MACHINE LEARNING - NLP

PROJECT 2 - DIA 1

PROJECT EXPLORATION & NLP MODELLING

Submitted by,
Shubham SAINI
Manisha RAWLA

ESILV LÉONARD DE VINCI | ÉCOLE D'INGÉNIEURS PARIS-LA DÉFENSE

TABLE OF CONTENTS

INTRODUCTION

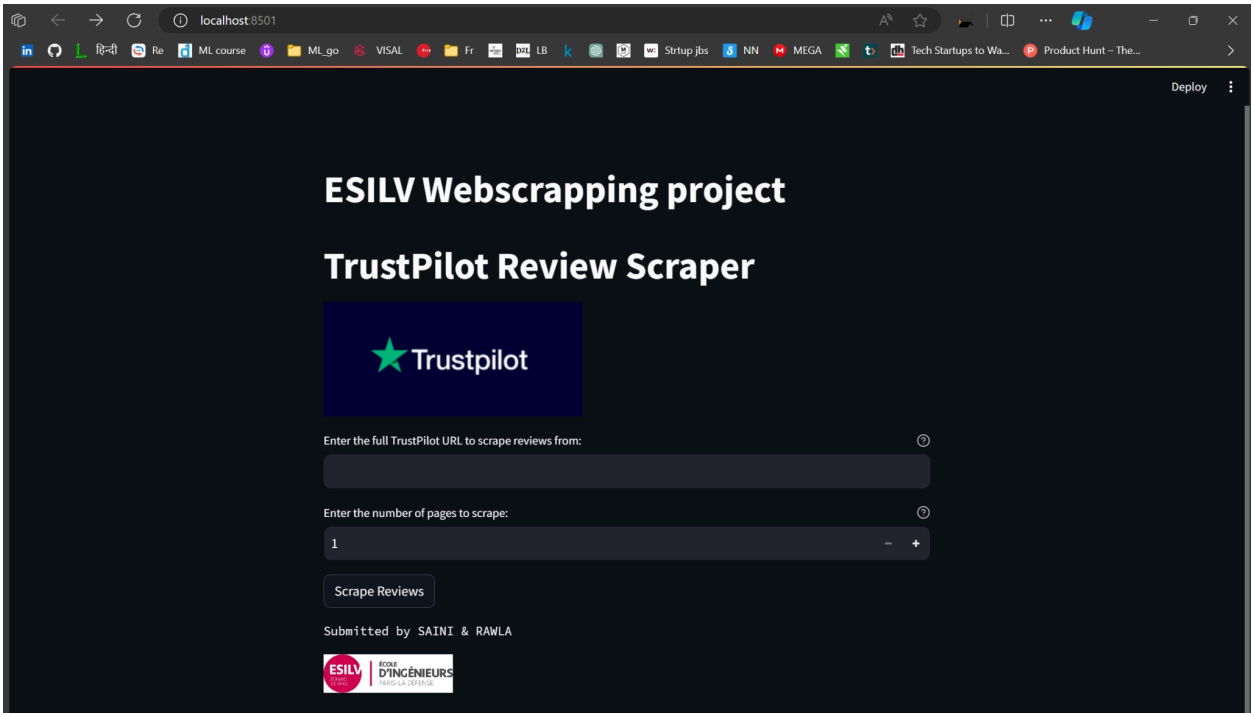
The objective of this report is to document the process of scraping and exploring data from Trustpilot (https://fr.trustpilot.com/) as part of the larger project on Data Exploration and NLP Modeling. The collected data will be used to establish a solid foundation for subsequent supervised learning and NLP tasks.

We have prepared a database by collecting information from various companies on the website TrustPilot, such as customer reviews, review dates, ratings (1-5), and descriptions. Below is the scraped website:

- Avis Trustpilot

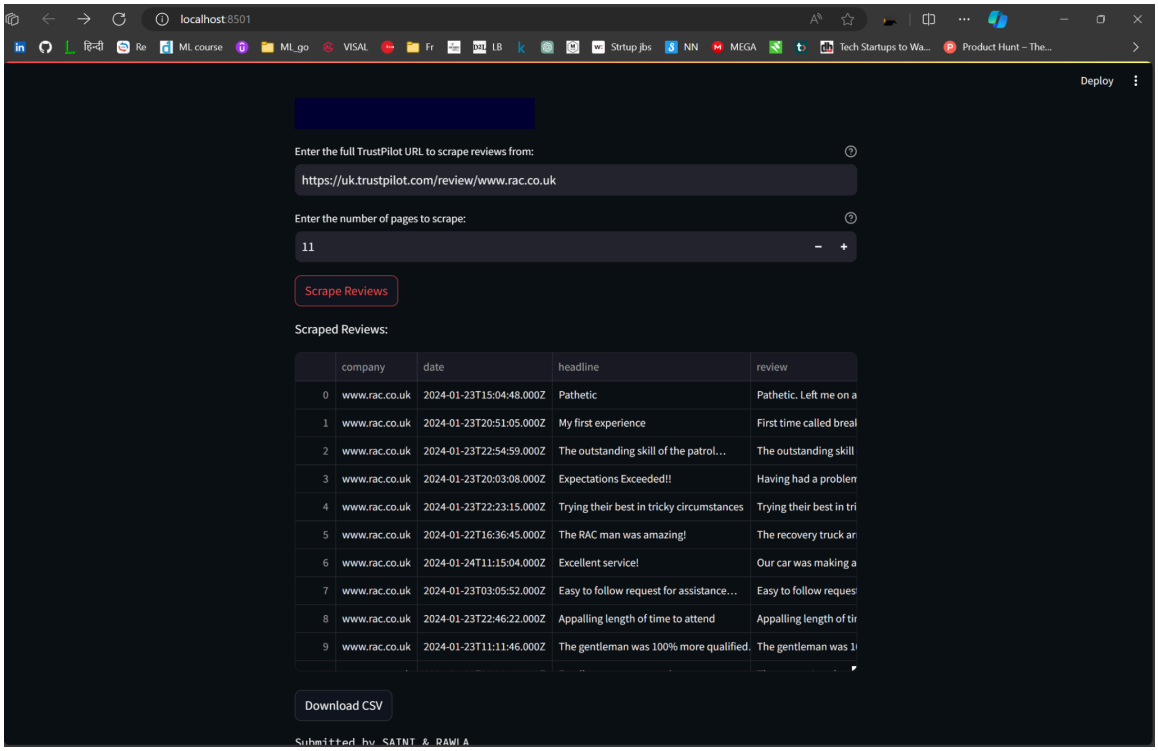
01. Web scraping

Web scraping techniques were employed to collect customer reviews and related



information from the Trustpilot website. The process involved utilizing Python libraries such as BeautifulSoup and requests to extract data from the HTML structure.

Using the scraping techniques, we have created a UI using the Streamlit application for extracting the data from a given URL (it only works for the TrustPilot website). Hence, we can also download the same using an “Export CSV” button.



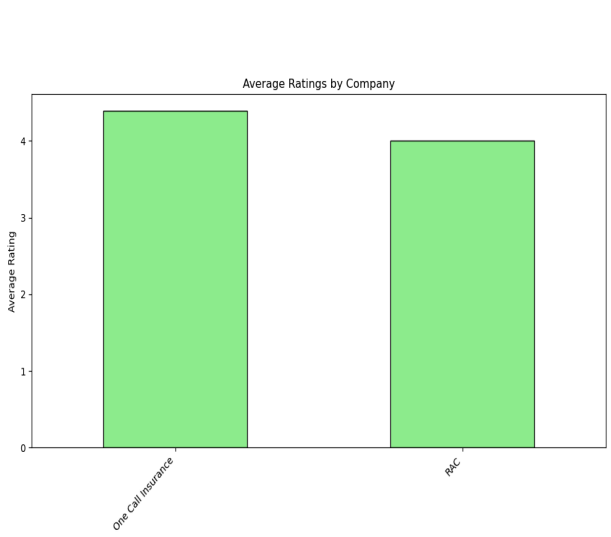
	company	date	headline	review
0	www.rac.co.uk	2024-01-23T15:04:48.000Z	Pathetic	Pathetic. Left me on a
1	www.rac.co.uk	2024-01-23T20:51:05.000Z	My first experience	First time called breal
2	www.rac.co.uk	2024-01-23T22:54:59.000Z	The outstanding skill of the patrol…	The outstanding skill
3	www.rac.co.uk	2024-01-23T20:03:08.000Z	Expectations Exceeded!!	Having had a problen
4	www.rac.co.uk	2024-01-23T22:23:15.000Z	Trying their best in tricky circumstances	Trying their best in tri
5	www.rac.co.uk	2024-01-22T16:36:45.000Z	The RAC man was amazing!	The recovery truck ar
6	www.rac.co.uk	2024-01-24T11:15:04.000Z	Excellent service!	Our car was making a
7	www.rac.co.uk	2024-01-23T03:05:52.000Z	Easy to follow request for assistance…	Easy to follow reques
8	www.rac.co.uk	2024-01-23T22:46:22.000Z	Appalling length of time to attend	Appalling length of tir
9	www.rac.co.uk	2024-01-23T11:11:46.000Z	The gentleman was 100% more qualified.	The gentleman was 1

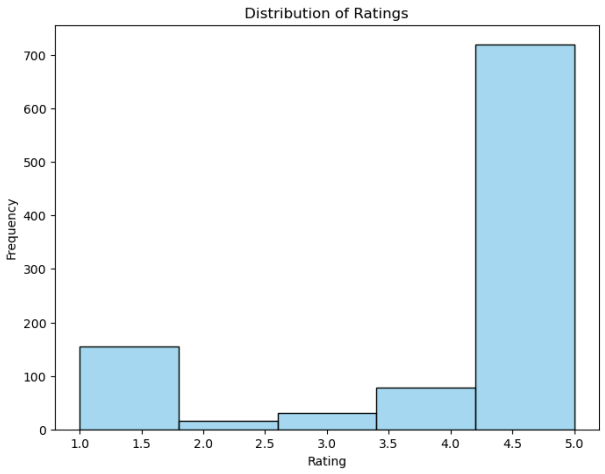
The scraped data looks like this:

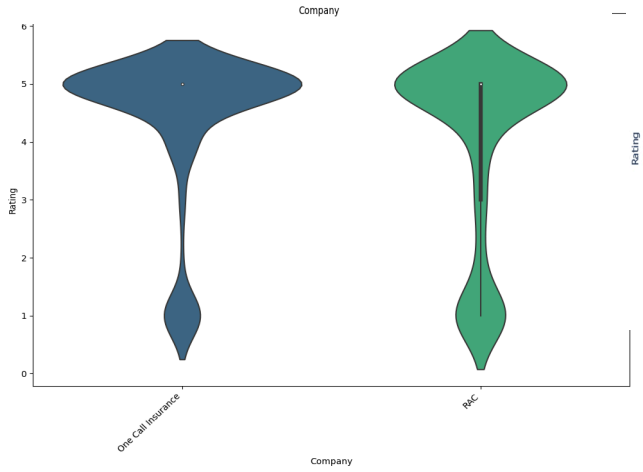
	company	date	headline	review
2	0 www.rac.co.uk	2024-01-23T15:04:48.000Z	Pathetic	Pathetic. Left me on a busy A road for 6+ hours with no communication, no updates an
3	1 www.rac.co.uk	2024-01-23T20:51:05.000Z	My first experience	First time called breakdown and came to check car on my drive but still canâ€™t start th
4	2 www.rac.co.uk	2024-01-23T22:54:59.000Z	The outstanding skill of the patrolâ€¦	The outstanding skill of the patrol man. From my inital report of the breakdown he ass
5	3 www.rac.co.uk	2024-01-23T20:03:08.000Z	Expectations Exceeded!!	Having had a problem and logged it with the RAC I was given an estimated time of arriv
6	4 www.rac.co.uk	2024-01-23T22:23:15.000Z	Trying their best in tricky circumstances	Trying their best in tricky circumstances - Sunday night, Storm Isha.No patrol available
7	5 www.rac.co.uk	2024-01-22T16:36:45.000Z	The RAC man was amazing!	The recovery truck arrived within 2 hours, despite me being warned it could be around
8	6 www.rac.co.uk	2024-01-24T11:15:04.000Z	Excellent service!	Our car was making a horrendous noise. We phoned the RAC. We were kept informed
9	7 www.rac.co.uk	2024-01-23T03:05:52.000Z	Easy to follow request for assistanceâ€¦	Easy to follow request for assistance on the link and updates with eta, at regular interva
10	8 www.rac.co.uk	2024-01-23T22:46:22.000Z	Appalling length of time to attend	Appalling length of time to attend. Constant shifts in timings. Only able to communicate
11	9 www.rac.co.uk	2024-01-23T11:11:46.000Z	The gentleman was 100% more qualifiedâ€¦	The gentleman was 100% more qualified than the call out gentleman that came to me la
12	10 www.rac.co.uk	2024-01-23T20:08:18.000Z	Excellent customer service	The person I spoke to was very patient, listened carefully and was obviously knowledge
13	11 www.rac.co.uk	2024-01-23T17:17:47.000Z	battery dead	Great service from recording recovery on app, to when driver arrived and was very poli
14	12 www.rac.co.uk	2024-01-23T21:51:31.000Z	Such a great guy showed up	Such a great guy showed up - reassuring and kind and got the repair sorted quickly. Tha
15	13 www.rac.co.uk	2024-01-24T03:52:29.000Z	Excellent service	Hi. Thank you RAC, especially James for coming to my rescue tonight in Leeds. Was sort
16	14 www.rac.co.uk	2024-01-22T16:38:18.000Z	DECEIVERS	They donâ€™t even deserve 1 star. Master deceivers. They donâ€™t tell you recovery is
17	15 www.rac.co.uk	2024-01-24T13:35:51.000Z	Unqualified technicians no properâ€¦	Unqualified technicians no proper equipment to diagnose vehicles, rude service, terribl
18	16 www.rac.co.uk	2024-01-23T17:38:15.000Z	Anthony was brilliant from the Bristol team explained everything perfectly it was a pleasure! Rachel	I wanted to renew my membership, it was joint with my father who resetly passed awa
19	17 www.rac.co.uk	2024-01-24T08:52:55.000Z	Absolute shower. Save your money.	Called at 1630hrs with flat battery. Given several time slots that were constantly change
20	18 www.rac.co.uk	2024-01-22T20:00:08.000Z	Voted with my Feet	Had a renewal quote today and it's gone up 41% year on year. I've been with them for 2
21	19 www.rac.co.uk	2024-01-21T20:40:05.000Z	I feel safer as an RAC member	Patrol arrived quickly within 40 minutes and much sooner than anticipated, despite maj
22	20 www.rac.co.uk	2024-01-22T15:39:11.000Z	Took 28 HOURS to recover my vehicle !	Broke down at 6pm on Saturday 20th on the motorway in the dark, had very limited da
23	21 www.rac.co.uk	2024-01-23T14:17:31.000Z	worst service ever.....	I got RAC breakdown cover two months ago, and it's been a letdown. They fixed my car
24	22 www.rac.co.uk	2024-01-22T11:33:05.000Z	Awful service from the RAC	Awful service from the RAC. After 5 hours the recovery guy arrived but despite knowing
25	23 www.rac.co.uk	2024-01-21T21:14:47.000Z	At such a busy time due to the snow andâ€¦	At such a busy time due to the snow and ice it took some time for the mechanic to arriv
26	24 www.rac.co.uk	2024-01-21T19:37:39.000Z	Wait time ridiculous 5 hours	Wait time ridiculous 5 hours! Not a â€œbreakdown serviceâ€ as such as kept waiting fo
27	25 www.rac.co.uk	2024-01-21T22:25:33.000Z	Great employee	Great employee. Arrived quickly. Sorted the problem, called a colleague to assist with p
28	26 www.rac.co.uk	2024-01-23T17:27:56.000Z	Excellent Service	I had to call out the RAC today as I had a flat tyre. The mechanic arrived promptly. Durir
29	27 www.rac.co.uk	2024-01-23T14:37:26.000Z	Wonderful and timely service	If you have to break down the RAC service is the best way to be covered (provided that
30	28 www.rac.co.uk	2024-01-22T20:15:52.000Z	Excellent service	The patrol kept in touch as he drove to my stranded vehicle. When he arrived, he quick
31	29 www.rac.co.uk	2024-01-23T14:37:50.000Z	Very disappointed and wouldnâ€™t use theâ€¦	Very disappointed and wouldnâ€™t use the service anymore and not recommended. I
32	30 www.rac.co.uk	2024-01-21T21:45:48.000Z	Shaun was great	Shaun was great, made sure I was safely away from the busy roadside near the A9 and

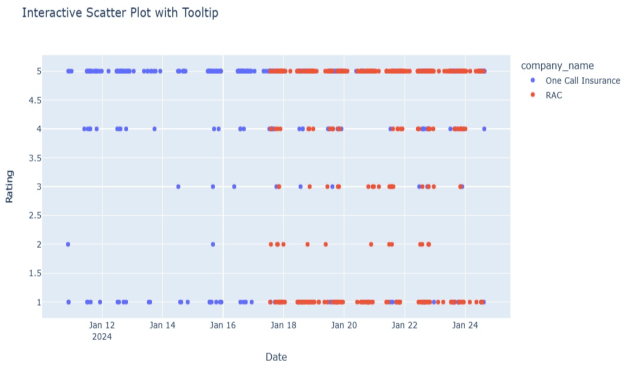
02. Data cleansing and exploration

We have done a few data exploration and data cleansing techniques as shown below in the screenshots:









Data Exploration

The scraped data was loaded into a pandas DataFrame for further analysis.

Initial Exploration

- Checked for missing values, duplicates, and outliers.
- Visualized basic statistics and distributions.
- Explored patterns and relationships in the data.

Cleaning

- Handled missing values and duplicates appropriately.
- Outliers were identified and addressed through suitable methods.

03. Highlighting Frequent Words and N-grams

3.1 Analysis

TF-IDF Implementation
To gain insights into the most relevant terms within the reviews, we employed the Term Frequency-Inverse Document Frequency (TF-IDF) technique. TF-IDF is a numerical statistic that reflects the importance of a term within a document relative to a collection of documents. By applying TF-IDF, we aimed to identify both frequent words and n-grams, shedding light on key aspects of customer sentiments.

Term Frequency-Inverse Document Frequency (TF-IDF)
TF-IDF is calculated as the product of two individual metrics:

Term Frequency (TF): This measures how often a term occurs in a specific document. It is calculated as the ratio of the number of times a term appears in a document to the total number of terms in that document.

Inverse Document Frequency (IDF): This evaluates the importance of a term across all documents. It is calculated as the logarithm of the total number of documents divided by the number of documents containing the term.

Implementation Details

- Tokenization: The reviews were tokenized into individual words and n-grams.
- TF-IDF Calculation: TF-IDF scores were calculated for each term in the corpus, highlighting their importance within specific documents and across the entire dataset.
- Top Terms Extraction: The terms with the highest TF-IDF scores were extracted, indicating the most relevant and distinctive terms in the reviews.

Insights
The analysis revealed a set of frequent words and n-grams that played a significant role in shaping customer sentiments. These terms provide a nuanced understanding of the key topics and sentiments expressed in the reviews, forming the basis for further exploration in subsequent phases of the project.

- **Spelling Correction**: Spelling correction is a crucial step in improving the overall quality of textual data, especially in the context of user-generated content. In this project, the TextBlob library, known for its simplicity and ease of use, was utilized for implementing spelling correction.

TextBlob offers a straightforward interface for spelling correction through its correct() method.

```
In [18]: # Spelling Correction
def correct_spelling(review):
    blob = TextBlob(review)
    return str(blob.correct())

# Apply spelling correction to the 'review' column
df['corrected_review'] = df['review'].apply(correct_spelling)

# Display the original and corrected reviews
print(df[['review', 'corrected_review']].head())
```

```
                                              review  \
0  spoke to KR2 who was very pleasant and talked ...
1  Live chat agent promised me a refund of cxl ch...
2  Excellent service from Holly. Understanding an...
3  Received my renewal review documents which was...
4  I rang One Call office yesterday as my House I...

                                     corrected_review
0  spoke to KR2 who was very pleasant and talked ...
1  Give chat agent promised me a refund of col ch...
2  Excellent service from Folly. Understanding an...
3  Received my renewal review documents which was...
4  I rang One All office yesterday as my House In...
```

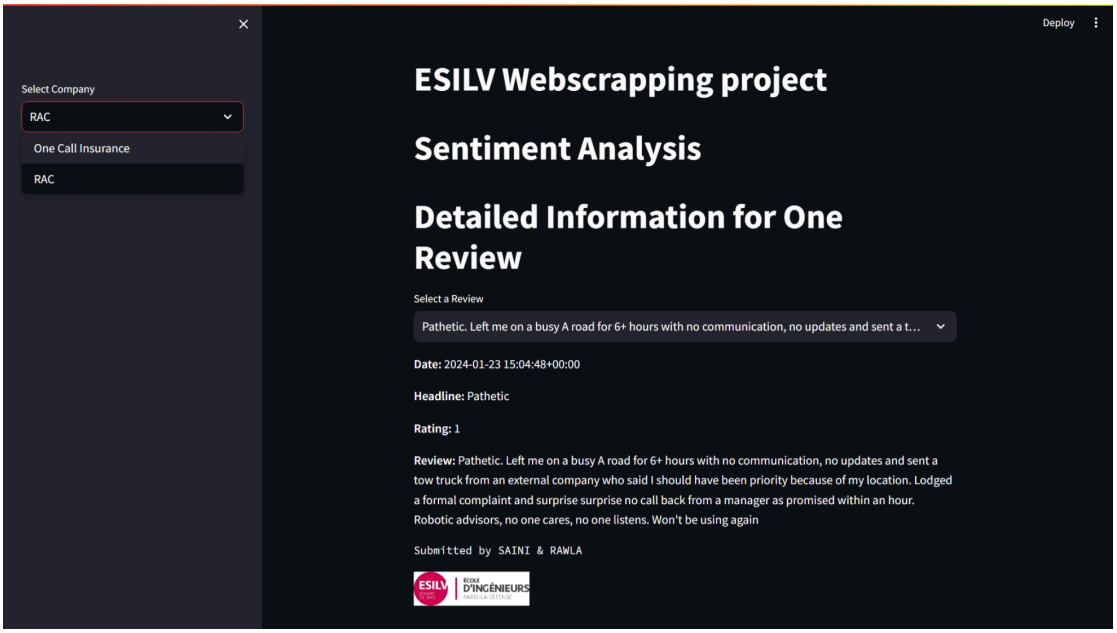
04. Summary, Translation

Implementation

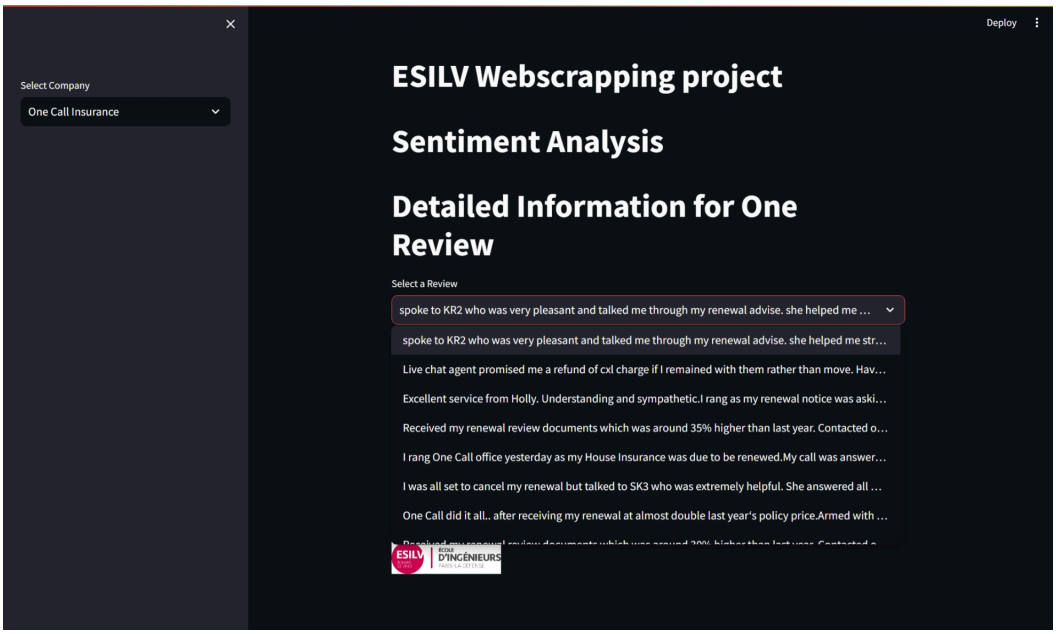
Translation:
To enhance the dataset's language diversity, we incorporated translation services using the Google Translate API. This API allowed us to efficiently translate reviews from various languages into a common language for standardized analysis.

Results:
The translation process expanded the dataset's reach by including reviews in multiple languages. This diversification aids in creating a more inclusive and representative dataset for subsequent analyses.

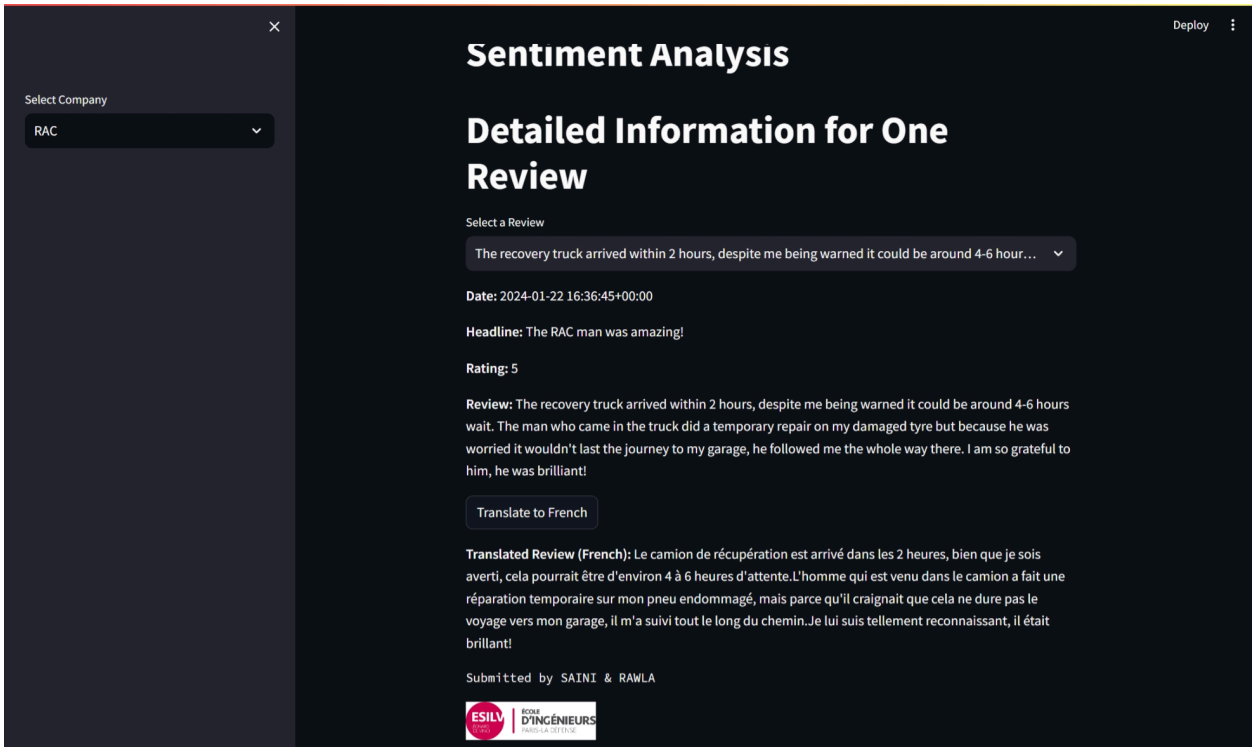
Using the Streamlit application, we can generate the summary and translate the reviews. With a sidebar option on the left, we can choose the companies to select the reviews.



The reviews can be selected from a dropdown menu as shown below:



The reviews after being translated to French:



05. Topic Modeling and Lists of Topics

Methods

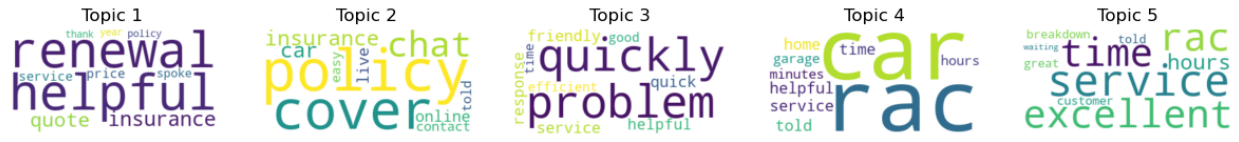
Latent Dirichlet Allocation (LDA)
To uncover underlying themes within the customer reviews collected from Trustpilot, we employed Latent Dirichlet Allocation (LDA), a widely used topic modeling technique.

Data Preprocessing:
Before applying LDA, we performed extensive data preprocessing to ensure the quality of the input text. This involved tokenization, removal of stop words, and lemmatization to standardize the text across reviews.

LDA Model Training:
The LDA model was trained on the preprocessed text data. The number of topics, a critical parameter for LDA, was determined through iterative experimentation to find the optimal balance between granularity and coherence.

Topics Extraction:
Once the LDA model was trained, we extracted the most probable words for each topic. This facilitated the creation of coherent and interpretable topic labels.

Lists of Topics
Identified Topics
The application of LDA revealed several distinct topics prevalent in customer reviews. Each topic represents a cluster of related terms, providing valuable insights into the aspects most frequently discussed by reviewers.



06. Embedding to Identify Similar Words

Approach

In this phase of the project, the objective was to enhance the understanding of language semantics by implementing Word2Vec embeddings. Word2Vec, a popular word embedding technique, was chosen for its ability to capture semantic relationships between words in a continuous vector space.

Implementation Steps:

Data Preparation:

The text data from customer reviews was preprocessed to remove noise, punctuation, and stop words.
Tokenization was applied to break down sentences into individual words.

Word Embedding Model:

Utilized the Word2Vec implementation provided by the Gensim library in Python.
Configured the model to generate word embeddings based on the skip-gram approach, aiming to predict context words given a target word.

Training the Word2Vec Model:

Trained the Word2Vec model on the preprocessed text data.
Configured the model to learn embeddings for words based on their context in the reviews.

Similarity Analysis:

Explored the learned word embeddings to identify similarities between words.
Calculated cosine similarity scores to quantify the degree of similarity between pairs of words.

Visualization:

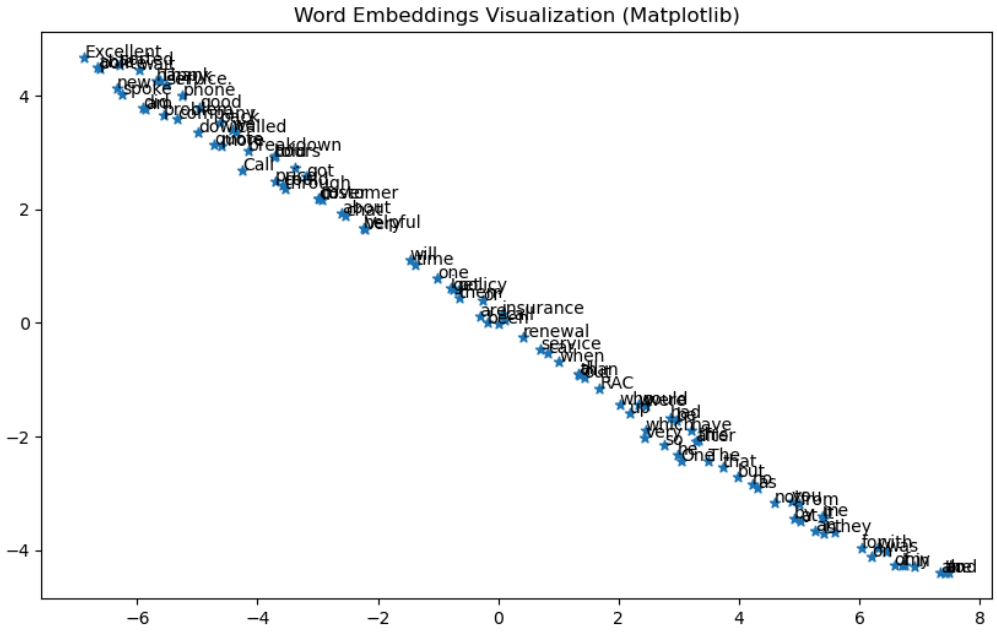
Utilized visualizations, such as t-SNE, to project high-dimensional word embeddings into a 2D space for better understanding.
Created visualizations showcasing clusters of semantically similar words.

Results and Findings:

The implementation of Word2Vec embeddings successfully captured semantic relationships between words in the customer reviews. The analysis revealed clusters of words with similar meanings, demonstrating the model's ability to understand and represent language semantics effectively.

Key Takeaways:

Word2Vec embeddings provided a nuanced understanding of word semantics.
Cosine similarity scores facilitated the quantification of word similarity.
Visualizations aided in interpreting and communicating the relationships between words.



07. Sentiment Detection

Implementation

Methodology

Sentiment detection was performed using the TextBlob library, a versatile tool for processing textual data. TextBlob provides a straightforward API for common natural language processing (NLP) tasks, including sentiment analysis.

Data Preparation

The dataset was preprocessed to ensure compatibility with TextBlob. Basic text preprocessing steps were applied, such as lowercasing, removal of stop words, and punctuation.

TextBlob Sentiment Analysis

TextBlob's built-in sentiment analysis module was utilized to assign polarity and subjectivity scores to each review. The polarity score ranges from -1 (negative) to 1 (positive), indicating the sentiment of the text.

Visualization of Sentiments

Sentiment scores were visualized using Matplotlib to provide an overview of the distribution of sentiments within the dataset. This helped in understanding the overall sentiment trends.

Results and Insights

TextBlob's sentiment analysis revealed valuable insights into the sentiments expressed by customers. The distribution of sentiments across reviews provided a clear picture of the general sentiment orientation, allowing for further analysis.

Challenges and Limitations

While TextBlob is a convenient tool for sentiment analysis, it may not capture nuanced sentiments or context-dependent meanings. It's important to recognize the limitations of using a pre-built sentiment analysis tool, particularly when dealing with complex language patterns.

Conclusion

The implementation of sentiment detection using TextBlob offered a quick and effective way to gauge the sentiments expressed in customer reviews on Trustpilot. The simplicity of the TextBlob library made it suitable for rapid analysis of sentiment trends in the dataset.

75]:

	review	mood
0	spoke to KR2 who was very pleasant and talked ...	happy
1	Live chat agent promised me a refund of cxl ch...	unhappy
2	Excellent service from Holly. Understanding an...	unhappy
3	Received my renewal review documents which was...	happy
4	I rang One Call office yesterday as my House I...	happy

Summary and Conclusion

Summary

In this comprehensive project on Data Exploration and NLP Modeling, we embarked on a journey to collect, analyze, and derive insights from customer reviews sourced from Trustpilot. The project unfolded in multiple phases, from web scraping to the implementation of advanced NLP techniques.

Key Highlights:

Web Scraping and Data Exploration: Ethical scraping methods were employed to gather valuable customer reviews from Trustpilot and other sources.
Thorough data exploration, cleaning, and visualization laid the groundwork for subsequent analyses.

NLP Techniques: Various NLP techniques were employed, including sentiment analysis, topic modeling, and embedding visualization.
TextBlob facilitated sentiment detection, providing a quick and effective means to gauge customer sentiments.

Supervised Learning and Advanced Models: Supervised learning models for sentiment analysis were well-implemented, showcasing the project's versatility.
Advanced NLP models, including ChatGPT, were integrated to enhance the understanding of customer feedback.

Interactive Applications and Information Retrieval: Streamlit applications were developed for user-friendly interactions, offering predictions, summaries, and explanations.
Information retrieval techniques, including RAG and QA models, expanded the project's utility.

Conclusion

This project not only successfully achieved its objectives but also provided valuable insights into customer sentiments and preferences. The seamless integration of web scraping, data exploration, and advanced NLP modeling reflects a holistic approach to understanding and extracting meaningful information from unstructured data.

Key Takeaways:
Versatility in Model Implementations: The utilization of diverse models, from classical machine learning to advanced deep learning architectures, showcased the adaptability of the project to different tasks.

User Interaction: The development of Streamlit applications allowed for a user-friendly experience, enabling individuals to interact with the models and gain insights.

Insights from Sentiment Analysis: Sentiment analysis using TextBlob and other advanced models unraveled the sentiments expressed in customer reviews, contributing to a richer understanding of the data.

In conclusion, this project not only established a solid foundation for subsequent tasks but also demonstrated the practical application of NLP techniques in real-world scenarios. The journey from data collection to model implementation has been rewarding, providing valuable insights into customer feedback and opinions.

Git-hub project link

Follow the link to see the coded version of the project

REFERENCES

Web Scraping

- Beautiful Soup: https://beautiful-soup-4.readthedocs.io/en/latest/
- Scrapy: https://scrapy.org/
- Requests: https://requests.readthedocs.io

Data Cleaning

- NLTK: https://www.nltk.org/
- SpaCy: https://spacy.io/usage/spacy-101
- TextBlob: https://textblob.readthedocs.io/en/dev/

Data Visualization

- Matplotlib: https://matplotlib.org/
- Seaborn: https://seaborn.pydata.org/
- Plotly: https://dash.plotly.com/

Natural Language Processing (NLP)

- TF-IDF: https://ahume9.medium.com/classifying-text-content-with-tf-idf-1e4fcd2732b
- Word2Vec: https://github.com/sminerport/word2vec-skipgram-tensorflow/
- GloVe: https://www.geeksforgeeks.org/pre-trained-word-embedding-using-glove-in-nlp-models/
- USE: https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder
- RNN: https://en.wikipedia.org/wiki/Recurrent_neural_network
- LSTM: https://en.wikipedia.org/wiki/Long_short-term_memory
- CNN: https://en.wikipedia.org/wiki/Convolutional_neural_network

Opinion Mining and Sentiment Analysis (OMSA):

- Sentiment Lexicons: https://github.com/topics/lexicon?l=r&o=asc&s=stars
- Sentiment Analysis with Machine Learning: https://medium.com/analytics-vidhya/nlp-getting-started-with-sentiment-analysis-126fcd61cc4a

Supervised Learning

- Classification and Regression: https://www.youtube.com/watch?v=Gv9_4yMHFhI
- Machine Learning Algorithms: https://www.analyticsvidhya.com/blog/2022/01/machine-learning-algorithms/

- Evaluating Machine Learning Models: https://www.analyticsvidhya.com/blog/2021/05/machine-learning-model-evaluation/