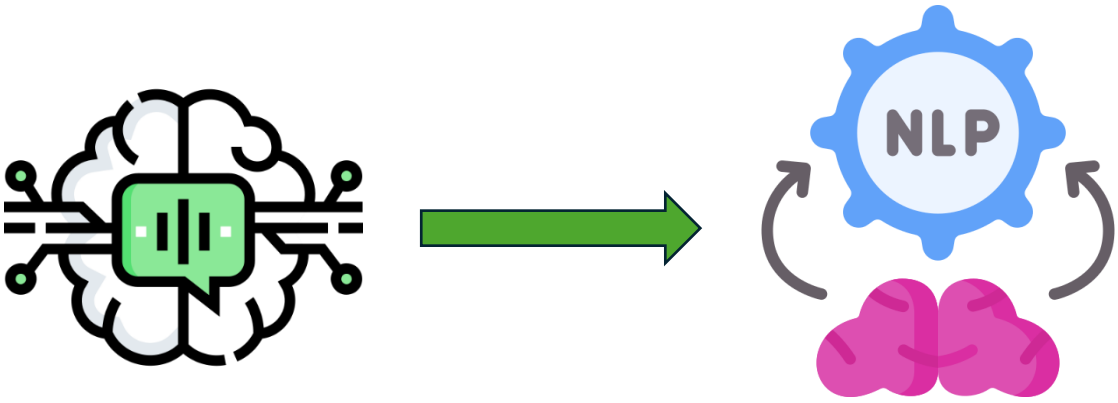
Pattern Recognition Project – Text Sentiment Analysis with Natural Language Processing

Summary



Introduction

Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on enabling machines to understand, process, and respond to human language in a meaningful way.

Language, being inherently complex and unstructured, presents significant challenges for pattern recognition.

For this project, the team selected NLP as a core focus to explore its potential in bridging the gap between human communication and machine understanding.

Two impactful applications of NLP were chosen for this project: sarcasm detection and restaurant review sentiment analysis.

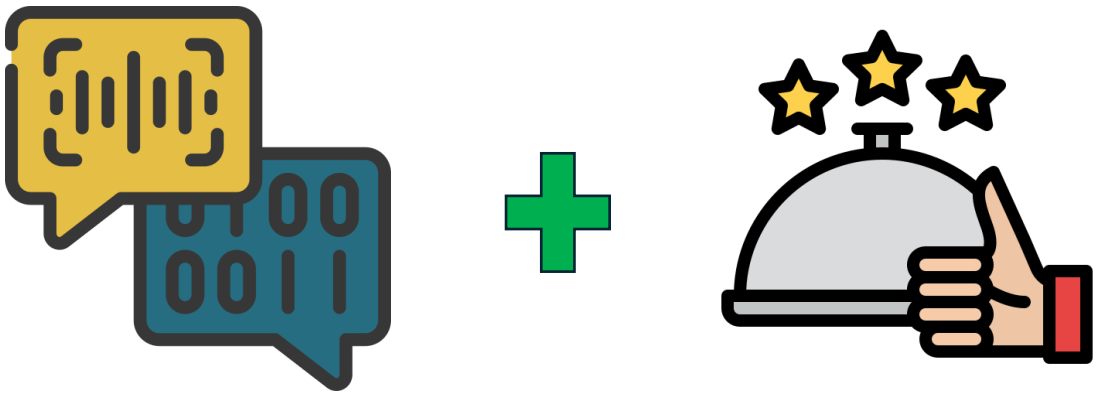
These applications were selected because they demonstrate how NLP can handle subtle linguistic nuances, such as tone and sentiment, and provide meaningful insights for decision-making processes.

The primary goals of this project include:

- **Developing an NLP pipeline for sarcasm detection using the News Headlines Dataset.**
- **Analyzing restaurant reviews using sentiment classification powered by the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model.**

This report outlines the processes undertaken for data collection, preprocessing, model development, evaluation, and validation.

The results demonstrate the effectiveness of supervised learning and deep learning models in tackling pattern recognition problems in NLP.



Overview of Natural Language Processing

NLP is an interdisciplinary field that combines linguistics, computer science, and machine learning to analyze and interpret human language.

Early methods relied on rule-based systems, which required manually defined syntactic and semantic rules.

These approaches, while foundational, struggled with the scalability and variability of real-world language data.

Modern NLP leverages advancements in machine learning, particularly deep learning, to process text and speech more effectively.

Models such as BERT, GPT, and XLNet have revolutionized NLP by using transformers and large-scale datasets to capture contextual information in language.

These models rely on techniques like attention mechanisms to achieve state-of-the-art performance in various NLP tasks, including sentiment analysis, machine translation, and text summarization.

NLP applications are integral to many industries, enabling functionalities like real-time language translation, sentiment prediction, and conversational AI.

Tools such as chatbots, virtual assistants, and recommendation systems benefit from NLP's ability to analyze complex patterns in text and speech.

Data Collection and Preparation

Sarcasm Detection

The **News Headlines Dataset** was used for sarcasm detection. This dataset contains three attributes:

1. **is_sarcastic**: A binary label indicating whether the headline is sarcastic (1) or not (0).
2. **headline**: The text of the news headline.
3. **article_link**: A link to the original dataset source.

To prepare the data, we loaded the JSON file, iterated through its records, and extracted headlines and labels into Python lists.

Training and testing datasets were split using list slicing, ensuring a balanced representation.

Restaurant Review Analysis

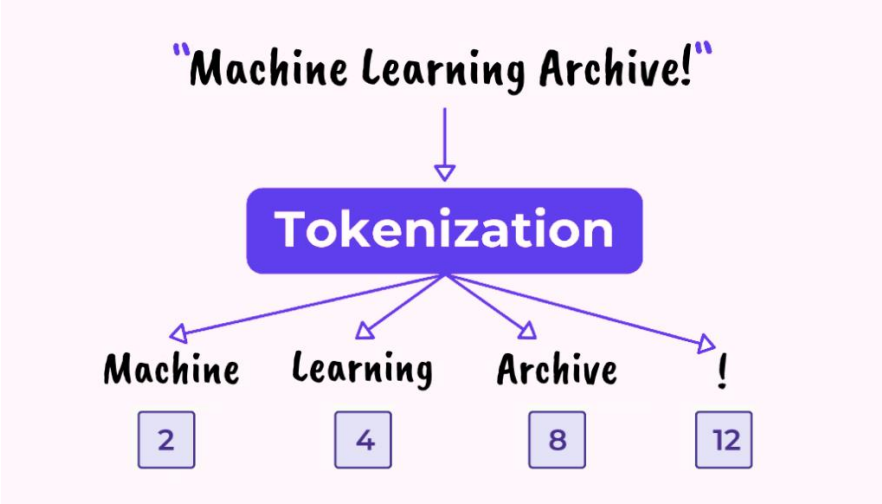
The sentiment analysis task utilized customer reviews from Din Tai Fung restaurant in Kuala Lumpur using Yelp website.

This dataset was collected using the BeautifulSoup library for web scraping and included reviews written in six languages: English, Dutch, German, French, Spanish, and Italian.

Each review was associated with a star rating (1 to 5) representing sentiment polarity, with 1 being negative and 5 being positive.

Preprocessing Steps:

- **Data Cleaning**: Removed HTML tags, URLs, emojis, and non-alphanumeric characters.
- **Tokenization**: Split reviews into individual words for analysis.



Tokenization is the process of splitting text into smaller components, such as words or subwords, to enable computational analysis.

For example, the phrase "Machine Learning Archive!" is broken into tokens: "Machine," "Learning," "Archive," and "!", each assigned a unique identifier (e.g., 2, 4, 8, 12).

These numerical representations allow machine learning models to process text efficiently, bridging the gap between human language and machine-readable data.

Tokenization is a foundational step in natural language processing, making textual data structured and analyzable for various applications.

Model Development

Sarcasm Detection

Sarcasm detection is a challenging NLP task because it requires the model to interpret contextual cues that often contradict the literal meaning of the text.

A supervised learning approach was adopted using the News Headlines Dataset.

The sarcasm detection model used TensorFlow's Tokenizer to preprocess the text by tokenizing sentences and padding sequences to a uniform length.

This step ensured consistent input dimensions for the neural network. The architecture of the neural network included:

1. **An embedding layer that converted tokens into dense vectors of a fixed size.**
2. **A GlobalAveragePooling1D layer that reduced the dimensionality by averaging vector values.**
3. **A Dense layer with 24 units and ReLU activation for feature extraction.**
4. **A final Dense layer with a sigmoid activation function for binary classification.**

The model was compiled using the Adam optimizer and binary cross-entropy loss function.

It was trained over 30 epochs, achieving a training accuracy of 96% and a testing accuracy of 84%.

A decoding function was implemented to reverse the tokenization process.

It maps numeric sequences back to words using the word index, providing interpretable outputs.

For example, the sentence "Former store clerk sues over secret 'black '" was reconstructed from its padded sequence.

We also examined the embedding layer’s weights to understand the dimensional representation of each word.

The embedding matrix, with a shape of (10,000, 16), encodes semantic relationships within the vocabulary.

Restaurant Review Analysis

For sentiment analysis, we used a pre-trained BERT model from Huggingface.

BERT, fine-tuned for sentiment analysis, predicted review sentiment on a scale of 1 to 5. The process included:

Web Scraping: Reviews were extracted using the BeautifulSoup library by targeting relevant HTML elements from Yelp pages.

Text data was parsed, and unnecessary tags were removed using regular expressions.

Sentiment Scoring: The AutoTokenizer and AutoModelForSequenceClassification from Huggingface were used to tokenize the reviews and generate predictions.

Sentiments were encoded on a scale of 1 to 5, with the highest score representing the predicted sentiment.

Storing Results: Extracted reviews and their corresponding sentiment scores were stored in a pandas DataFrame for further analysis.

Each review's sentiment score was computed using a custom function that mapped the tokenized text to the BERT model's output.

Model Evaluation and Validation

Sarcasm Detection

The performance of the sarcasm detection model was analyzed using training and validation metrics:

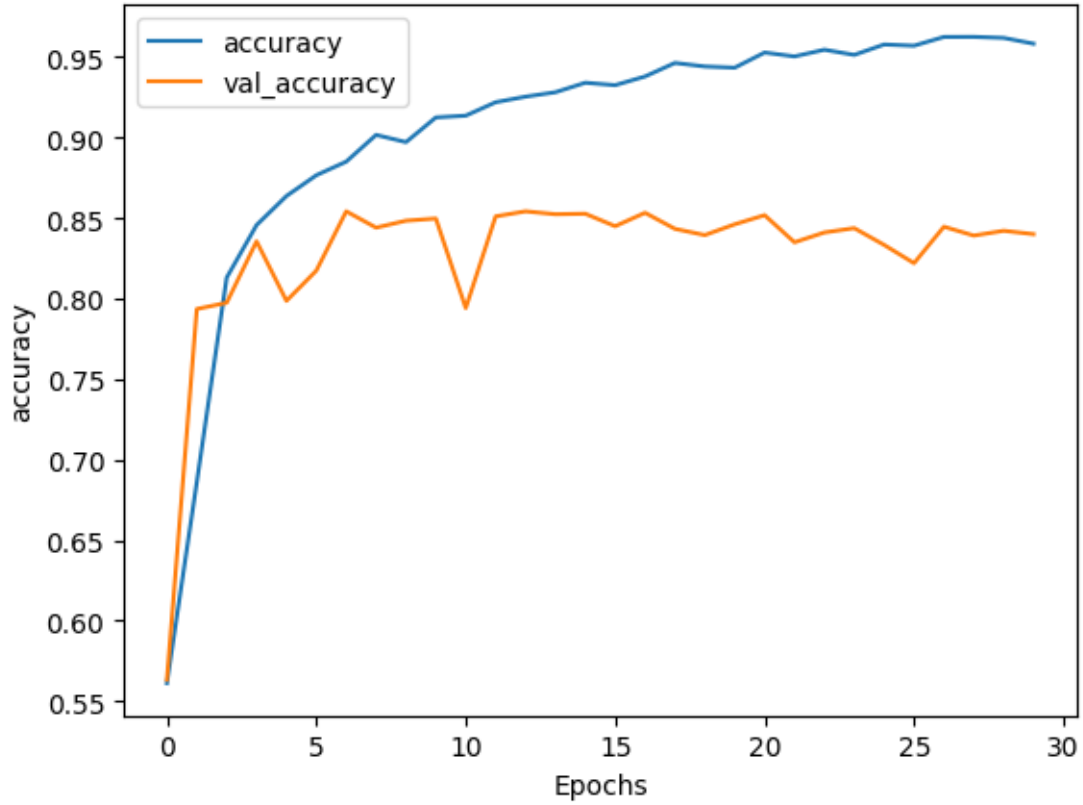
Accuracy Plot: The accuracy plot shows the model’s training and validation accuracy across 30 epochs.

The training accuracy steadily increased, reaching approximately 96% by the final epoch.

Validation accuracy stabilized around 84%, indicating the model’s ability to generalize well on unseen data.

The divergence between training and validation accuracy suggests slight overfitting, which could be mitigated with techniques like dropout or regularization.

Figure : Accuracy Plot regarding Sarcasm Detection

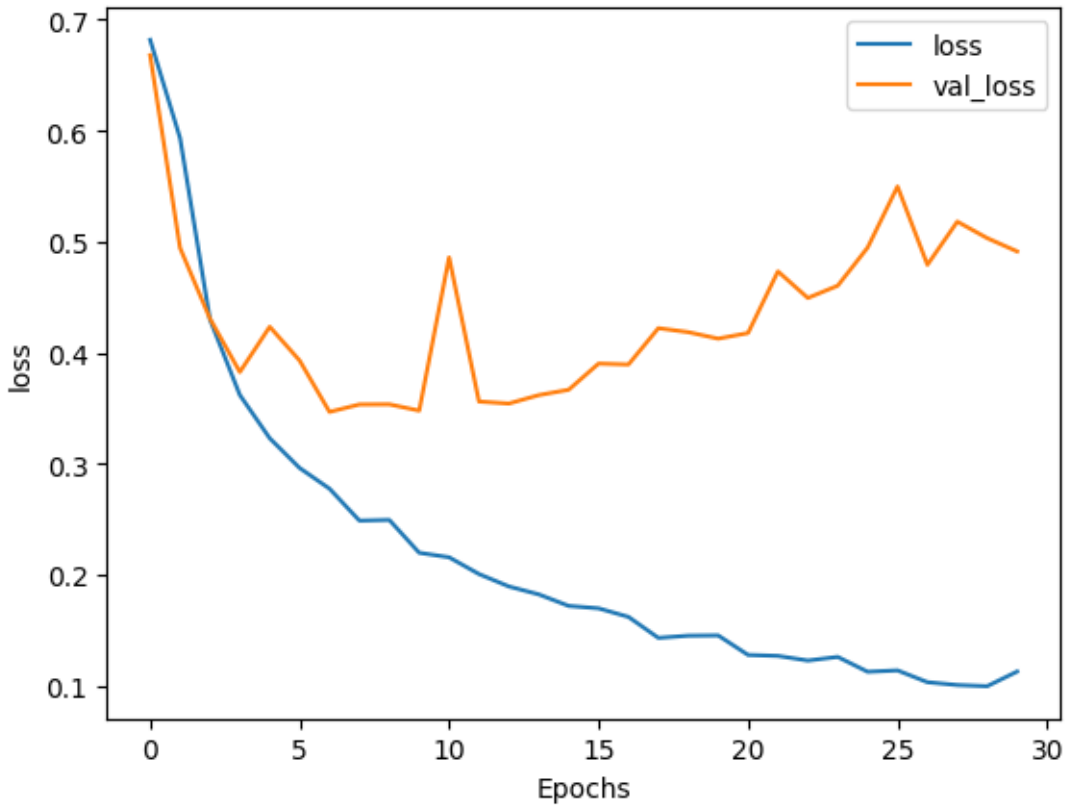


Loss Plot: The loss plot demonstrates the model's reduction in both training and validation loss over epochs.

Training loss consistently decreased, reflecting effective optimization of the model's parameters.

Validation loss, while fluctuating slightly, showed an overall downward trend before stabilizing, indicating that the model learned meaningful patterns from the data without significant overfitting.

Figure : Loss Plot regarding Sarcasm Detection



Sentence Predictions:

The sentence predictions illustrate the practical application of the sarcasm detection model and its ability to discern nuanced meanings within text.

The model's prediction for the sentence, "Sure, because binge-watching TV shows totally counts as exercise," accurately identified sarcasm with a high confidence score of 81%, demonstrating its ability to recognize ironic or contextually contradictory language patterns.

On the other hand, the prediction for "Attack On Titan season finale showing this Sunday night at cinema" as non-sarcastic with a confidence of less than 1% highlights the model's capability to distinguish straightforward, factual statements from sarcasm.

These examples emphasize the model’s strengths in interpreting tone and context, essential for sarcasm detection tasks.

However, further analysis could investigate its performance across more diverse and challenging contexts, such as multi-sentence sarcasm or culturally-specific expressions.

This capability showcases the potential for real-world applications, such as sentiment analysis in social media, where sarcasm often obscures the true intent of statements.

Restaurant Review Analysis

The BERT model achieved a 93% accuracy rate in predicting star ratings.

Reviews were scored between 1 and 5, with predictions reflecting customer sentiment.

The results stored in a DataFrame demonstrated consistent performance in distinguishing positive and negative reviews. For example:

- "This is a Taiwanese restaurant that's now famous around Asia and the US west coast..." received a sentiment score of 5.
- "I have had good experiences at franchises in the Seattle, Washington USA, Taipei 101, Taiwan..." received a sentiment score of 2, reflecting a more critical tone.

This approach highlights the effectiveness of using pre-trained models for multi-language sentiment analysis.

Conclusion and Future Work

NLP has immense potential to transform how machines interpret and interact with human language, as demonstrated by our project's focus on sarcasm detection and restaurant review sentiment analysis.

While we achieved promising results, there are areas for improvement, including enhancing sarcasm detection by incorporating contextual information from additional datasets, expanding sentiment analysis to include real-time review monitoring, and exploring multi-modal NLP applications that combine text, audio, and visual inputs.

As NLP continues to evolve, these advancements will further enrich human-machine interactions, solidifying its role as a vital area for future exploration.

References

1. Rishabh Misra, "News Headlines Dataset for Sarcasm Detection." Available at https://rishabhmisra.github.io/NewsHeadlinesDataset.pdf
2. Huggingface Documentation: BERT Model for Sentiment Analysis.
3. Din Tai Fung Restaurant in Kuala Lumpur reviews from Yelp website: https://www.yelp.com/biz/din-tai-fung-kuala-lumpur-2?osq=Restaurants