Journal of Social Media Research
ISSN: 3062-0945
2025, 2(4), 319-331
https://jsomer.org

REVIEW ARTICLE OPEN ACCESS

Comparison of commercial decoder-only large language models for multilingual sentiment analysis of short text

John Corcoran Burns[1]* and Tom Kelsey[1]

[1]University of St Andrews, School of Computer Science, United Kingdom
* **Correspondence:** Jack Cole Building, North Haugh, St. Andrews, KY16 9SX, UK, e-mail: burnsjack45@gmail.com Tel: +1-718-702-7724

Article Info
DOI: 10.29329/jsomer.38
Article History:
Received: 17/04/2025
Revised: 17/10/2025
Accepted: 09/11/2025
Keywords:
Multilingual Analysis
Sentiment Analysis
Social Media
Artificial Intelligence
Large Language Models

Highlights:

- Commercial Decoder - only LLMs achieve high accuracy for sentiment analysis, even though it is not their specialty.
- Higher sentiment accuracy in original language text versus translated to English text
- Different LLMs performed better on different languages; no one model was the best.
- Slang and Short texts common on social media can still cause issues for LLMs in terms of sentiment accuracy, especially in non-English languages.

Abstract

This article explores multilingual sentiment analysis of short texts using three commercial decoder-only Large Language Models ("LLMs"): OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini. The training data for these models is approximately 90% English, and it remains an open question whether it is better to evaluate text data in its original language or translate it into English first. We build on previous research on sentiment analysis of multilingual short texts, such as those found on social media, using 1000 short text samples in seven languages (English, Spanish, French, Portuguese, Arabic, Japanese, and Korean) translated into English using Google Translate. We processed these samples with decoder-only LLMs and compared their results with those from other methods (encoder-only LLMs, RNNs, lexicons). We found that decoder-only LLMs achieved the highest accuracy across all sentiment analysis methods when working with the original language data. The only exception was with the French data, where an RNN was the most accurate. Among the three decoder-only LLMs, ChatGPT had the highest accuracy in four of the seven languages, Claude in two, and Gemini, which ranked second in six of the seven languages.

Citation: Burns, J. C., Kelsey, T. (2025). Comparison of commercial decoder-only large language models for multilingual sentiment analysis of short text. *Journal of Social Media Research, 2*(4), 319-331. https://doi.org/10.29329/jsomer.38

©The Author(s). This is an **open-access** article under the **CC BY-NC-ND license** https://creativecommons.org/licenses/by-nc-nd/4.0/. The authors agree that the text, figures, and documents in the article do not violate the copyrights of third parties and that the publisher is not responsible for any claim or lawsuit by third parties due to copyright infringement. The authors declare that the publisher is not responsible for the article's content and that all responsibility for the article belongs to the authors.

1. Introduction

ChatGPT and other commercial decoder-only Large Language Models ("LLMs") have taken the world by storm with their text generation capabilities. As of October 2025, among the countries that use ChatGPT, two of the top five, the USA and India, use English either as an official language or as a widely spoken language, accounting for approximately 25% of worldwide users (Duarte, 2025). However, this creates a challenge for many commercial decoder-only LLMs, as English dominates their training data. For example, English accounts for 89% of Meta's Llama-2 model training data (Wendler et al., 2024, p. 3) and 93% of GPT -3's training data (Ochieng et al., 2024). As Nicholas and Bhatia state: "The hegemony of English data means that most large language models, even multilingual ones, are built predominantly using Standard English language text and work best in Standard English language contexts." (Nicholas & Bhatia, 2023, p. 9). This has implications, as shown by a recent paper by Mondal, which discusses mapping Covid-19 infections in California: "over 50% California's farm workers are undocumented, speaking a combination of Spanish and seven indigenous languages" (Mondal, 2021, p. 1). This highlights the need for accurate multilingual resources, especially with new LLMs, even in countries where English is widely spoken, since English might not adequately serve people.

Therefore, we aimed to analyze the multilingual capabilities of popular commercial decoder-only LLMs by examining their ability to perform sentiment analysis across different languages. Sentiment analysis involves assigning a sentiment value, such as positive, negative, or neutral, to a piece of text based on its content. This technique has many practical applications, including customer analytics, brand monitoring, market research, crime prediction, political campaigns, and financial market analysis (Rodríguez-Ibánez et al., 2023; WhyLabs, Inc., 2024). As Miah et al. describe, highly accurate sentiment analysis with decoder-only LLMs has only been demonstrated in English. This creates a significant gap in text analysis, since many languages lack labeled data for training and evaluating these models (Miah et al., 2024). However, translating text data into English can cause issues, as Nicholas and Bhatia note: "multilingual language models often rely on machine-translated text that can contain errors or terms native language speakers do not actually use" (Nicholas & Bhatia, 2023, p. 25). Therefore, enabling these models to work in their original language is important for obtaining the most accurate results.

To that end, we leverage the results of Burns, which analyzed the sentiment of short multilingual texts using encoder-only LLMs, RNNs, and lexicons in both the original language and translated into English. Burns evaluated short texts in English, Spanish, French, Portuguese, Arabic, Japanese, and Korean because, at the time of the research, these languages accounted for 85%–90% of the posts on X/Twitter (Vicinitas, 2018). Since the Vicinitas study was conducted in 2018, we used more recent data on the countries with the most Twitter users for comparison (Statista, 2025). The Statista report showed that the top countries with the most users largely overlapped with the primary languages spoken in 2018 (top five: United States, Japan, Indonesia, India, United Kingdom; however, Mexico, Spain, Brazil, Saudi Arabia, France, Canada, and South Korea also made the Top 20), so we felt confident using the same languages (Statista, 2025). For a proper comparison, we will use the same data and evaluate it with three commercial decoder-only models: OpenAI's ChatGPT-3.5, Anthropic's Claude Sonnet, and Google's Gemini-1.5-flash.

Rodríguez-Ibánez et al. demonstrated that GPT models struggled with sentiment analysis of short texts, such as tweets, in English. Therefore, we sought to investigate whether this is also the case for other widely used languages. Thus, with this article, we aim to answer three research questions: 1. Which popular decoder-only LLM model performs multilingual sentiment analysis the best? 2. Does translating the data to English improve the performance of the LLM? 3. Do LLMs perform sentiment analysis better than other methods?

2. Concepts Overview

First, a definition of LLMs will help define the terminology discussed. LLMs are machine learning/AI models based on the transformer architecture (Vaswami et al., 2017) that have been trained on vast amounts of data using deep learning and a stacked structure (multiple layers of encoders and/or decoders). These transformer LLMs have become the gold standard in all aspects of Natural Language Processing ("NLP") and Natural Language Understanding ("NLU"), encompassing tasks such as text and code generation, translation, and text classification, including sentiment analysis (IBM, 2024). However, a transformer LLM can define a wide variety of the most advanced models used today for the different tasks described. For our purposes. we will break them down into three different types: encoder-only, decoder-only, and encoder-decoder models (Hoque, 2023). Encoder-only transformer LLM models take an input of sequences, such as words in sentences, and output a vector representation of that sequence, i.e., a label for the text. These models excel in text classification, sentiment analysis, and spam detection (Hoque, 2023), with the best-known models being BERT (Devlin et al.,

2018) and its variants. We will compare the results of our study with the results of Burns who evaluated their test data with BERT models.

Decoder-only LLM models are the focus of our study. They take in a vector of text and use that vector to generate a sequence of text, successively building off the vector (in this case, a prompt) and what has already been generated (Hoque, 2023). These decoder-only models excel in text generation tasks, such as code generation, text summarization, and prose construction. However, as these models have expanded, they can also perform text classification tasks, such as sentiment analysis. Lastly, the encoder-decoder model is the final type of transformer LLM, which leverages the encoder's ability to process a sequence and transform it into a vector, and the decoder's ability to generate text based on that encoded vector. Therefore, the encoder-decoder architecture is well-suited for translation tasks (Hoque, 2023). Famous encoder-decoder transformer LLMs include Google Translate (Wu et al., 2016), which uses a hybrid of transformer and RNN architecture (Vivien, 2025), Google's T5 model (Xue et al., 2021), and BART (Lewis et al., 2019).

3. Literature Review

With the rise of LLMs as widely accessible tools, many researchers have begun exploring their capabilities across various fields. For example, Uddin and Sarker used LLMs to detect phishing emails. Brand et al. used GPT-3.5 for marketing research purposes, having the model generate responses to a survey to identify potential consumer behaviors and preferences. Closer to our work, Wahidur et al. utilized encoder-only LLMs (MiniLM) to analyze sentiment in cryptocurrency-related tweets, achieving the highest sentiment accuracy. Other studies have also examined multilingual sentiment analysis with decoder-only LLMs. For instance, Koto et al. used 34 language lexicons, each containing words with sentiment values, to pretrain various LLMs (GPT–3.5, BLOOMZ, XGLM). They found that pretraining models achieved higher sentiment accuracy than models fine-tuned only on English sentiment datasets. Additionally, Sudhir and Suresh conducted a comprehensive study on different approaches to sentiment analysis and the accuracy of various methods, including lexicons, support vector machines, RNNs, and encoder-based LLMs like BERT.

However, four studies provided the main background for our research. Following Abbas et al., we also utilized OpenAI's ChatGPT–3.5, Anthropic's Claude Sonnet, and Google's Gemini-1.5-flash, which were among the most popular LLM models available at the time of our study (Abbas et al., 2024; Luna, 2024). Because these are three of the most widely used LLM models, we aimed to explore their capabilities. Second, we sought to build on the work of Kheiri and Karimi, as well as Rodríguez-Ibánez et al., both of whom compared LLMs like GPT to other models and methods, such as BERT and LSTM layers in RNNs. These methodologies served as the basis for comparing our results to those of Burns, expanding the analysis to a multilingual context. Finally, we incorporated Miah et al.'s study, in which they translated their data into English using Google Translate to assess the sentiment accuracy of the LLMs, including ChatGPT. As described by Miah et al., the neural network driving Google Translate has been fine-tuned for translation across multiple languages, with a focus on complex sentences and subtle details in the text (Miah et al., 2024, p. 7). This approach also enables us to compare our findings with Burns, who similarly translated their data into English using Google Translate.

Through a combination of these methodologies, we aim to expand our understanding of the research on LLMs' multilingual capabilities. As we have demonstrated, this is an important field because many languages—and thus parts of the world—risk being left behind as LLMs continue to advance. By comparing the sentiment accuracy of the baseline capabilities of decoder-only LLMs across some of the most widely used models, we will gain a clearer understanding of their usefulness and potential areas for improvement. Our research differs from previous work by taking a broader approach to testing these models' sentiment analysis abilities: comparing sentiment analysis across seven languages, examining different text lengths, and assessing how well they analyze original-language text compared to translated English text.

4. Data and Methodology

For our methodology section, we divided it into our data sources and how we employed the various LLMs and other models on the data.

4.1. Data

For the sentiment analysis data, we used data created for the research conducted in Burns and adapted it as input to the decoder-only LLMs for the seven languages of our study (English, Spanish, French, Portuguese, Arabic, Japanese, and Korean). These seven languages are among the most widely spoken languages globally, with diverse writing systems, which should adequately test the various LLM models. Each data set consisted of

1000 labeled samples, with approximately 50% Positive and 50% Negative text. For English, we used Hammer's dataset, which included both positive and negative customer tweets directed at airlines. For Spanish, Bello used a labeled corpus of short hotel-review texts. For French, the Gamebusterz short text data sets were used. For Portuguese, Dias had annotated samples of short texts for use. For Arabic, Alomari had sufficient data samples to create the 1,000 required samples. For the Japanese, the Chakki-works team had the "chABSA-dataset" on GitHub, which contained enough text data for testing. For Koreans, Park used the "testing" dataset for sampling. Lastly, we performed basic data cleaning on the English translation of the data, removing punctuation, stop words, and non-letter characters. This data-cleaning process improves evaluation efficiency, but it may affect text coherence, especially in short texts like those found on social media. However, words with sentiment value are kept, and modifier words such as "not" or "but" remain to ensure the true meaning of the text is preserved.

Additionally, we translated the texts into English for this study because, as Dave explains, "the tools continue to work the best in English" (Dave, 2023). We wanted to see if there have been any improvements on this front since Dave's stated position in 2023. We used Google Neural Machine Translation (i.e., Google Translate) as our translator to enable a better comparison with the results from Burns and Miah, who also used Google Translate. It is worth noting that Google Translate is not perfect. The study by Taira et al. (2021) shows that Google Translate's overall accuracy for hospital discharge forms across six languages was 82.5% (with Spanish at 94% and Korean at 82.5%) (Taira et al., 2021, p. 1). Google Translate also struggles with colloquial and slang words often found on social media, with one study finding only an average accuracy of 72% for colloquial speech across three European languages (Mykhalevych, 2024). These issues will affect the models' sentiment accuracy. However, since each model will work off the same English text data, they will all be affected equally, allowing for comparison. Nonetheless, this inaccuracy should be noted, as it could be a confounding factor in any potential decrease in accuracy during this subset of tests.

4.2. Methodology

We employed three of the most widely used commercial LLM models: ChatGPT 3.5, Claude Sonnet 3, and Gemini-1.5-Flash. For ChatGPT[1] and Claude[2], both models had websites that let us access the LLMs, enabling us to import parts of the datasets for evaluation. Following Zhang et al., we used zero-shot learning for our sentiment analysis evaluation, including only "essential components in the prompt, namely the task name, task definition, and output format" (Zhang et al., 2023). Thus, for our prompts, we used the phrase: "Can you do sentiment analysis on these sentences if it is positive or negative, returning only the sentence number and the sentiment label, please?" with subsets of 200 data points for each language (these subsets were used because the chat functions could not handle all 1000 data points at once). A total of 1,000 data points were used, as this number was consistent with the analyses by Miah et al. and Zhao et al. Additionally, each language was evaluated individually, and we did not mix languages within the text or within the dataset. Additionally, this prompt was chosen for its simplicity. We wanted to see how well the models would handle a simple prompt without much instruction; could they follow the prompt and output answers within expectations? The models' chats would output "Positive," "Negative," or "Neutral" sentiment for each text. However, our test data had only "Positive" or "Negative" labels. Therefore, when "Neutral" was labeled by the chat, we counted this as an incorrect label and marked it as the opposite of the original data label. While this could affect overall accuracy, it was necessary because the models still produced the "Neutral" label even when prompted only to return "Positive" or "Negative," indicating they do not always follow our prompt. Since these experiments were all zero-shot, we did not rerun the batches multiple times. As described by Naveed et al. LLM models have shown improvements in few-shot learning because they are meta-learners, meaning they can learn from their previous outputs (Naveed et al., 2024, p. 23). While few-shot learning might improve model performance, we focused on examining the zero-shot capabilities of all three LLMs. As Rusnachenko found, "zero-shot application of LLMs is already effective for simpler sentiment classification tasks, such as binary and trinary classification." (Rusnachenko et al., 2024, p. 2).

While this process worked for ChatGPT and Claude, Gemini's chat functionality could not process the data in the same way. Thus, we created a Python program that called the Gemini-1.5-flash model API to do the sentiment analysis. For Gemini, we used the prompt: "Tell me whether the following sentences have sentiment that is positive or negative. If the text violates safety ratings, skip it and go on to the next sentence. Can you order by sentence number as well?" This change came from Gemini not analyzing any of the subset text data if

[1] https://chatgpt.com
[2] https://claude.ai/chats

one line of text violated the safety guidelines, for example, if the text mentions violence or contains hate speech. Thus, it was necessary to update the prompt to allow Gemini to continue analyzing the text data; however, strangely, it would also analyze the sentiment of the offending text.

Lastly, we calculated accuracy using a confusion matrix, where accuracy equals the sum of True Positives and True Negatives divided by the Total Count. We collated these results into tables and compared them with results from Burns, Chapter 3, and Miah et al. (2024) with other sentiment analysis methods.

5. Results

Below are the results from the different LLMs sentiment analysis runs. Table 1 compares the performance of all three LLMs with the original language, while Table 2 contrasts the findings from the translated text to English.

Table 1: The results from the original language sentiment data analysis with LLMs.

Language	ChatGPT	Claude	Gemini
English	**93.0%**	76.4%	89.6%
Spanish	83.0%	**84.6%**	83.8%
French	**64.6%**	56.5%	57.6%
Portuguese	**95.0%**	83.8%	90.8%
Arabic	**95.9%**	80.9%	92.3%
Japanese	87.6%	**90.4%**	79.5%
Korean	**84.9%**	76.2%	84.7%

Note: The grey highlight and bold text represent the most effective method for the language overall, as indicated across all four tables. The bold text only represents the best accuracy across the three LLM models and between the original text data and translated-to-English text data but was not the best accuracy overall across the four tables.

Table 2: The results from the translated-to-English sentiment data analysis with LLMs.

Language	ChatGPT	Claude	Gemini
Spanish	83.3%	80.3%	83.2%
French	59.5%	**57.1%**	56.7%
Portuguese	91.9%	**88.6%**	**92.0%**
Arabic	81.4%	78.8%	91.1%
Japanese	**89.6%**	87.8%	50.7%
Korean	80.7%	64.7%	82.8%

Note. The blue and bold text is the model-language combination that had a higher accuracy when the text was translated to English.

Below are the results from Burns on other sentiment analysis methods with different languages. We contain the information for these different models in the Appendix. Table 3 shows the results of the original language sentiment analysis with an RNN trained on the original language, a BERT model trained on the original language, and a multilingual BERT model trained on tweets from multiple languages (Barbieri et al., 2022).

Table 3: The other methods original language sentiment analysis results.

Language	Original RNN	Original BERT	Multilingual BERT
English	66.1%	68.5%	80.8%
Spanish	71.7%	61.4%	56.6%
French	**81.2%**	58.0%	65.1%
Portuguese	73.8%	22.9%	70.1%
Arabic	72.6%	84.6%	79.1%
Japanese	66.8%	68.9%	22.0%
Korean	82.9%	4.0%	72.5%

French for Original RNN is highlighted because it was the best overall across the four tables. Note that the Korean Original BERT was trained on Korean financial data, which differed significantly from our testing data and may have contributed to the low accuracy (Burns, 2024). Additionally, the Multilingual BERT model was not fine-tuned on Japanese text data, which also contributed to the low accuracy (Barbieri et al., 2022).

For Table 4, the results from the other sentiment analysis methods applied to the translated-to-English text are presented: one is a sentiment lexicon called VADER (Hutto & Gilbert, 2014), another is an RNN trained on English text, and a BERT model also trained on English text. A full description of these models used in Table 4 can be found in the Appendix.

Table 4: The other sentiment analysis method applied to the translated-to-English data.

Language	English VADER	English RNN	English BERT
English	75.4%	66.1%	68.5%
Spanish	69.6%	78.8%	80.2%
French	64.8%	58.4%	57.2%
Portuguese	74.9%	66.2%	73.7%
Arabic	74.7%	66.5%	72.8%
Japanese	67.0%	57.2%	53.3%
Korean	64.9%	71.3%	61.8%

For our first research question, we found that out of the seven languages, ChatGPT 3.5 had the overall best model in five out of seven languages for LLM results in the original language (English, French, Portuguese, Arabic, and Korean). Claude Sonnet 3 obtained the best results for Japanese and Spanish, while Gemini-1.5-flash was second most accurate for six out of seven languages (only in Japanese did Gemini come third). The average accuracy for the three LLMs was very high, with ChatGPT at 86.3%, Claude at 78.4%, and Gemini at 82.6%. These averages are brought down by the outlier of French, which we explain later; however, once that outlier is removed, the average accuracy increases to 89.9% for ChatGPT, 82.1% for Claude, and 86.8% for Gemini. Additionally, the ChatGPT English results aligned with the findings of Ye et al., who also achieved high accuracy with ChatGPT 3 models.

Regarding the second research question, whether translating to English would improve results, we obtained somewhat mixed outcomes. For ChatGPT 3.5 and Gemini, five out of six languages showed higher accuracy when evaluated in their original languages (only Japanese for ChatGPT and Portuguese for Gemini were more accurate when translated into English). Conversely, Claude had slightly better results when translated into English, with two out of six languages showing increased accuracy. Regarding the overall translated-to-English results, ChatGPT 3.5 and Gemini tied for the most languages with the highest accuracy, each with three out of six (ChatGPT: Spanish, French, Japanese; Gemini: Portuguese, Arabic, Korean). The average accuracy was 81.1% for ChatGPT, 76.2% for Claude, and 76.1% for Gemini. Again, once the French outlier was removed, the accuracy increased to 85.4%, 80.0%, and 80.0%, respectively. These results align with the findings of Miah et al. (2024), who also used Google Translate; ChatGPT had accuracy scores ranging from 84% to 88%.

The results of the third research question were also slightly surprising. All but one language-method combination from the previous research was worse than the decoder-only LLM models. This was surprising, as encoder-only LLMs, such as BERT models, excel in sentiment analysis and can better understand the context of text (Staff, 2024), while decoder-only LLMs, like ChatGPT, focus more on text generation (Staff, 2024). There could be several factors for this, which we detail in the discussion section. Overall, these results demonstrate that LLMs outperform other methods.

6. Discussion

Overall, our results align with those of Ochieng et al., Zhang et al., and Zhao et al., who also found that GPT models performed best in sentiment analysis across multiple languages, achieving similar levels of accuracy. These decoder-only LLMs were more accurate than previous models, such as BERT models or RNNs, with a roughly 10% increase in average accuracy for each LLM (once outliers were removed). Interestingly, for languages with very short, tweet-sized text examples, such as French, we obtained similar results to those of Rodríguez-Ibánez et al. They found that GPT models had below-average recall rates (57.0% for GPT-J and 58.4% for GPT-3) for short texts (Rodríguez-Ibánez et al., 2023, p. 12) and performed worse than other models, such as RoBERTa (Liu et al., 2019). Similar discrepancies were found by Zhang, et al., who found that while the decoder-only LLMs they investigated performed on the document level, when it came to short text/sentence level text, such as a tweet, the LLMs performed badly, with ChatGPT only having 69.4% accuracy when compared to 94.2% accuracy for longer / document level text (Zhang, et al., 2023, p. 8). One potential reason for these low accuracy results across the decoder-only LLMs for shorter text is our restriction to zero-shot learning. Since we provide no examples of the short text beforehand, we give the LLMs significantly less context to work with compared to longer text analysis, making the classification of positive or negative examples more difficult, which leads to inaccuracies.

An example of this is the outlier in the French data. The relatively low accuracy for French could be due to the French text data itself rather than a failure in the French sentiment analysis by the three LLMs. As shown in Table 3, the original language RNN created for the French analysis achieved the highest accuracy across all

methods. This RNN had fewer parameters than the BERT or GPT models (full details in the Appendix), but it was trained on data more similar to the testing data than the BERT or GPT models. Additionally, this French data was generally shorter (average character length was 78, compared to 340 for Spanish) and potentially more ambiguous. Therefore, identifying sentiment might have been more challenging for the more complex models, which may not have been trained on this type of data. That said, we had a human annotator cross-reference the French text data to assess how accurately the original sentiment was captured and to remove any data that was not clearly positive or negative. After removing such data, we rechecked the accuracy of the decoder-only LLMs on the French data and found it increased to 87.2% for ChatGPT, 78.5% for Claude, and 85.5% for Gemini. These results aligned more closely with other languages. This suggests that zero-shot learning with short text might be the issue in the analysis, as less information from the text makes it harder for decoder-only LLMs to determine a clear positive or negative sentiment.

Another anomaly worth briefly mentioning is the accuracy of Gemini's Japanese-to-English sentiment analysis. The results of this test were unusually different compared to other decoder-only LLMs and when translating the same content from other languages into English that Gemini analyzed. When we reran a sample of this test data, we found that 67.5% of the text was labeled as “Neutral,” despite clear instructions to only mark texts as “Positive” or “Negative.” The Japanese short texts primarily referenced financial data, which were slightly longer (with an average of 239 characters) and contained more neutral words, potentially contributing to the high proportion of “Neutral” responses and significantly reducing accuracy according to our methodology. However, due to Gemini's “black box” design, we can only hypothesize why this happened, as the other two decoder-only LLMs easily labeled the data as “positive” or “negative,” and Gemini itself is known for its strength in analyzing financial texts (Leechewyuwasorn & Wangpratahm, 2024, p. 7).

An interesting finding was that the decoder-only LLMs performed significantly better than encoder-only models like BERT, which are typically expected to have higher accuracy for text classification compared to decoder-only models like ChatGPT. We found this result surprising because of BERT's bidirectional architecture. BERT models can process text in both directions—left-to-right and right-to-left—simultaneously, achieving a more complete understanding of the text. In contrast, decoder-only LLMs, such as ChatGPT, are designed for text generation and are unidirectional, which may limit their understanding of the text. Therefore, we would expect BERT models to exhibit greater accuracy in sentiment analysis tasks. However, our results suggest otherwise. We believe this difference is due to the size of the training data and the complexity of the models. BERT-Large has 340 million parameters (Devlin, et al., 2018), while ChatGPT 3.5 has 175 billion parameters (Brown, 2020; Ye, 2023), and Claude Sonnet 3 is estimated to have 70 billion parameters (Thompson, 2024). Unfortunately, at the moment of writing, there is no information on the parameter count for Gemini-1.5-flash, but for another Gemini model, Nano-2, it has 3.25 billion parameters (Epoch AI, 2024). These models, with parameter counts ranging from 10 to 2,000 times larger than BERT-Large and additionally trained on sentiment analysis data, could explain why decoder-only LLMs tend to be significantly more accurate than BERT models. They are trained on much larger datasets, which leads to a better understanding of text and improved classification.

There could be several reasons why the translated-to-English results were inferior to the original language. As Miah et al. describe in their paper, when translating their data into English, translation can cause a potential loss of context. Additionally, they found that translation bias can occur and lead to mistranslations when dealing with irony or slang (Miah et al., 2024, pp. 15-16). These issues are particularly relevant to our research involving short text data. With a slight change in the meaning of a word in the text, once translated, the text's meaning can change, leading to decreased accuracy, despite LLMs having significantly more training data in English.

6.1. Limitations

A potential limitation of our study relates to our chosen methodology of labeling neutral responses as incorrect. We believe this is the best way to handle neutral responses, as specified in the prompts to the LLMs, which instruct them to only return positive or negative labels, making a neutral one technically incorrect. Other options included removing neutrals altogether, but this would have artificially increased the overall accuracy of the LLMs by reducing the amount of test data. We could have matched neutral labels to either positive or negative based on our initial data, but this would have also artificially inflated the accuracy. Therefore, to avoid inflating the results, we decided to mark neutral answers as incorrect. It is essential to note that this method may decrease the factual accuracy of the decoder-only LLMs, as some of our test data could genuinely be neutral despite being labeled as either “positive” or “negative”.

In addition to our methodology, several general limitations of LLMs should be noted. One is the Hallucination problem, where decoder-only LLMs can generate incorrect information (Zhao et al., 2023, p. 62). This Hallucination issue could lead to mislabeling sentiment data and is currently being addressed across all LLM types. As described by Kheiri and Karimi, there is also a social bias component in LLMs (Kheiri & Karimi, 2023, p. 10). When trained on large amounts of text data, biases embedded in the text can influence the model's predictions, potentially causing biased answers. This could be another reason for inaccuracy in sentiment analysis. Building on the social bias issue, Wendler et al. explored what language a computer 'thinks' in. Would an LLM use English to pivot between translations, such as from German to Spanish—i.e., translate from German to English, then from English to Spanish—potentially losing context along the way? Using English as a pivot language could further decrease accuracy in sentiment analysis for languages other than English. However, Wendler et al. found that LLMs "think" in concepts and not just in English, but these concepts are biased toward English. This still introduces bias, but not as severely as the previous scenario.

The final major limitation of LLMs, arguably the most serious for this type of research, is "data contamination" (Zhao et al., 2023, p. 68). As Zhao explains, as these LLMs grow, more data is being added to their training datasets. This is a significant problem in sentiment analysis research because if the test data being analyzed were already included in the training data, the results would show very high accuracy, as the model would already have the answer. Unfortunately, we have no way of knowing whether some of the datasets we used were part of the training data, as these are all closed-source models, and we cannot verify this. This could explain the differences in accuracy between languages when using LLMs, for example, Portuguese and Arabic achieving roughly 95% accuracy with ChatGPT, while Claude only reaches around 80%-85%. If our testing data were used during training, the reported accuracy would be artificially high. Given the limited availability of labeled sentiment analysis datasets in languages other than English and the size of these decoder-only LLMs, this is a likely scenario. This problem will only worsen as ChatGPT-4 is rumored to have 1.5 trillion parameters, and the unreleased GPT-5 is expected to be even larger (Ferrer, 2024). Therefore, some of our sentiment analysis accuracy results, as well as all studies examining LLM accuracy, may be affected by this issue and should be interpreted with caution.

6.2. Future Work

While this was an extensive study of LLMs' sentiment analysis capabilities across multiple short texts in various languages, we identified several avenues for future work. One possibility is to conduct a similar study using few-shot learning. It would be interesting to see if providing examples of positive or negative answers in the prompt improves accuracy results, as Naveed et al. found. We briefly explored multi-shot training and observed mixed results. The English text analyzed by ChatGPT achieved 97% accuracy, while the Japanese text analyzed by ChatGPT resulted in 55% accuracy. In contrast, the Portuguese text analyzed by Claude achieved a 99% accuracy. A comprehensive study on multi-shot training could yield valuable insights.

Expanding to other LLM models and methods, using open-source models like Meta's Llama, similar to Ochieng et al., or as other types of models improve, such as Encoder-only models like variations of BERT, such as RoBERTa or XLM-T (Barbieri et al., 2020; Devlin et al., 2018; Liu et al., 2019), or Encoder-Decoder models like BART (Lewis et al., 2019) or mT5 (Xue et al., 2021). Some of these models are designed to handle text classification more effectively than Decoder-only models, such as ChatGPT; however, the non-decoder-only models have significantly smaller parameter sizes than ChatGPT. If these other model types increase in parameter size, they may achieve greater accuracy in sentiment analysis.

Another focus could be that most of the languages in our study are high-resource (Arabic, French, Japanese, Spanish) or extremely high-resource (English), with two being medium-resource (Portuguese and Korean), as described by Nicholas and Bhatia. A future study could examine the same accuracy tests in low-resource languages, such as Swahili, Maltese, or Burmese, to determine whether LLMs' accuracy remains high. The challenge of this future research would be finding or creating testing datasets, as it was challenging to locate sentiment analysis-specific datasets, even for high-resource languages. Additionally, we could leverage the translation capabilities of decoder-only LLMs to assess whether sentiment accuracy improves compared to Google Translate (Dreibelbis, 2024), similar to how Burns used the OPUS-MT translator (Tiedemann & Thottingal, 2020) to compare results with Google Translate.

Finally, there is an intriguing study that extends beyond sentiment analysis accuracy and expands into a multilingual context using GPT-3 (Brown et al., 2020). In Section 6 of the paper, the authors report an experiment on bias related to gender, race, and religion, specifically using sentiment analyses of GPT-3 responses to racial prompts. It would be interesting to see this explored in a multilingual setting. Would GPT

and other decoder-only LLMs respond the same way, with similar biases across different languages, or would the language influence the sentiment of the response?

7. Conclusion & Practical Implications

We have demonstrated that among the popular commercial decoder-only large language models (LLMs), ChatGPT generally exhibits the highest proficiency in multilingual sentiment analysis; however, other LLMs outperform it for specific languages. Additionally, our findings indicate that translating data does not, on average, enhance performance. Nevertheless, certain models, such as Claude with Portuguese, do achieve higher accuracy when the data is translated into English. Finally, these decoder-only LLMs outperform alternative methods in sentiment analysis, notwithstanding some outliers.

As the abundant future work highlighted above shows, this study only begins to explore the multilingual aspects of Large Language Models. We demonstrate that decoder-only LLMs achieve very high accuracy across multiple languages, outperforming other model types such as RNNs and BERT. Additionally, these decoder-only LLMs can process text in various languages with generally higher accuracy than when the text is translated to English and then analyzed. However, we found that while short texts can be analyzed with high accuracy, very short texts, such as brief tweets and more ambiguous sentiment data, can cause issues in evaluation, leading to significantly decreased accuracy for these decoder-only LLMs. Overall, at the document level, decoder-only LLMs on “medium” to “high” (Nicholas & Bhatia, 2023) resource languages will yield highly accurate sentiment analysis results; however, the smaller the text, the less effective these models become.

Statement of Researchers	
Researchers’ contribution rate statement:	**Burns:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, writing- original draft preparation, writing - review & editing, data curation. **Kelsey:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, writing - review & editing
Conflict statement:	The authors declare that they have no conflict of interest.
Data Availability Statement:	The data supporting this study's findings are openly available in https://github.com/jb370/Comparison-of-Commercial
Funding:	This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Ethical Considerations:	All procedures followed were by the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000. Informed consent was obtained from all participants for being included in the study.” This research was approved by the University of St Andrews School of Computer Science Ethics Committee No. 13, dated 11/04/2022.
Author Biographies	
Dr. John Corcoran Burns	PhD Graduate from the School of Computer Science at the University of St Andrews. Specializes in natural language processing, multilingual analysis, geopolitical risk, and AI / ML.
Tom Kelsey	Professor in the School of Computer Science at the University of St Andrews. Specializes in data-driven research, using advanced computational techniques to identify, retrieve, classify, and analyze data in order to generate hypotheses and develop models. His primary interest is the modelling of human fertility, both for the healthy population and for survivors of childhood cancer.

6. References

2018 Research on 100 Million Tweets: What It Means for Your Social Media Strategy for Twitter. (2018) Vicinitas https://web.archive.org/web/20191122044543/https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets#language.

Abbas, A., Rehman, M. S., & Rehman, S. S. (2024). Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. *Cureus*, vol. 16, no. 3, doi:10.7759/cureus.55991.

AI, Epoch. (2024). *Notable Ai Models.* Epoch AI, Published: June 1st, 2024. https://epochai.org/data/notable-ai-models?view=table#explore-the-data.

Alomari, K. (2016). *Arabic-twitter-corpus-AJGT*, GitHub, https://github.com/komari6/Arabic-twitter-corpus-AJGT.

Anthropic. (2024). *Introducing the Next Generation of Claude. Announcements*. Anthropic https://www.anthropic.com/news/claude-3-family. Accessed June 1st, 2024.

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022) Xlm-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. edited by Cardiff NLP Snap Inc., *ArXiv Cornell University*, doi:10.485/arxiv.2104.12250.

Bello, H. J. (2021). *Sentiment-Spanish*. sentiment-analysis-sentiment. GitHub https://github.com/sentiment-analysis-spanish/sentiment-spanish.

Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for Market Research. edited by *Harvard Business School*, https://www.hbs.edu/faculty/Pages/item.aspx?num=63859.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kalpan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Kruger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models Are Few-Shot Learners. edited by *Johns Hopkins University & Open AI*, https://arxiv.org/abs/2005.14165.

Burns, J. C. (2024). Automatic Evaluation of Geopolitical Risk. *School of Computer Science, vol. PhD Doctor of Philosophy*, *University of St Andrews*, general editor, Tom; Donovan Kelsey, Carl, doi:https://doi.org/10.17630/sta/918.

Dave, P. (2023). *Chatgpt Is Cutting Non-English Languages out of the Ai Revolution.* Wired, May 31st, 2023. https://www.wired.com/story/chatgpt-non-english-languages-ai-revolution/.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. edited by *Google AI Language*, *ArXiv,,* doi:10.48550/arxiv.1810.04805.

Dias, F. (2021). *Brazilian Portuguese Sentiment Analysis Datasets* Kaggle, https://www.kaggle.com/datasets/fredericods/ptbr-sentiment-analysis-datasets?select=olist.csv

Dreibelbis, E. (2024). *Google Translate Vs. Chatgpt: Which One Is the Best Language Translator?* Contented Traveler. PCMag https://www.pcmag.com/articles/google-translate-vs-chatgpt-which-is-the-best-language-translator. Accessed June 1st, 2024.

Duarte, F. (2025). *Number of Chatgpt Users (October 2025).* explodingtopics. https://explodingtopics.com/blog/chatgpt-users. Accessed October 17th, 2025.

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., & Kyriazis, D. (2023) Transforming Sentiment Analysis in the Financial Domain with Chatgpt. *Machine Learning with Applications*, vol. 14, https://www.sciencedirect.com/science/article/pii/S2666827023000610.

Ferrer, J. (2024). *Everything We Know About Gpt-5.* Artificial Intelligence (AI). Data Camp https://www.datacamp.com/blog/everything-we-know-about-gpt-5. Accessed June 1st, 2024.

Gamebusterz (2017) *xae, French-Sentiment-Analysis-Dataset*, GitHub, https://github.com/gamebusterz/French-Sentiment-Analysis-Dataset/blob/master/xae

Gamebusterz (2017) *xak*, *French-Sentiment-Analysis-Dataset*, GitHub, https://github.com/gamebusterz/French-Sentiment-Analysis-Dataset/blob/master/xak

Google (2024). *Gemini Apps Faq.* Google Gemini https://gemini.google.com/faq?gad_source=1&gclid=EAIaIQobChMIy9ffvv7jhgMV8FBHAR3GfAjyEAAYASABEgKyf_D_BwE. Accessed June 1st, 2024.

Hammer, B. (2016). *crowdflower-airline-twitter-sentiment,* Kaggle, crowdflower.com, https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

Hoque, M. (2023). *A Comprehensive Overview of Transformer-Based Models: Encoders, Decoders, and More.* Medium, https://medium.com/@minh.hoque/a-comprehensive-overview-of-transformer-based-models-encoders-decoders-and-more-e9bc0644a4e5. Accessed June 1st, 2024.

Hutto, C. J., & Gilbert, E. (2014). Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, http://eegilbert.org/papers/icwsm14.vader.hutto.pdf.

IBM. (2024) *What Are Large Language Models (Llms)?* IBM https://www.ibm.com/topics/large-language-models. Accessed June 1st, 2024.

Kheiri, K., & Karimi, H. (2023) Sentimentgpt: Exploiting Gpt for Advanced Sentiment Analysis and Its Departure from Current Machine Learning. edited by *Utah State University*, *ArXiv,* 2023. https://arxiv.org/abs/2307.10234.

Koto, F., Beck, T., Talat, Z., Gurevych, I., & Baldwin, T. (2024) Zero-Shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon. edited by *MBZUAI Department Natural Language Processing*, *ArXiv,* https://arxiv.org/abs/2402.02113.

Kubo, T., Nakayama, H., & Kamura, J. (2018) *chABSA-dataset* Github, https://github.com/chakki-works/chABSA-dataset

Leechewyuwasorn, D. & Wangpratahm, N. (2024) Comparative Analysis of Financial Sentiment Analysis Models for the Thai Stock Market: Traditional NLP vs. GPT vs. Gemini. *QuantCorner Laboratory* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4921837

Leading countries based on number of X (formerly Twitter) users as of February 2025 (2025). Statista https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019) Bart: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. edited by *Facebook AI, ArXiv*, https://arxiv.org/pdf/1910.13461.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O, Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019) Roberta: A Robustly Optimized Bert Pretraining Approach. edited by *Facebook AI, ArXiv*. https://arxiv.org/pdf/1907.11692.

Luna, J. C. (2024) *The Top 10 Chatgpt Alternatives You Can Try Today*. Artifical Intelligence. datacamp https://www.datacamp.com/blog/10-chatgpt-alternatives. Accessed June 1st, 2024.

Mykhalevych, N. (2024) *How Accurate is Google Translate?* Preply. https://preply.com/en/blog/how-accurate-is-google-translate/

Miah, M.S.U., Kabir, M.M., Sarwar, T.B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024) A Multimodal Approach to Cross-Lingual Sentiment Analysis with Ensemble of Transformer and Llm. *Scientific Reports*, vol. 14, doi:https://doi.org/10.1038/s41598-024-60210-7.

MindsDB, M., P. C., & Slawinska, M. (2024). Updated June 2024: A Comparative Analysis of Leading Large Language Models. *MindsDB* https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models. Accessed June 1st, 2024.

Mondal, A. (2021). The Importance of Community Engagement on Covid-19 Vaccination Strategy: Lessons from Two California Pilot Programs. *EClinicalMedicine, Part of The Lancet Discovery Science*, vol. 32, https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00034-1/fulltext.

Naveed, H., Khan, A. U., Qui, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024) A Comprehensive Overview of Large Language Models. edited by *Lahore University of Engineering and Technology (UET), Pakistan*, *ArXiv*, https://arxiv.org/pdf/2307.06435.

Nicholas, G., & Bhatia, A. (2023) Lost in Translation: Large Language Models in Non-English Content Analysis. *AI Policy & Goverance, Center for Democracy & Technology,* https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf.

Ochieng, M., Gumma, V., Sitaram, S., Wang, J., Chaudary, V., Ronen, K., Bali, K., & O'Neill, J (2024) Beyond Metrics: Evaluating Llms' Effectiveness in Culturally Nuanced, Low-Resource Real-World Scenarios. edited by *University of Washington Microsoft Corporation*, *ArXiv,* https://arxiv.org/html/2406.00343v2.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I, Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I.,… Zoph, B. (2023) Gpt-4 Technical Report. edited by *Open AI, ArXiv,* https://arxiv.org/pdf/2303.087742.

Park, L. (2015) *Nsmc,* GitHub https://github.com/e9t/nsmc.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, L., Li, W., & Liu, P (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, vol. 21, pp. 1-67, https://arxiv.org/pdf/1910.10683.

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P.M. (2023) A Review on Sentiment Analysis from Social Media Platforms. *Expert Systems with Applications*, https://www.sciencedirect.com/science/article/pii/S0957417423003639.

Rusnachenko, N., Golubev, A., & Loukachevitch, N. (2024) Large Language Models in Targeted Sentiment Analysis. edited by *England Newcastle Upon Tyne, United Kingdom*, *ArXiv*, https://arxiv.org/abs/2404.12342.

Sherstinsky, A. (2020) Fundamentals of Recurrent Neural Network (Rnn) and Long Short-Term Memory (Lstm) Network. *Physica D: Nonlinear Phenomena*, vol. 404, https://arxiv.org/pdf/1808.03314.

Staff, Coursera. (2024) *Bert Vs. Gpt: What's the Difference?* AI and Machine Learning. Coursera https://www.coursera.org/articles/bert-vs-gpt. Accessed June 1st, 2024.

Sudhir, P., & Suresh, V. D. (2021) Comparative Study of Various Approaches, Applications and Classifiers for Sentiment Analysis. *Global Transitions Proceedings*, vol. 2, no. 2, pp. 205-11, doi:https://doi.org/10.1016/j.gltp.2021.08.004.

Taira, B. R., Kreger, V., Orue, A., & Diamond, L. C. (2021). A Pragmatic Assessment of Google Translate for Emergency Department Instructions. Journal of General Internal Medicine, vol. 36. https://link.springer.com/article/10.1007/s11606-021-06666-z

Tiedemann, J., & Thottingal, S. (2020). Opus-Mt — Building Open Translation Services for the World. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT),* https://github.com/Helsinki-NLP/Opus-MT.

Thompson, A. (2024). The Memo - Special Edition: Claude 3 Opus. *LifeArchitect* https://lifearchitect.substack.com/p/the-memo-special-edition-claude-3. Accessed June 1st, 2024.

Uddin, M. A., & Sarker, I. H. (2024). An Explainable Transformer-Based Model for Phishing Email Detection: A Large Language Model Approach. edited by *BGC Trust University Bangladesh Department of Computer Science and Engineering*, *ArXiv*. https://arxiv.org/abs/2402.13871.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I (2017). Attention Is All You Need. edited by *Google Brain Google Research*, *ArXiv,* https://arxiv.org/abs/1706.03762.

Vivien, L. (2025). *Inside Google Translate – The Technology Behind*. La Vivien Post. https://www.lavivienpost.com/google-translate-and-transformer-model/ Accessed October 17th, 2025

Wahidur, R. S., Tashdeed, I., Kaur, M. & Lee, H. N. (2024). Enhancing Zero-Shot Crypto Sentiment with Fine-Tuned Language Model and Prompt Engineering. *IEEE Access,* https://ieeexplore.ieee.org/iel7/6287639/10380310/10382518.pdf.

Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024). Do Llamas Work in English? On the Latent Language of Multilingual Transformers." *ArXiv*, doi:https://doi.org/10.48550/arXiv.2402.10588.

WhyLabs, I. (2024). *Sentiment Analysis with Large Language Models (Llms). Use Cases of Large Language Models (LLMs).* WhyLabs Inc https://whylabs.ai/learning-center/llm-use-cases/sentiment-analysis-with-large-language-models-llms.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., & Norouzi, M. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. edited by *Google, ArXiv*, https://arxiv.org/pdf/1609.08144.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). Mt5: A Massively Multilingual Pre-Trained Text-to-Text Transformer. edited by *Google Research, ArXiv*, https://arxiv.org/abs/2010.11934.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., & Huang, X (2023). A Comprehensive Capability Analysis of Gpt-3 and Gpt-3.5 Series Models. edited by *Fudan University, ArXiv*, https://arxiv.org/abs/2303.10420.

Zelieska, L. (2024). *Top Models for Natural Language Understanding (Nlu) Usage.* Quantpedia https://quantpedia.com/top-models-for-natural-language-understanding-nlu-usage/#:~:text=RoBERTa%20(A%20Robustly%20Optimized%20BERT,extensive%20and%20optimized%20pretraining%20process.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment Analysis in the Era of Large Language Models: A Reality Check. Edited by *Alibaba and Nanyang Technological University*, *ArXiv,* https://arxiv.org/abs/2305.15005 .

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y,, Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X,. Liu, Z., Liu, Z., Liu, P., Nie, J. Y., & Wen, J. R. (2023) A Survey of Large Language Models. Edited by *Renmin University of China, Gaoling School of Artificial Intelligence and School of Information, ArXiv*. https://arxiv.org/pdf/2303.18223.

Appendix

Below are the breakdowns of the Recurrent Neural Networks and BERT models that were used in the comparison in Tables 3 and 4. As Table 5 shows, most RNNs only have roughly 117 million parameters, which are significantly fewer parameters than the LLMs. The different BERT models were all fine-tuned on various languages but based on the base-BERT model detailed in Table 6. Except for the multilingual model, which has 550 million parameters, as it contains many languages and was fine-tuned on multilingual Twitter data (Barbieri, 2022). A note for Table 4, the results were produced by the English model in Table 5, and the English model of Table 6. As for VADER mentioned in Table 4, it is a sentiment lexicon, not a model, validated by humans, with various heuristics to better label speech, especially the kind found in social media, for full details, please reference Hutto & Gilbert, 2014 (Hutto & Gilbert, 2014, pp. 1, 6).

Table 5: These are the hyperparameters for the RNNs models that were constructed in Burns. While the data size varies for each of the RNNs models this was done to make sure that most amount of data that could be used to create the most accurate RNN model with a balanced data set. For full details please reference Burns, 2024.

Language	Total Params	Data Size	Layers	Batch Size	Epochs	Data Balance
English	116,965,633	50,000	4	4	5	50%
Spanish	116,913,121	5,764	5	2	7	50%
French	116,965,633	415,702	4	128	7	50%
Portuguese	116,938,081	200,000	6	4	5	50%
Arabic	117,012,193	102,196	6	4	5	50%
Japanese	117,558,017	20,000	5	5	7	50%
Korean	117,012,193	150,000	6	5	5	50%

Table 6: These are the hyperparameters for the BERT models that were implemented for our study. The question mark under Fine-Tuned Count for Japanese is due to the creator of the model not specifying how much data was used to fine-tune the Japanese sentiment analysis. For full details please reference Burns, 2024.

Language	Total Params	Vocab	Fine Tuned Count	Batch Size	Learning Rate	Layers
English	110M	30k	11.8k	128	3e-5	12
Spanish	110M	32k	8.4k	2,048	0.0001	12
French	110M	30k	140k	128	3e-5	12
Portuguese	110M	29.8k	.5k	16	2e-5	12
Arabic	110M	30k	23.3k	32	3e-5	12
Japanese	110M	30k	?	128	3e-5	12
Korean	102M	20k	50k	32	5e-5	12
Multilingual	550M	150k	24.3k	8,192	0.0001	24