End-to-End NLP for Automotive Insights: Sentiment Analysis, Translation, Question Answering, and Summarization of Car Reviews

Arezoo Bybordi
Professor Rozovskaya
Course: Advanced Natural Language Processing

Abstract

This project develops and evaluates a car review pipeline including the Helsinki-NLP's opus-mt-en-es model to translate English car reviews into Spanish, with translation quality assessed through BLEU scores. By implementing data preprocessing, sentiment classification, translation, question answering and summarization, we deliver high-quality sentiment analysis and reliable Spanish translations suitable for customer insights. The results demonstrate the model's effectiveness for multilingual applications within the automotive industry.

Introduction

In the automotive industry, understanding and responding to customer feedback is fundamental to enhancing product quality, customer satisfaction, and brand loyalty. Online car reviews provide a wealth of unstructured data reflecting consumer opinions on aspects like performance, aesthetics, and reliability. However, manual analysis of such reviews presents significant challenges due to their sheer volume, diversity of languages, and varying review lengths, especially for multinational companies operating across linguistic and cultural boundaries. Natural Language Processing (NLP) offers powerful tools to address these challenges, enabling the automation of sentiment analysis, machine translation, question answering, and summarization. While each of these tasks individually provides valuable insights, there has been a lack of cohesive solutions that integrate all these functionalities within a single framework;especially when applied to large datasets of multilingual car reviews.
To bridge this gap, this project introduces a unified Large Language Model (LLM)-based chatbot designed specifically for the automotive industry. The system empowers employees to determine whether a review conveys a positive or negative sentiment, translates reviews for non-English-speaking users, answers specific questions about a customer's opinion (e.g., their thoughts on color or purchase date), and summarizes lengthy reviews to extract the essence of customer feedback. This integrated approach not only simplifies the analysis of diverse feedback but also sets the stage for scalable, multilingual insights in future iterations.
The core of our solution is an end-to-end NLP pipeline leveraging state-of-the-art pretrained models from Hugging Face. These models were carefully selected for their ability to handle domain-specific language tasks effectively. By consolidating these capabilities into a single interface, we provide automotive companies with a versatile tool for extracting actionable insights, improving customer support, and streamlining feedback analysis.

This report outlines the methodology for each NLP task, evaluates model performance on the car review dataset, and demonstrates the pipeline's applicability to real-world scenarios in the automotive domain. Our contribution lies in developing an integrated NLP system that is both scalable and extendable, addressing a pressing industry need with unprecedented breadth and precision.

Related Work

Large Language Models (LLMs) and transformer-based architectures have revolutionized the field of Natural Language Processing (NLP) by delivering state-of-the-art performance across various tasks, including sentiment analysis, translation, question answering, and summarization. These advancements have also enabled LLMs to extend their applications beyond traditional NLP domains, into industries such as automotive and biology. For instance, LLMs have been employed to predict differentially methylated cytosines in TET and DNMT3 knockout mutants, demonstrating their applicability in biological research [2]. These innovations highlight the transformative potential of LLMs in addressing domain-specific challenges [1].

Sentiment analysis is a key component of this project, and extensive work has been done in this area using transformer models like BERT, RoBERTa, XLNet, ELECTRA, DistilBERT, ALBERT, T5, and GPT and showcase adaptability and high performance across various datasets, cementing their role in modern sentiment analysis tasks [3]. Specific architectures such as the TRABSA model, which combines transformer-based frameworks, attention mechanisms, and BiLSTM networks, have achieved remarkable accuracy in these tasks [6]. Moreover, LLMs demonstrate strong sentiment analysis capabilities, particularly in zero-shot settings for simpler tasks, though they face challenges with more complex structured sentiment information [7].

Machine translation has also benefited greatly from advancements in LLMs. Recent work has highlighted how models like GPT-4 enhance translation capabilities, especially in low-resource languages, by leveraging fine-tuning techniques [8]. Furthermore, these models now focus on semantic and contextual understanding rather than syntax alone, ensuring higher-quality translations suitable for real-world applications [11]. In the automotive industry, these advancements can improve access to multilingual customer feedback and enhance customer satisfaction.

Question answering has seen significant progress with the introduction of models such as InstructGPT, which has been shown to improve the accuracy of open-domain question answering tasks through fine-tuning and evaluation using human judgment [9]. Additional research in this domain has explored leveraging external knowledge, such as extracting sub-graphs from ConceptNet, to enhance a model's ability to answer open-domain questions by aligning contextually relevant entities with the input [12]. These approaches demonstrate the potential of LLMs to address domain-specific question-answering tasks, such as understanding specific aspects of car reviews.

Text summarization is another area where LLMs have excelled, supporting both extractive and abstractive techniques. Methods like LexRank, for graph-based summarization, and BART, for abstractive summarization, have shown promising results in condensing information while maintaining factual consistency [13]. These techniques are crucial in scenarios where long-form reviews must be distilled into concise and meaningful insights, as required by this project. Recent investigations further confirm that LLMs deliver robust summarization performance

across different NLP tasks, ensuring the extraction of reliable summaries from complex datasets [10].
Beyond core NLP tasks, the versatility of LLMs has been demonstrated in fields like biology and automotive. For example, ProtGPT2, based on the GPT-2 architecture, has been used to integrate protein sequence and expression data, enhancing molecular characterization in cancer research [4]. Similarly, LLMs have been employed to predict biological markers like methylation patterns [2] and improve access to information in traditional vehicle manuals, underscoring their potential for domain-specific applications [5].
This body of work demonstrates the versatility and potential of LLMs and transformer-based architectures in addressing complex and diverse NLP challenges. By leveraging state-of-the-art advancements in sentiment analysis, machine translation, question answering, and summarization, this project integrates these capabilities into a unified framework tailored for extracting actionable insights from car reviews, addressing multilingual barriers, and enhancing customer support in the automotive industry.
Despite significant progress, no prior research has unified sentiment analysis, translation, question answering, and summarization into a single framework tailored to a dataset as large and diverse as the one used in this study. By integrating these tasks into a cohesive pipeline, this project advances the state of the art, offering a scalable solution to the unique challenges posed by multilingual car reviews. This work bridges the gap between existing methods and practical, end-to-end applications in the automotive industry, paving the way for enhanced insights and customer support.

Methodology

For all of the tasks we used Hugging Face library, python programming and transformer-based models.

1. **Classifying Car Reviews:** For the task of classifying car reviews, we utilized the pretrained model and tokenizer distilbert-base-uncased-finetuned-sst-2-english [16]. This model is based on DistilBERT, a compact, faster, and lighter version of BERT (Bidirectional Encoder Representations from Transformers) [15],[18]. DistilBERT is designed to retain 97% of BERT's language understanding capabilities while being 60% faster and requiring 40% fewer parameters, making it suitable for real-time applications. The model is developed using knowledge distillation, a technique where a smaller model (the "student") learns from the outputs of a larger pretrained model (the "teacher").
This specific version of DistilBERT is fine-tuned on the Stanford Sentiment Treebank v2 (SST-2) dataset [17], which is a widely used benchmark for binary sentiment classification. The dataset comprises movie reviews labeled as either positive or negative. Since the model is "uncased," it ignores capitalization during preprocessing, reducing complexity for tasks where case sensitivity is not critical. Fine-tuning involved adapting the pretrained DistilBERT weights to the SST-2 dataset by training on sentiment classification tasks, which required adding a classification head (a dense layer) to predict sentiment polarity. The model was optimized on the labeled data using loss functions such as cross-entropy.

We applied this model to classify customer opinions about cars as either positive or negative. The Edmunds-Consumer Car Ratings and Reviews dataset from Kaggle [19] served as our primary dataset for this task. Initially, the dataset contained customer ratings on a scale of 1 to 5. We transformed these ratings into binary sentiment labels using the following formula: “Positive opinion: Ratings strictly greater than 4 and Negative opinion: Ratings of 4 or less.” These binary labels served as the ground truth. The reviews' textual content was then processed using the Hugging Face Transformers library’s AutoTokenizer class. This tokenizer converted raw textual data into smaller subword tokens and numerical IDs corresponding to a learned vocabulary.
To preprocess the data, a maximum text length of 512 tokens was set for classification. Using the review text, we predicted sentiment and compared the predictions against the ground truth to evaluate the model’s performance. This approach leverages a pretrained model fine-tuned on task-specific sentiment knowledge, providing both general-purpose language understanding and efficiency suitable for practical applications.

2.**Translating a Car Review:** The goal of this task was to translate car reviews from English to Spanish, enabling non-English-speaking employees or app users, particularly Spanish speakers, to utilize the application effectively. For this purpose, we employed the pretrained model Helsinki-NLP/opus-mt-en-es from Hugging Face. This model is part of the OPUS-MT project, which is based on a transformer architecture. It is specifically fine-tuned for English-to-Spanish translation, leveraging multilingual data to produce accurate and contextually relevant translations. The model's efficiency lies in its ability to learn from extensive parallel corpora curated for linguistic diversity.
To evaluate the performance of this model, we compared it with two of the best multilingual machine translation models available on Hugging Face: M2M100 models by Meta AI. These models are designed for direct translation between 100 languages without using English as an intermediary ("pivot"). Built on the Transformer architecture, they feature a shared vocabulary across languages. The 418M parameter version prioritizes efficiency, while the 1.2B parameter version offers higher accuracy due to its larger model size. These models were trained on a balanced dataset to ensure high-quality translations across both high- and low-resource languages [20],[21].
For the comparison, the first review in the dataset was translated using both M2M100 models and Helsinki-NLP/opus-mt-en-es. A reference list of translations was created based on the outputs from the M2M100 models (with a maximum sequence length of 30 tokens). The performance of Helsinki-NLP/opus-mt-en-es was then assessed against this reference list using the BLEU (Bilingual Evaluation Understudy) score, a widely used metric for evaluating the quality of machine-translated text by comparing it to human-generated reference translations. The BLEU score calculation was limited to the first review of the dataset to illustrate comparative performance.

3.**Question Answering About a Car Review**: This task focuses on question answering, where the goal is to answer a specific question based on the provided context — in this case, the car review. This feature is particularly beneficial for users or employees who wish to retrieve specific information from a review without reading it in its entirety.

We utilized the pretrained model deepset/minilm-uncased-squad2, available on Hugging Face. This model is a distilled version of BERT, optimized specifically for question-answering tasks. MiniLM leverages advanced techniques such as deep self-attention distillation and teacher-student frameworks, achieving a balance between computational efficiency and accuracy. It retains the core language understanding capabilities of larger models while significantly reducing resource consumption, making it ideal for deployment in resource-constrained settings [22].
The model was trained on the SQuAD2.0 (Stanford Question Answering Dataset v2.0) [23], a benchmark dataset for question answering. SQuAD2.0 introduces unanswerable questions alongside answerable ones, enhancing the model's ability to discern when no valid answer exists in the context provided.
For evaluation, we tested the model on the third review in the dataset. A specific question was posed regarding the review, and the model successfully retrieved a relevant answer. This demonstrated its effectiveness in processing review-based question-answering tasks.
To evaluate, the third review in the data, in the code, we asked a question and got relevant answer:
Question: `"How long ago was the product purchased?"`
and this is the context: `"havin a blast! gotta remind myself to be careful because it takes off like a rocket and handles like its riding on rails. just purchased 2 months ago , so don't have any comments regarding durability or reliability but so far its great. would recommend for anyone havin a midlife crisis , lol."`
Predicted Answer in the code: 2 months

4.**Summarize and Analyze a Car Review**:This task focuses on text summarization, designed to generate concise summaries of lengthy car reviews. This feature is intended to help users quickly grasp the essence of a review without needing to read it in full.
The model utilized for this task is facebook/bart-large-cnn, available in the Hugging Face library. This model is a fine-tuned variant of the BART (Bidirectional and Auto-Regressive Transformers) architecture, optimized specifically for abstractive text summarization. BART integrates the advantages of bidirectional and autoregressive models by employing a denoising autoencoder strategy during pretraining. It encodes input text into a latent representation, which is then autoregressively decoded into a coherent summary.
This architecture is particularly effective in generating human-like summaries, as it not only extracts key information but also rephrases it in a natural manner. The facebook/bart-large-cnn model has demonstrated strong performance across various natural language processing tasks, including text summarization, question answering, and text generation [24].
To demonstrate the model's capabilities, we applied it to the last review in the dataset. The output summary successfully condensed the content while retaining the core message, showcasing the model's utility for summarization tasks. This is the original review:
`"I have had several high end cars, my main focus is 4 door luxury sedan, combined with performance, this is my 2nd flying spur, 1st one was a 2006 model, 25k miles, this car gave me a lot of problems, mainly confined to the info-tainement system, though a few major engine problems did occur.Luckily it was under warranty, good job because it cost 20k plus had`

the front struts replaced, i sold it within 2 years.I purchased a high end Mercedes AMG SUV, what a mistake, how i missed the bentleys serene ride,composed road manners, the overall feeling of luxury, and secure ride. I had to buy another, there are no words to describe how it feels, other cars just don't compare.The Bentley is a very simple car , there is no lane departure, brake assist,forward cameras, in fact you feel short changed in that department, but it you will not miss them, the is a simple elegance the other cars don't have, or replicate.So i purchased the best i could, 2012 Speed, what an improvement on the 2006, much better car all round, more power, better seats, handling, just great.Buy warranty if you can, or good after market service repair, low miles, good service records are essential, and then enjoy the best ride of you life, in a way i regret getting the car, as everything else seems boring, and mundane, for the price of a new Cadillac Escalade, a used one makes sense, my car was $240k when new 3 years ago, i saved $135K over new, what a bargain. You only live once, go for it."

And this is the summarized version by the model:

"The Bentley is a very simple car. There is no lane departure, brake assist,forward cameras, in fact you feel short changed in that department, but it you will not miss them. The is a simple elegance the other cars don't have, or replicate."

Overall, this is a model that can do various tasks of sentiment analysis, translation, question answering and summarization. Which enables users to be able to do a wide range of tasks.

Dataset

The dataset used in this study is the Edmunds-Consumer Car Ratings and Reviews dataset. During the preprocessing stage, issues with file formats led to the removal of 123 rows, and some files could not be included. After cleaning and combining the data, the final dataset consisted of 1,978 rows.

This dataset contains reviews and ratings on a scale of 1 to 5. To facilitate sentiment analysis, the ratings were converted into binary sentiment labels: (Ratings above 4 were classified as positive opinions. Ratings of 4 or below were classified as negative opinions.)

For the question-answering task, the MiniLM (deepset/minilm-uncased-squad2) model was trained on the SQuAD2.0 (Stanford Question Answering Dataset) [23]. SQuAD2.0 is an extension of the original SQuAD dataset, specifically designed for machine reading comprehension. It comprises over 100,000 answerable questions and an additional 50,000 unanswerable questions, which were carefully crafted to resemble answerable ones.

This dataset challenges models to not only locate correct answers within a context but also to recognize when the context does not provide sufficient information to answer a given question. This dual requirement makes SQuAD2.0 a benchmark for developing intelligent systems capable of both answering questions accurately and abstaining when necessary.

Results

Car Review Classification (Sentiment Analysis):
The sentiment analysis task achieved an accuracy of 80% and an F1 score of 84%, demonstrating strong performance in classifying customer reviews as positive or negative.
English to Spanish Translation:
For the translation task, the BLEU score for the first review (translated from English to Spanish) was 0.49, indicating a good level of translation quality [25].
Question Answering and Summarization:
For the question answering task, the model was able to successfully answer a question based on the review context, demonstrating its ability to provide relevant answers. In the summarization task, the model effectively generated a concise summary of a review, showcasing its capability to extract key information from longer texts.

Limitations

The translation model currently supports only English to Spanish, which limits its usefulness for users who do not speak Spanish. Additionally, the model's maximum input lengths are restricted, with a 512 token limit for sentiment analysis and a 30 token limit for translation, which may impact the handling of longer texts.

Conclusion and Future Work

In this work, we developed a comprehensive framework for analyzing car reviews that encompasses several key tasks: sentiment classification, English-to-Spanish translation, question answering, and text summarization. Our sentiment analysis model achieved an accuracy of 80% and an F1 score of 84%, demonstrating the system's effectiveness in understanding user opinions about cars. The translation model, evaluated using the BLEU score, showed promising results with a score of 0.49 for the first review. The question answering and summarization tasks also proved to be successful, with examples provided to showcase the model's ability to answer specific queries and generate concise summaries from reviews.
In future work, the translation model can be expanded to support additional languages, enhancing the tool’s accessibility for a wider user base. Additionally, we aim to optimize the model’s computational efficiency by increasing the maximum text length, currently limited to 512 tokens for sentiment analysis and 30 tokens for translation. This enhancement would allow the system to handle larger inputs more effectively, especially for long reviews. As the system currently runs on Google Colab with limited storage, leveraging more robust computational resources would further improve performance and scalability. Furthermore, incorporating evaluation metrics for question answering and summarization tasks on the entire dataset would provide a more comprehensive assessment of model performance, ensuring its robustness across various types of reviews.

References

[1] Raeini, Mohammad. "A Survey of Large Language Models: Applications, Challenges, and the Future." Challenges, and the Future (September 08, 2024) (2024).

[2] Sereshki, S., & Lonardi, S. (2024). Predicting differentially methylated cytosines in TET and DNMT3 knockout mutants via a large language model. bioRxiv. https://doi.org/10.1101/2024.05.02.592257

[3] Bashiri, H., & Naderi, H. (2024). Comprehensive review and comparative analysis of transformer models in sentiment analysis. Knowledge and Information Systems, 66(12), 7305-7361. https://doi.org/10.1007/s10115-024-02214-3

[4] Sholehrasa, Hossein. "Integrating Protein Sequence and Expression Level to Analysis Molecular Characterization of Breast Cancer Subtypes." arXiv preprint arXiv:2410.01755 (2024).

[5] Medeiros, T.; Medeiros, M.; Azevedo, M.; Silva, M.; Silva, I.; Costa, D.G. Analysis of Language-Model-Powered Chatbots for Query Resolution in PDF-Based Automotive Manuals. Vehicles 2023, 5, 1384-1399. https://doi.org/10.3390/vehicles5040076

[6] Jahin, Md A., Shovon, Md S. H., Mridha, M. F., Islam, Md R., & Watanobe, Y. (2024). A hybrid transformer and attention-based recurrent neural network for robust and interpretable sentiment analysis of tweets. Scientific Reports, 14(1), 24882. https://doi.org/10.1038/s41598-024-76079-5

[7] Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024). Sentiment analysis in the era of large language models: A reality check. In K. Duh, H. Gomez, & S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 3881-3906). Mexico City, Mexico: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.246

[8] Coleman, J., Krishnamachari, B., Rosales, R., & Iskarous, K. (2024). LLM-assisted rule-based machine translation for low/no-resource languages. In M. Mager, A. Ebrahimi, S. Rijhwani, A. Oncevay, L. Chiruzzo, R. Pugh, & K. von der Wense (Eds.), Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024) (pp. 67-87). Mexico City, Mexico: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.americasnlp-1.9

[9] Kamalloo, E., Dziri, N., Clarke, C., & Rafiei, D. (2023). Evaluating open-domain question answering in the era of large language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5591-5606). Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.307

[10] Laban, P., Kryscinski, W., Agarwal, D., Fabbri, A., Xiong, C., Joty, S., & Wu, C.-S. (2023). SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 9662-9676). Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.600

[11] Baek, Seungyun, Lee, Seunghwan, & Seok, Junhee (2024). Strategic Insights in Korean-English Translation: Cost, Latency, and Quality Assessed through Large Language Model. In 2024 Fifteenth International Conference on Ubiquitous and Future Networks (ICUFN) (pp. 551-553). https://doi.org/10.1109/ICUFN61752.2024.10625181

[12] Koshti, Dipali, Gupta, Ashutosh, & Kalla, Mukesh (2023). Knowledge Blended Open Domain Visual Question Answering using Transformer. In 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 823-828). https://doi.org/10.1109/ICAIS56108.2023.10073911

[13] Patil, Sanika, Nandvikar, Shraddha, Pardeshi, Aakash, & Kurhade, Prof. Swapnali (2024). Automatic Devanagari Text Summarization for Youtube Videos. In 2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP) (pp. 16-21). https://doi.org/10.1109/INNOCOMP63224.2024.00014

[14] "Analyzing Car Reviews with LLMs," DataCamp, [Online]. Available: https://www.datacamp.com. [Accessed: Dec. 10, 2024].

[15] Sanh, V. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[16] HF Canonical Model Maintainers, "distilbert-base-uncased-finetuned-sst-2-english (Revision bfdd146)," Hugging Face, 2022. [Online]. Available: https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english. doi: 10.57967/hf/0181.

[17] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, Oct. 2013, pp. 1631–1642. Association for Computational Linguistics. [Online]. Available: https://aclanthology.org/D13-1170

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, Jun. 2019, pp.

4171–4186. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423

[19] A. Singh, "Edmunds Consumer Car Ratings and Reviews," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/ankkur13/edmundsconsumer-car-ratings-and-reviews/data. [Accessed: Nov. 19, 2024].

[20] A. Fan, S. Bhosale, H. Schwenk, et al., “Beyond English-Centric Multilingual Machine Translation,” arXiv preprint arXiv:2010.11125, 2020. Available: https://arxiv.org/abs/2010.11125.

[21] Meta AI, “facebook/m2m100_418M and facebook/m2m100_1.2B,” Hugging Face, [Online]. Available: https://huggingface.co/facebook/m2m100_418M. [Accessed: Dec. 4, 2024].

[22] W. Wang, F. Wei, L. Dong, et al., “MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers,” arXiv preprint, arXiv:2002.10957, 2020. Available: https://arxiv.org/abs/2002.10957.

[23] P. Rajpurkar, J. Jia, R. Liang, and P. Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD,” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Melbourne, Australia, 2018, pp. 784–789. Available: https://arxiv.org/abs/1806.03822.
4o

[24] M. Lewis, L. Liu, V. Goyal, et al., “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020, pp. 7871–7880. Available: https://arxiv.org/abs/1910.13461.

[25] E. Reiter, "A Structured Review of the Validity of BLEU," Computational Linguistics, vol. 44, no. 3, pp. 393–401, 2018, doi: 10.1162/coli_a_00322.