NLP - Natural Language Processing

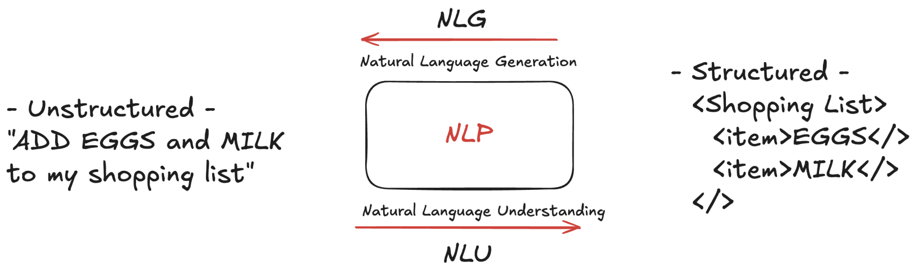
Resources used for the following notes:

1. IBM Technology - What is NLP (Natural Language Processing)?
    a. https://www.youtube.com/watch?v=fLvJ8VdHLA0
2. IBM - What is NLP (natural language processing)?
    a. https://www.ibm.com/topics/natural-language-processing?utm_medium=OSocial&utm_source=Youtube&utm_content=000027BD&utm_term=10004432&utm_id=YTCard-101-What-is-NLP-LH-Natural-Language-Processing-Guide
3. Images and Drawings:
    a. https://excalidraw.com/

A subfield of Computer Science and Artificial Intelligence (AI) that uses Machine Learning to enable computers to understand and communicate with human language.

NLP enables computers and digital devices to recognize, understand and generate text and speech by combining computational linguistics—the rule-based modeling of human language—together with statistical modeling, machine learning and deep learning....

Let's look at an example:



Above is a simple exam how NLP operated on human text.... The computer can process information from speech, lists, writing and come up with a structured output that we can see on the right with the help of NLP.

Some use cases of NLP are as follows:

Sentiment Analysis `(OUR USE CASE)` :

Sentiment Analysis is a powerful application of NLP that aims to determine the emotional tone or polarity of text. It involves identifying and categorizing opinions expressed in a piece of text to gauge whether the sentiment is positive, negative, or neutral. `In algorithmic trading, sentiment analysis is highly valuable` as it can provide insights into public perception around specific stocks, companies, or broader market trends.

In the context of our usage , sentiment analysis can help gauge investor sentiment on equities, such as **TSLA** (Tesla). By implementing NLP on financial news, social media, and other sources, our platform can assess market mood and incorporate this sentiment data to influence trading decisions. For instance, if sentiment around TSLA turns negative due to a major news story, our model may decide to hold back or short the asset, whereas positive sentiment could trigger a buy signal. This approach offers a more nuanced, data-driven input into investment strategies, enhancing the platform's ability to make informed decisions.

Virtual Assistant / Chatbot

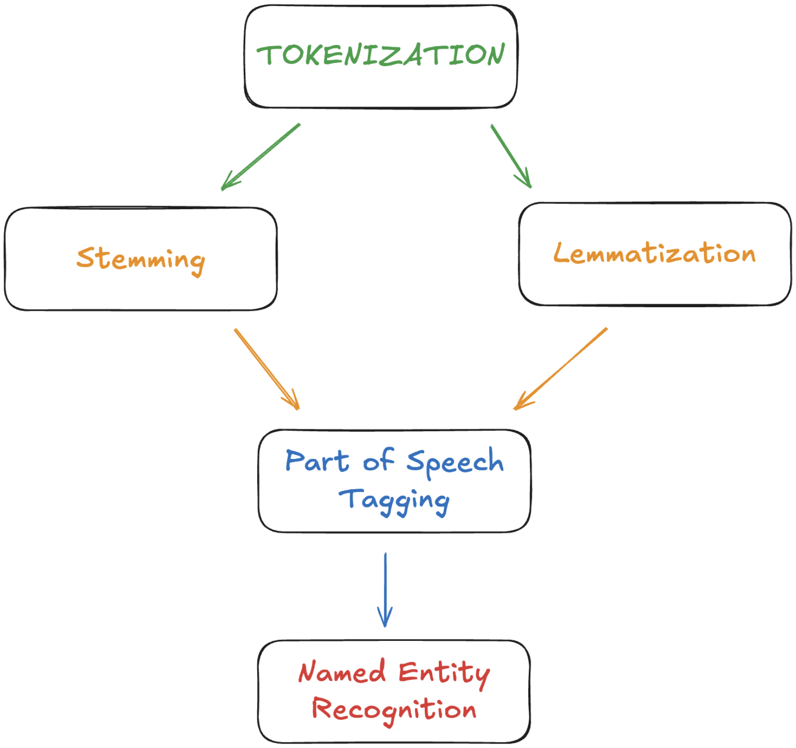
Virtual assistants and chatbots leverage NLP to interact with users in a conversational manner. Through NLP, these systems can understand user requests, respond appropriately, and even perform tasks or answer complex questions. Virtual assistants are now widely used across industries—from customer support chatbots on websites to AI-driven personal assistants like Siri and Alexa. They work by parsing user inputs, identifying intent, and generating contextually appropriate responses, often using a combination of pre-trained models and continuous learning from user interactions.

Machine Translation

Machine Translation is the automated process of translating text or speech from one language to another, making information accessible to a global audience. Advanced machine translation systems, like Google Translate, use deep learning and NLP techniques to understand the context of sentences and provide accurate translations beyond word-for-word replacement. Machine translation can involve large-scale language models trained on extensive multilingual datasets, and improvements in this area have enabled more seamless communication across language barriers.

Spam Detection

Spam detection is an NLP-based application that identifies and filters out unsolicited, irrelevant, or harmful messages, particularly in email. By analyzing the content, structure, and linguistic features of messages, NLP models can classify emails as spam or not. These models are trained on datasets of spam and non-spam messages, learning to recognize patterns, phrases, and characteristics typical of spam. Modern spam detection systems incorporate NLP with additional techniques, such as machine learning and rule-based algorithms, to prevent unwanted messages from reaching users' inboxes, thus ensuring better email security and user experience.



1. Tokenization

Tokenization splits financial news, social media posts, and other text data into individual words or phrases. This is crucial in our project because it allows our model to identify and process key terms relevant to stock sentiment, like “profit,” “risk,” or “growth” in relation to TSLA. Tokenizing text ensures that each word or phrase can be separately analyzed for sentiment.

2. Stemming

Stemming reduces words to their base form, so different variations (like "investing," "invested," and "invests") are all recognized as "invest." For our platform, this simplifies the vocabulary and allows the sentiment model to treat similar words as one concept, ensuring consistency in how it interprets sentiment across different word forms.

3. Lemmatization

Lemmatization provides even more accuracy by reducing words to their root form based on context, such as converting "better" to "good." This is important in financial text, where subtle differences in words can change sentiment. For example, “growth” and “growing” are related but may imply different market conditions. Lemmatization helps our model understand these nuances, improving sentiment accuracy.

4. Part of Speech (POS) Tagging

POS tagging identifies the function of each word (e.g., noun, verb), helping our model focus on sentiment-carrying words like adjectives and verbs. For instance, knowing that “robust” in “TSLA reported robust growth” is an adjective describing growth allows the model to weigh it as positive sentiment. This helps our algorithm distinguish positive or negative tones, impacting our model’s decision on whether to invest.

5. Named Entity Recognition (NER)

NER detects specific entities within text, such as companies, locations, or dates. For our project, NER will help identify mentions of TSLA or other stocks, competitors, or events that could impact sentiment. This enables our model to track sentiment specific to TSLA or similar entities, adding precision to trading decisions.

How NLP Works?

NLP uses computational methods to analyze, understand, and generate human language. Here’s an overview of the typical steps in an NLP pipeline:

1. **Text Preprocessing:** Raw text is prepared by breaking it down into tokens (words or phrases), converting text to lowercase, and removing stop words (e.g., “is,” “the”). Stemming or lemmatization reduces words to their base forms (e.g., “running” to “run”). This cleans and standardizes text for analysis.
2. **Feature Extraction:** Text is converted into numerical data. Techniques like Bag of Words and TF-IDF quantify word presence, while word embeddings (e.g., Word2Vec) capture deeper semantic relationships by representing words in a continuous vector space.
3. **Text Analysis:** The processed text is analyzed to extract meaning. This involves:
   - **POS Tagging**: Identifies grammatical roles of words.
   - **NER**: Detects entities like names and locations.
   - **Sentiment Analysis**: Determines emotional tone.
   - **Dependency Parsing**: Understands sentence structure.
4. **Model Training:** The structured data is used to train machine learning models, which learn patterns in the data for making predictions on new text.

Tools like NLTK (for text processing) and TensorFlow (for model training) support these processes. Together, these steps allow machines to interpret and derive insights from human language.

Challenges of NLP

NLP models, even advanced ones, face limitations due to the complexities of human language. Key challenges include:

1. **Biased Training**: If training data is biased, the NLP model’s outputs will reflect these biases. This is especially problematic in applications like government or healthcare, where diverse users need fair treatment. Web-sourced datasets, commonly used for training, often contain biases.
2. **Misinterpretation:** Speech recognition struggles with dialects, slang, incorrect grammar, or background noise, which can lead to inaccurate transcriptions.
3. **New Vocabulary:** Language constantly evolves with new words and changing grammar. NLP models can struggle with unfamiliar terms, sometimes making incorrect guesses or showing uncertainty.

4. **Tone of Voice:** Sarcasm, emphasis, or emotional tone are challenging for NLP to interpret, as they can change the intended meaning, complicating reliable semantic analysis.

These challenges make it difficult for NLP applications to consistently provide accurate and meaningful insights, requiring continuous improvements in model training and data handling.

Usage of NLP in our Project

In finance, speed and precision are critical, as the market can shift in a matter of seconds. NLP helps financial professionals quickly analyze vast amounts of information, including financial statements(10-K Reports of companies), regulatory filings, news reports, and even social media. By automatically extracting key data and insights, NLP enables faster decision-making, helping analysts identify trends, assess risks, and respond to market events in real-time. This can be especially useful in algorithmic trading, where NLP-driven sentiment analysis of news or social media can trigger buy or sell signals, potentially giving firms a competitive edge.