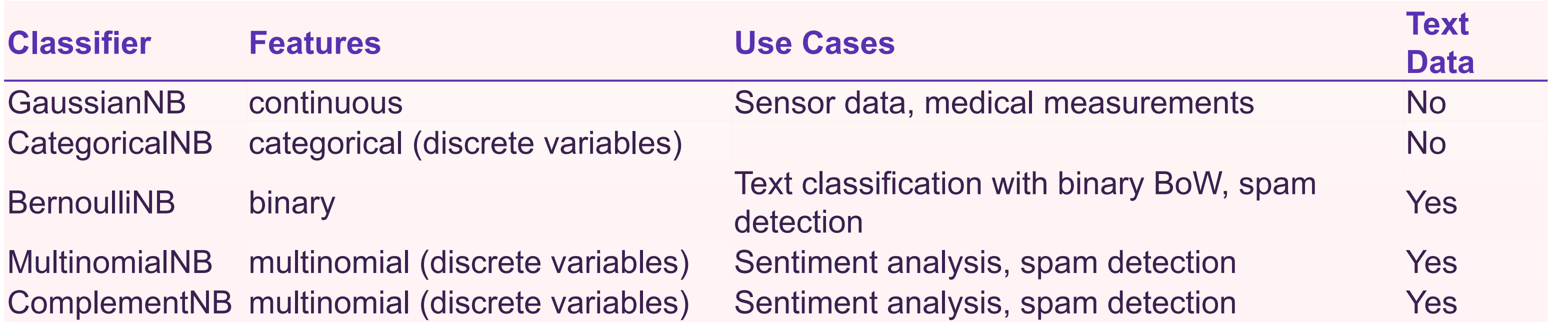
Baseline Model

- KNN with TF-IDF
- Consider 5 neighbors
- Baseline accuracy value:

KNN-English	0.497556
KNN-French	0.522222

Naive Bayes

- ❑ ComplementNB
- ❑ MultinomialNB [6]



Classifier	Features	Use Cases	Text Data
GaussianNB	continuous	Sensor data, medical measurements	No
CategoricalNB	categorical (discrete variables)		No
BernoulliNB	binary	Text classification with binary BoW, spam detection	Yes
MultinomialNB	multinomial (discrete variables)	Sentiment analysis, spam detection	Yes
ComplementNB	multinomial (discrete variables)	Sentiment analysis, spam detection	Yes

Naive Bayes

□ ComplementNB

□ MultinomialNB [6]

Model	**Accuracy**
ComplementNB-BoW-English	0.583111
ComplementNB-TF-IDF-English	0.578
ComplementNB-Word2Vec-English	0.507111
MultinomialNB-BoW-English	0.588222
MultinomialNB-TF-IDF-English	0.580667
MultinomialNB-Word2Vec-English	0.507111

MultinomialNB-BoW-English

Class	Precision	Recall	F1-Score	Support
0 (Negative)	0.63	0.65	0.64	1500
1 (Neutral)	0.48	0.49	0.49	1500
2 (Positive)	0.66	0.62	0.64	1500

Model	**Accuracy**
ComplementNB-BoW-French	0.609333
ComplementNB-TF-IDF-French	0.607556
MultinomialNB-BoW-French	0.616667
MultinomialNB-TF-IDF-French	0.616444

MultinomialNB-BoW-French

Class	Precision	Recall	F1-Score	Support
0 (Negative)	0.66	0.65	0.66	1500
1 (Neutral)	0.5	0.49	0.49	1500
2 (Positive)	0.68	0.71	0.7	1500

SVM Model

- Support Vector Classifier: Finds optimal boundary (hyperplane) that maximally separates classes in Data.

- Why SVC?

✔ non-linear

✔ Effective in high-dimensional data.

- Vectorization Techniques:

✔ BOW

✔ TF-IDF

✔ Word2Vec

- Hyperparameter Tuning
- GridSearchCV
- Grid Search Method

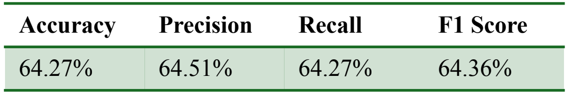
✔ C (regularization strength)

✔ kernel type

- Drawbacks :
- Computationally intensive:
- Large datasets
- Complex kernels (e.g., RBF)
- Extensive grid search

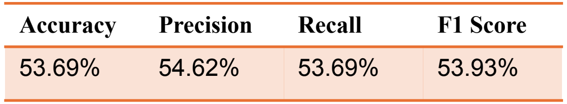
- Key insights :

✔ High : French – TF-IDF →



Accuracy	Precision	Recall	F1 Score
64.27%	64.51%	64.27%	64.36%

✔ Low : English –Word2Vec →



Accuracy	Precision	Recall	F1 Score
53.69%	54.62%	53.69%	53.93%

Random Forest

- An ensemble learning method based on multiple decision trees.

- ✔ Uses bagging and feature randomness to build diverse trees and reduce overfitting.
- ✔ Final prediction is determined by Majority voting for classification.

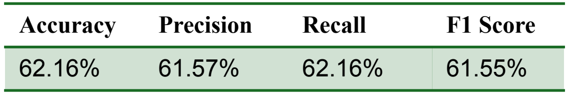
- Vectorization Techniques:
- ✔ BOW
- ✔ TF-IDF
- ✔ Word2Vec

- Drawbacks :
- Computationally intensive:
- Large datasets or many trees
- interpretability can decrease with many trees.

- Hyperparameter Tuning
- ✔ Grid Search Method :
- ✔ n_estimators
- ✔ max_depth
- ✔ min_samples_split

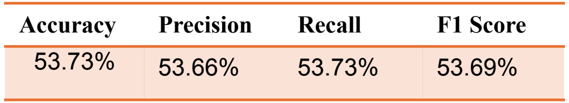
- Key insights:

- ✔ High : French- BOW →



Accuracy	Precision	Recall	F1 Score
62.16%	61.57%	62.16%	61.55%

- ✔ Low : English- Word2Vec →



Accuracy	Precision	Recall	F1 Score
53.73%	53.66%	53.73%	53.69%

LSTM Neural Network

- A deep learning method based on bidirectional LSTM networks.

✔ Processes sequences in both forward and backward directions to capture comprehensive context.
✔ Final prediction is determined by Majority voting for classification.

- Vectorization Techniques:

✔ Word2Vec
  ✔ Into an embedding layer in the LSTM

- Drawbacks :
  - Computationally intensive
  - More complex to setup and interpret

- Hyperparameter Tuning

✔ Embed size (small for low computation time)
✔ Dense layers
✔ Dropout
✔ Activation functions
✔ Optimizer functions

- Key insights:

✔ High : French- BOW →

Accuracy	Precision	Recall	F1 Score
70.00%	66.60%	58.30%	52.10%

✔ Low : English- Word2Vec →

Accuracy	Precision	Recall	F1 Score
68.50%	45.70%	57.00%	50.70%

Bert

Model: "nlptown/bert-base-multilingual-uncased-sentiment"

🧠 A pretrained multilingual BERT model on product and restaurant reviews.
🌍 Supports multiple languages (including English, French, Spanish, German, and Italian).

- ✔ Better overall performance in French, especially in Neutral class (Class 1), where English struggles with recall.
- ✔ Positive class (Class 2) has high recall and F1 in both languages, showing robustness in detecting positive sentiment.
- ✔ Multilingual BERT handles both languages well.

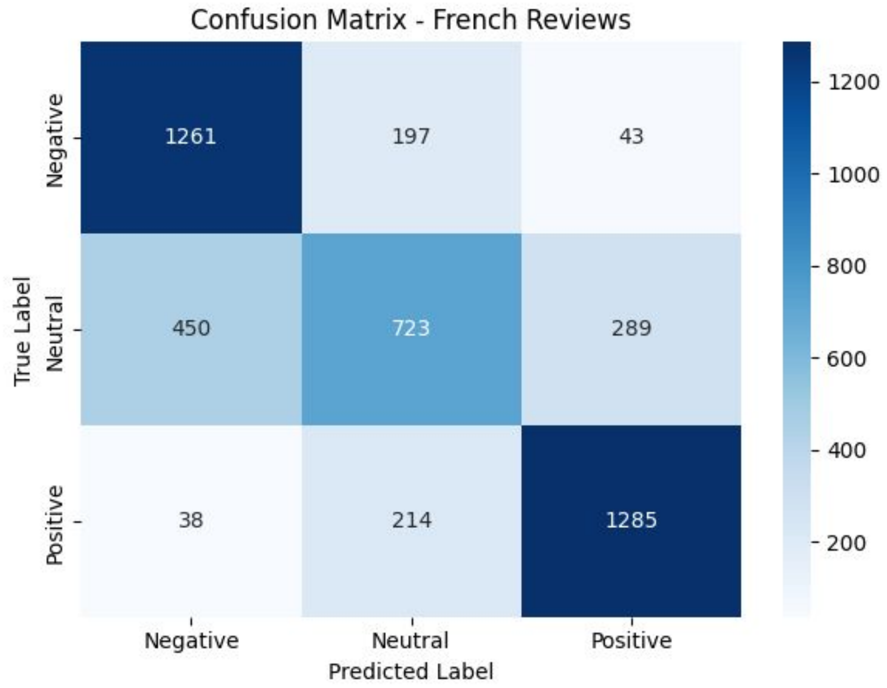
BERT-English

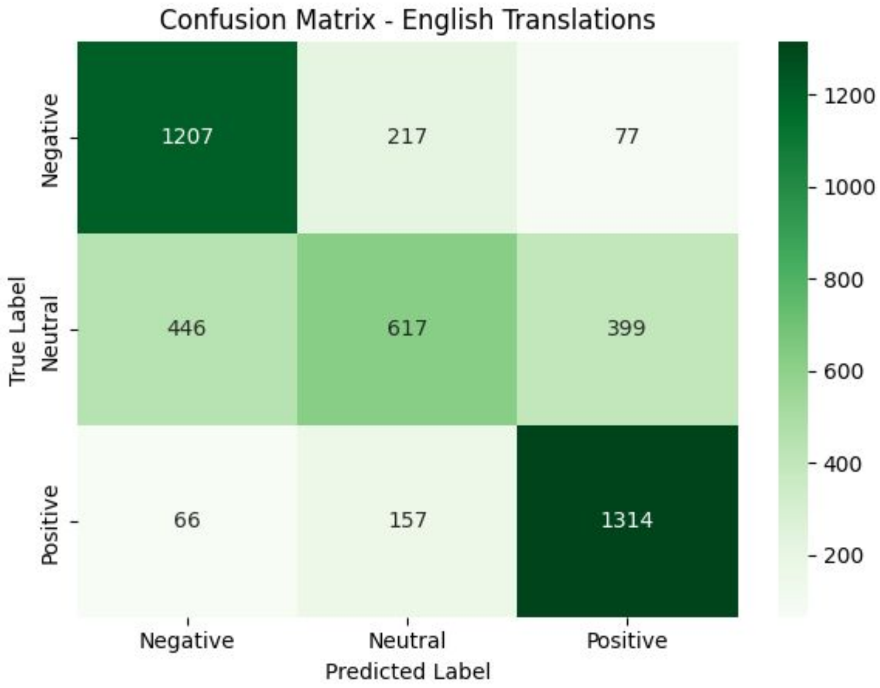
Class	Precision	Recall	F1-Score	Accuracy
0 (Negative)	0.702	0.804	0.750	69.7 %
1 (Neutral)	0.623	0.422	0.503	69.7%
2 (Positive)	0.734	0.855	0.790	69.7%

BERT-French

Class	Precision	Recall	F1-Score	Accuracy
0 (Negative)	0.721	0.840	0.776	72.6%
1 (Neutral)	0.638	0.495	0.557	72.6%
2 (Positive)	0.795	0.836	0.815	72.6%

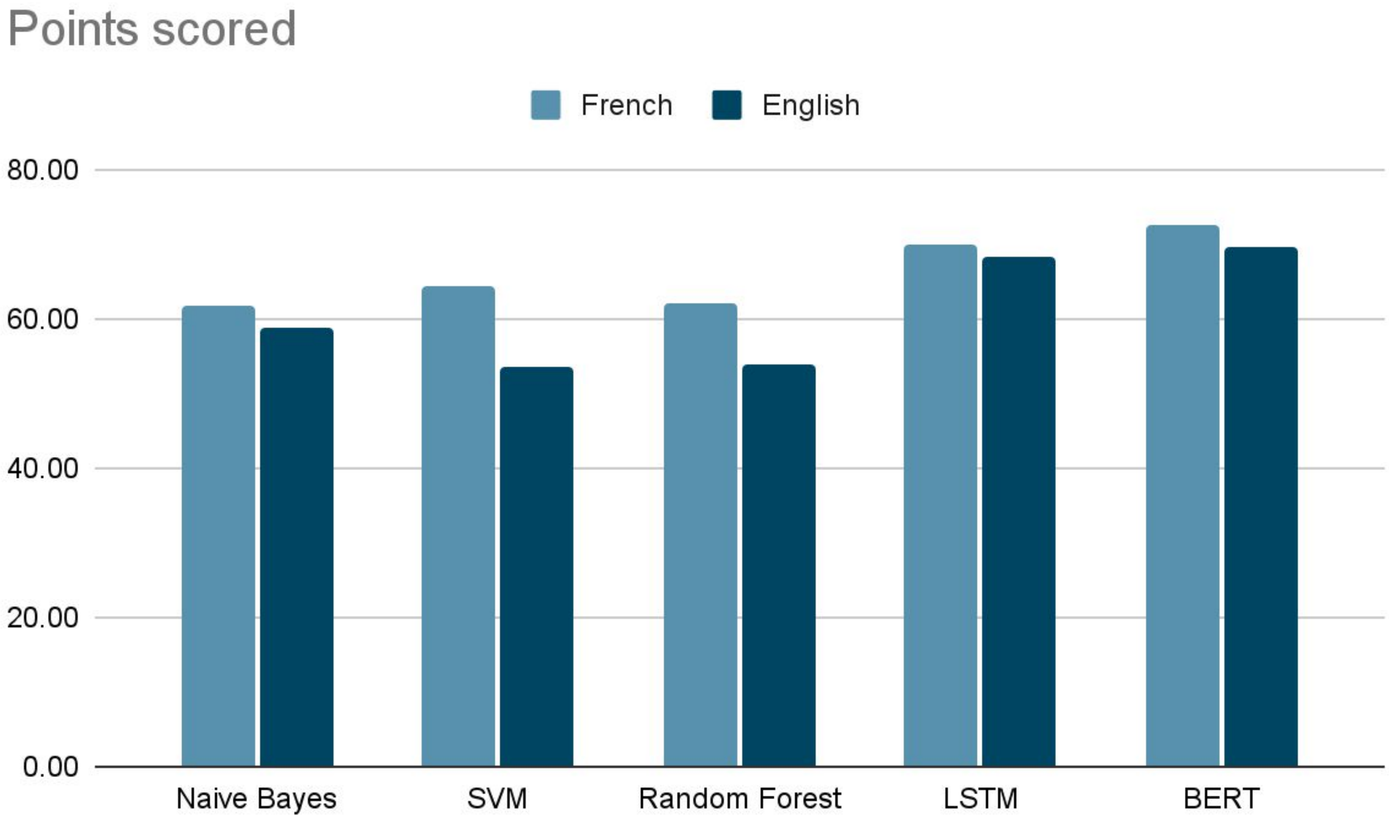
Bert


Confusion Matrix - French Reviews
True Label
Negative
Neutral
Positive
1261
197
43
450
723
289
38
214
1285
Negative
Neutral
Positive
Predicted Label
1200
1000
800
600
400
200


Confusion Matrix - English Translations
True Label
Negative
Neutral
Positive
1207
217
77
446
617
399
66
157
1314
Negative
Neutral
Positive
Predicted Label
1200
1000
800
600
400
200

Conclusion & Results

Benchmark of the models



Benchmark Conclusion

The BERT model performed the best on the dataset used.

- Self-attention mechanism
- Extensive pre-training
- Specification of language subtleties
  - Expressions
  - Sarcasm
- Transformer Architecture

<u>French data got better predictions than the English data</u>

- The initial data is French and the used data for English comes from a direct translation of the French data
- Words in english have many meanings, sometimes switching the sentiment from positive to negative
- French has more words and synonyms that give a better indication whether the meaning is positive or negative

some synonyms were maybe translated differently which changed/destabilized the prediction.

Conclusion

Our research question was…

How does sentiment vary across different languages, and can a machine learning model be trained to perform accurate sentiment analysis across multiple languages while addressing linguistic and cultural differences?

Sentiment predictions do vary depending on the language.
Whether it is because of :

- Richness of language
  - Synonyms
  - Expressions
  - Sarcasm
  - Slang
- Technological limitations
  - Negation detection
  - Lack of large datasets containing all of the above for many languages
  - Computationally intensive to preprocess (emojis, special characters, punctuation, …)

Models lack the training and methods to be versatile across languages.

Observations

What we have found with our models is that the predicted sentiment already vary with relatively simple languages such as French and English, and with already complex models implementing a part of a language’s subtleties.

Future Direction

•**Fine-Tuning Existing Models**
➤ Continue tuning hyperparameters for models like LSTM, NB, SVM, RF, and BERT to boost performance across languages.

•**Exploring New Architectures**
➤ Expand our study with additional ML and deep learning models to compare effectiveness in multilingual sentiment classification.

•**Improving Dataset Scale & Diversity**
➤ Increase the number of samples and find richer, more balanced multilingual review datasets.

•**Improving translation models and considering expressions meaning in each language**
➤ Considering word usage differences between languages and chat words and expressions while dealing with review texts

•**Utilizing HPC Resources**
➤ Plan to leverage High-Performance Computing (HPC) clusters from the Digital Research Alliance of Canada to accelerate training and experimentation.

References

1.https://www.flaticon.com/search?word=languages

2.https://www.amazon.science/publications/the-multilingual-amazon-reviews-corpus

3.T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 2018, pp. 1-6

4.B. K. Shah, A. K. Jaiswal, A. Shroff, A. K. Dixit, O. N. Kushwaha and N. K. Shah, "Sentiments Detection for Amazon Product Review," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1-6

5.https://www.kaggle.com/code/suvroo/complete-nlp-pipeline#notebook-container

6.https://www.analyticsvidhya.com/blog/2023/01/naive-bayes-algorithms-a-complete-guide-for-beginners/

7. Conneau, Alexis & Khandelwal, Kartikay & Goyal, Naman & Chaudhary, Vishrav & Wenzek, Guillaume & Guzman, Francisco & Grave, Edouard & Ott, Myle & Zettlemoyer, Luke & Stoyanov, Veselin. (2020). Unsupervised Cross-lingual Representation Learning at Scale. 8440-8451. 10.18653/v1/2020.acl-main.747.

8. H. Kumar, A. Kumar, A. Anand, A. Sabu and T. Jain, "Sentiment Analysis on Amazon Electronics Product Reviews using Machine Learning Techniques," *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2023, pp. 1-6, doi: 10.1109/GCAT59970.2023.10353467.

9.“Amazon Product Reviews: Sentiment analysis using supervised learning Algorithms,” *IEEE Conference Publication | IEEE Xplore*, Sep. 14, 2021.
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9641243

10.M. K. Nazir, C. N. Faisal, M. A. Habib and H. Ahmad, "Leveraging Multilingual Transformer for Multiclass Sentiment Analysis in Code-Mixed Data of Low-Resource Languages," in *IEEE Access*, vol. 13, pp. 7538-7554, 2025, doi: 10.1109/ACCESS.2025.3527710.

11. Yanying Mao, Qun Liu, and Yu Zhang. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. J. King Saud Univ. Comput. Inf. Sci. 36, 4 (Apr 2024).
https://doi.org/10.1016/j.jksuci.2024.102048

12. Singh, Shubham & Singla, Neetu. (2023). Sentiment Analysis on IMDB Review Dataset. Journal of Computers, Mechanical and Management. 2. 18-29. 10.57159/gadl.jcmm.2.6.230108.

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

14. (n.d.). scikit-learn: machine learning in Python — scikit-learn 1.6.1 documentation. Retrieved March 10, 2025, from https://scikit-learn.org/stable/index.html

15. API Documentation. (2024, September 30). TensorFlow. Retrieved March 10, 2025, from https://www.tensorflow.org/api_docs

16. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. [*Chapter 10 consulted*] MIT Press. http://www.deeplearningbook.org

17. Fang, X., Zhan, J. Sentiment analysis using product review data. Journal of Big Data 2, 5 (2015). https://doi.org/10.1186/s40537-015-0015-2

Thanks for Your Attention!