CS 4063 – Natural Language Processing

Lecture Notes – week 3

Muhammad Hannan Farooq

Text Processing techniques

- Stop words
- Punctuation
- Emojis
- URL’s,
- short Convo’s
- Stemming
- Lemmatization

Stop words

- **Stop words** are common words in a language that are often filtered out before processing text in Natural Language Processing (NLP) tasks.
- These words typically carry little meaning on their own and are not essential for understanding the main content or topic of a document.
- Common stop words include articles, prepositions, conjunctions, and pronouns.

Stop words(Cont.)

- English Stop Words:
  - **Articles:** the, a, an
  - **Prepositions**: in, on, at, by
  - **Conjunctions:** and, or, but, so
  - **Pronouns:** he, she, it, they, we, you
  - **Other Common Words:** is, are, was, were, has, have, of

Stop words(Cont.)

- Example 1 Sentence:
- Original Sentence:
  - "The quick brown fox jumps over the lazy dog.“
- Stop Words Removed:
  - "quick brown fox jumps lazy dog“
- In this example, common stop words like "the", "over", and "the" have been removed, leaving behind the more meaningful words like "quick", "brown", "fox", "jumps", "lazy", and "dog".

Stop words(Cont.)

- Example 2 Sentence:
- Original Sentence:
  - "He is going to the store to buy some groceries for dinner.“
- Stop Words Removed:
  - "going store buy groceries dinner“
- In this example, the stop words "He", "is", "to", "the", "some", "for" have been removed, leaving behind the more meaningful words that convey the key message of the sentence.

Stop words(Cont.)

- Why Remove Stop Words?
  - **Efficiency:** Removing stop words reduces the size of the text and helps focus on important words.
  - **Improved Model Accuracy:** Since stop words often don't contribute much to the meaning of the text, removing them can help NLP models, like text classifiers or search engines, focus on the more relevant terms.

Stop words(Cont.)

- **Text Classification:** Helps models focus on the significant words related to the category of the text.
- **Information Retrieval (Search Engines):** Improves search accuracy by discarding irrelevant words.
- **Topic Modeling:** Focuses on key topics by removing filler words that occur frequently across topics.

Punctuation

- Punctuation refers to the symbols used in writing to separate sentences, phrases, and clauses, and to indicate pauses, emphasis, or structure.
- Common punctuation marks include periods (.), commas (,), question marks (?), exclamation marks (!), quotes (" "), parentheses ( ), colons (:), semicolons (;), hyphens (-), and others.
- Punctuation plays a key role in the structure of text and impacts the way natural language is interpreted. In NLP, how punctuation is treated depends on the specific task and context.

Punctuation(Cont.)

- Breakdown of the role of punctuation in NLP:
  1. Preprocessing and Punctuation Removal
     - **Stop Words of Structure**: Often, punctuation is treated similarly to stop words (common words that don't carry much meaning, such as "the" or "and"). Many NLP tasks, such as text classification or topic modeling, involve removing punctuation to simplify the text and focus on meaningful words
     - **Tokenization:** Punctuation helps define word boundaries during tokenization (breaking text into smaller units like words or sentences). Some NLP systems remove punctuation during tokenization, while others keep it to preserve sentence structure.
     - **Example:** "Hello, world!" might be tokenized as ["Hello", "world"] if punctuation is removed, or ["Hello", ",", "world", "!"] if punctuation is retained.
     - **Normalization**: In some tasks, punctuation might be normalized (e.g., treating different types of quotes as the same symbol or converting hyphens to spaces).
  2. Sentence Segmentation
     - **Punctuation as Sentence Boundaries:** Punctuation, especially periods, question marks, and exclamation points, plays an important role in segmenting text into sentences. This helps in tasks like text summarization or machine translation, where sentence-level understanding is crucial.
     - **Example:** In the text "I went to the store. She stayed home.", the period signals the end of the first sentence and the start of the next one.
  3. Importance of Punctuation in Meaning
     - **Conveying Emotion and Intent**: Punctuation can influence the sentiment or emotion expressed in a sentence. For example, "What are you doing?" conveys a neutral question, whereas "What are you doing?!" conveys urgency or emotion.
     - **Example in Sentiment Analysis**: "I love this!" (positive) "I love this?" (doubtful/negative)
     - **Quotation Marks**: Quotation marks often denote direct speech or emphasize certain words, and this can affect how text is interpreted in tasks like entity recognition or sentiment analysis.

Punctuation(Cont.)

- Breakdown of the role of punctuation in NLP:
    4. Handling Punctuation in Specific NLP Tasks
        - **Named Entity Recognition (NER):** Punctuation may be used to help identify named entities, such as quotes around the names of people, or to distinguish between different parts of a sentence (e.g., separating locations from dates).
            - **Example:** "John Smith, the CEO of Acme Corp., will be visiting Paris next week.“
            - In this sentence, punctuation (commas and period) helps identify the named entities:
            - John Smith **(Person)**
            - Acme Corp. **(Organization)**
            - Paris **(Location)**
        - **Text Summarization:** In some summarization tasks, punctuation helps determine sentence importance. Long, complex sentences broken by commas or semicolons might be treated differently than short, simple sentences.
            - **Example:"**Artificial intelligence is advancing rapidly; it is transforming industries like healthcare, finance, and education. The impact is immense.“
            - In this example, the semicolon helps split the sentence into two parts. The first part contains important information about industries being transformed by AI, while the second part is shorter and less informative in comparison.
        - **Sentiment Analysis:** Punctuation can help capture tone or emotional intensity. Exclamation points, for example, might signal a stronger sentiment in a review or social media post.
            - **Example:"**I love this product!" (positive sentiment)“ I love this product?" (neutral or doubtful sentiment)

Punctuation(Cont.)

- Breakdown of the role of punctuation in NLP:
  5. Punctuation as a Feature in NLP Models
     - **Feature for Machine Learning Models:** In some NLP models, punctuation can be retained as a feature to help the model make better predictions, especially in tasks that depend on sentence structure or tone, such as sarcasm detection. Preserving Context: Some NLP tasks require preserving punctuation to maintain sentence structure, such as dependency parsing, where the relationships between words are mapped, and punctuation is important to understanding sentence syntax.
     - **Preserving Context**: Some NLP tasks require preserving punctuation to maintain sentence structure, such as dependency parsing, where the relationships between words are mapped, and punctuation is important to understanding sentence syntax.
  6. Challenges with Punctuation in NLP:
     - **Ambiguity in Punctuation Usage:** Different uses of punctuation across languages and contexts can make NLP more difficult. For example, a period may end a sentence, but it could also be part of an abbreviation (e.g., "Dr.").
     - **Example:** "I have an appointment with Dr. Smith tomorrow.“
     - Here, the period is part of the abbreviation "Dr." (Doctor) and does not signal the end of the sentence. The actual sentence continues after "Dr." with "Smith tomorrow.“
     - When performing tasks like sentence tokenization, a model may incorrectly interpret the period in "Dr." as the end of the sentence, causing improper splitting.
     - Incorrect segmentation: "I have an appointment with Dr." and "Smith tomorrow.“
     - Correct segmentation: "I have an appointment with Dr. Smith tomorrow.“
     - **Informal Writing (Social Media):** On platforms like Twitter, users often omit or misuse punctuation, which can confuse models trained on formal text. Emoticons, emojis, and excessive punctuation (e.g., "!!!") may need to be handled specifically.

Emojis

- Emojis are an increasingly important aspect of communication, particularly in informal and social media contexts.
- In Natural Language Processing (NLP), dealing with emojis poses unique challenges and opportunities since they can convey sentiment, emotion, and intent in ways that traditional words might not.

Emojis(Cont.)

- Challenges of Emojis in NLP:
  - **Multimodal Communication:** Emojis often provide contextual clues that complement or modify the meaning of text. **For example,** "I'm so happy 😊" expresses positive emotion more strongly than just "I'm so happy.“
  - **Ambiguity:** Many emojis can have multiple meanings depending on context. **For example**, A person from a culture that interprets the 🙏 emoji as a prayer gesture might send it to express thanks or gratitude: "Thank you for your help 🙏.“
  - **Sentiment Modification:** Emojis often modify the sentiment of a sentence. **For example,** the sentence "I'm fine 😢" expresses sadness, while "I'm fine 😁" expresses happiness or sarcasm.
  - **Standardization**: Emojis are pictographic symbols that have standardized Unicode representations, but they may appear differently across platforms (e.g., Apple, Android), which could lead to differences in interpretation.

Emojis(Cont.)

- Approaches to Handling Emojis in NLP:
  - Approaches to Handling Emojis in NLP:
    - **Emoji Tokenization:** Emojis are often treated as separate tokens, similar to words. Tokenizing them allows NLP models to identify and interpret their role in the text. **Example**: In a sentence like "I'm happy 😊", the emoji would be treated as its own token.
    - ["I'm", 'so', 'happy', '😊', '!’]
    - **Emoji Sentiment Analysis:** Emojis can be incorporated into sentiment analysis models. Many sentiment analysis tools include predefined sentiment scores for emojis (e.g., 😊 is positive, 😢 is negative). **Example:** A model can recognize that "I'm sad 😢" has a negative sentiment because of the emoji.
    - **Emoji Embeddings:** Emojis can be embedded in the same vector space as words using embeddings like Word2Vec or GloVe. This allows models to understand emojis in context and relate them to words with similar meanings.
    - **Emoji2Vec:** This is a pre-trained model that represents emojis as vectors, similar to Word2Vec, allowing models to interpret their meaning in relation to surrounding text.
    - **Emoji Sentiment Lexicons:** Prebuilt lexicons such as the Emoji Sentiment Ranking provide sentiment scores for common emojis based on their usage in large corpora.
    - **Example:**😊 (Smile): Positive sentiment😢 (Crying Face): Negative sentiment
    - **Contextual Meaning of Emojis:** NLP models like BERT and GPT-3 can understand emojis in the context of a sentence. For example, "Good job 👍" conveys positive feedback, whereas "Good job 👍😒" can convey sarcasm due to the combination of emojis.
    - **Transformers:** Emojis can be fed into transformer-based models to capture their context and meaning alongside words.
    - **Emoji Translation:** Emoji Prediction or Replacement: Some systems use NLP models to predict the next emoji based on the context of the sentence, or to translate emojis into textual descriptions.
    - **Example:**Input: "I'm so happy 😊"Translation: "I'm so happy [smiling face emoji]"

Emojis(Cont.)

- Example Use Cases of Emojis in NLP:
  - Sentiment Analysis on Social Media
  - Customer Feedback Analysis
  - Emoji-Based Chatbots and Virtual Assistant

URL’s

- URL’s (Uniform Resource Locators) are often found in textual data, such as web pages, social media posts, and documents.
- Handling URL’s in NLP is important because they can carry specific information or disrupt text processing if not managed correctly.

URL’s(Cont.)

- Common Challenges and Techniques for Handling URLs in NLP:
  - **Noise Removal:** URLs in text can be considered noise in many NLP tasks, such as sentiment analysis, text classification, and summarization. They often don’t contribute meaning to the context and can disrupt tokenization or word frequency analysis.
  - **Example:** In a tweet like "Check out this website: https://example.com", the URL might be irrelevant for sentiment analysis, so it’s removed during preprocessing.
  - **Tokenization Issues:** URLs can be problematic during tokenization, especially if they are treated as multiple words**.**
  - **For example**, "https://example.com" might be split into tokens like ["https", ":", "/", "/", "example", ".", "com"], which could distort analysis.
  - **Solution:** You can treat the entire URL as a single token or remove it completely, depending on the task.
  - **Feature for Information Extraction:** In some NLP tasks, such as information extraction or web scraping, URLs are important as they point to additional resources or provide metadata about the content.
  - **For example**, extracting URLs from HTML pages can help in web crawling or identifying hyperlinks in news articles or blogs.
  - **Text Normalization:** During text normalization (a preprocessing step), URLs can be replaced by a placeholder like <URL> to retain the structure of the sentence while removing the actual URL.
  - **Example:** Original Sentence: "Find more details at https://example.com."Normalized Sentence: "Find more details at <URL>."

URL’s(Cont.)

- Common Challenges and Techniques for Handling URLs in NLP:
  - **Named Entity Recognition (NER):** In tasks like Named Entity Recognition, URLs may be tagged or classified as entities, especially in technical documents or datasets where URLs provide important information.
  - **Example**: In a dataset of research papers, URLs might be treated as "RESOURCE" entities.
  - **Spam and Phishing Detection:** NLP can be applied to detect spam or phishing attempts in emails, social media, or messaging platforms by analyzing the occurrence of suspicious URLs.
  - **Example:** An NLP-based system might classify an email as spam if it contains URLs that match known patterns of phishing attacks.
  - **Social Media Analysis:** URLs are commonly included in social media posts (e.g., tweets). When analyzing text from platforms like Twitter, URLs can be either removed or retained depending on the analysis objective.
  - **For example**, when analyzing sentiment or engagement, the presence of a URL can indicate external content or references.
  - **Example of URL in a Tweet:** Tweet: "Amazing product! Check it out: https://example.com #bestproduct“
  - URL Removal for Sentiment Analysis: The URL is not important for determining sentiment and can be removed during preprocessing.

URL’s(Cont.)

- Common Challenges and Techniques for Handling URLs in NLP:
  - **Text Summarization and Topic Modeling:** URLs in documents (e.g., research papers, news articles) may provide important references or sources. In summarization or topic modeling, URLs might be retained to indicate the presence of external resources, or they may be replaced with a generic placeholder.
  - **Contextual Information from URLs:** Sometimes, URLs themselves contain valuable contextual information, such as the domain name or specific keywords in the URL path.
  - **For instance**, a URL like https://news.example.com/sports/football-highlights contains information about the type of content (sports, football highlights).
  - In this case, the domain or URL path could be useful for classification or understanding the topic of the text.

URL’s(Cont.)

- Techniques for Handling URLs in NLP:
  - **URL Detection and Removal:** Use regular expressions to detect and remove URLs from text during preprocessing.
  - **URL Tokenization:** You may choose to tokenize URLs as a single token, especially if the presence of a URL is important for the analysis.
  - **Example**: Before tokenization: https://example.com After tokenization: <URL>
  - **Extracting Features from URLs**: If URLs are meaningful, extract features like the domain name or path keywords for use in downstream tasks like classification or clustering.

URL’s(Cont.)

- Use Cases for URLs in NLP:
  - Web Scraping and Information Retrieval
  - Ad and Spam Detection
  - Content Classification

Short Convo’s

- In Natural Language Processing (NLP), short conversations (Convo's) refer to brief exchanges between individuals, typically seen in settings like customer support chats, text messaging, or social media interactions.
- These short convos can be challenging to process due to their brevity, informal language, and lack of context.

Short Convo’s(Cont.)

- Key Aspects of Short Conversations in NLP:
  - **Contextual Understanding:** Short conversations often lack detailed context, making it harder for NLP models to understand the full intent behind a message.
  - **Example:** Person A: "Got it.“ Person B: "Thanks!"Context: It’s difficult to know what "it" refers to without previous information.
  - **Informality and Slang:** Short conversations often include slang, abbreviations, and informal language, which can be difficult for traditional NLP models to interpret.
  - **Example:** "ttyl" (talk to you later), "idk" (I don't know), or emojis like 😊 or 👍.
  - **Elliptical Sentences:** Short convos frequently omit parts of speech, relying on the other person to understand the context.
  - **Example: "**Going to the store.“ This sentence lacks a subject but is understood in context.
  - **Response Generation:** NLP is often used to generate short responses in systems like chatbots or virtual assistants, requiring the model to understand the conversation's flow and intent.
  - **Example:** User: "What’s the weather?"Bot: "It’s sunny today.“
  - **Sentiment Detection:** Short texts, especially in conversations, are commonly analyzed for sentiment, particularly in customer service or social media monitoring.
  - **Example:** User: "This product is great!“ Sentiment: Positive

Short Convo’s(Cont.)

- Key Aspects of Short Conversations in NLP:
  - **Entity Recognition:** Extracting key entities (e.g., names, dates, products) from short texts is essential in customer service or messaging apps.
  - **Example**: "Meeting at Starbucks at 5pm.“ Entities: "Starbucks" (location), "5pm" (time).

Short Convo’s(Cont.)

- NLP Techniques for Handling Short Convo’s:
  - **Text Preprocessing:** Cleaning up informal language, abbreviations, and removing stop words to normalize short texts.
  - **Contextual Embeddings:** Models like BERT or GPT-3 can understand short texts better by capturing context from surrounding conversations.
  - **Dialogue Management**: Techniques like intent detection and entity recognition are used to manage short conversations and provide appropriate responses in chatbots.
  - **Sentiment Analysis:** Identifying emotions in short texts, even with sparse data, using pretrained models or lexicons.

Stemming

- Stemming is a technique in Natural Language Processing (NLP) that involves reducing a word to its root form (stem) by removing suffixes or prefixes.
- The goal of stemming is to group words with the same base meaning but different forms (e.g., "run," "running," "runner") together, allowing NLP models to treat them as the same entity.

Stemming(Cont.)

- Importance of Stemming in NLP:
  - **Reducing Variants of Words:** By converting different forms of a word to its root, stemming helps in normalizing text and reduces the number of unique tokens.
  - **Example:** "jumps," "jumped," "jumping" → "jump“
  - **Improving Search and Information Retrieval:** Stemming allows search engines or NLP systems to match documents or texts that use different word forms.
  - **Example:** Searching for "connecting" will also return results with "connect," "connected," and "connection.“
  - **Text Preprocessing:** Stemming is a common step in preprocessing for tasks like sentiment analysis, text classification, and information retrieval, where reducing the vocabulary size is beneficial.

Stemming(Cont.)

- Example of Stemming in NLP:
  - Common Stemming Algorithms:
  - Porter Stemmer: One of the most commonly used stemming algorithms, which applies a series of rules to remove suffixes.
  - Snowball Stemmer: A more advanced version of the Porter Stemmer, with language-specific improvements.
  - Lancaster Stemmer: A more aggressive stemming algorithm that tends to produce shorter stems.

- Benefits of Stemming:
  - **Dimensionality Reduction:** By converting multiple forms of a word to a single stem, stemming helps reduce the size of the vocabulary, making it easier for models to process.
  - **Efficiency in Search Engines:** When users search for a term, stemming ensures that documents containing variations of that term are also retrieved.

Stemming(Cont.)

- Challenges of Stemming:
  - **Over-Stemming:** Stemming can sometimes result in removing too many characters, which can change the meaning of the word.
  - **For example,** "university" might be stemmed to "univers," or “uni” which isn’t meaningful.
  - **Loss of Meaning:** Stems produced by the algorithms may not always be valid words (e.g., "easily" stemmed to "easili"), and this can result in a loss of interpretability.

Stemming vs. Lemmatization

- **Stemming** involves cutting off the ends of words to get to a base form, without considering the word's meaning or context.
- **Lemmatization** is more sophisticated, as it reduces words to their root forms (lemmas) based on their meaning and context (e.g., "better" becomes "good").

Lemmatization

- Lemmatization in Natural Language Processing (NLP) is the process of reducing a word to its base or dictionary form (lemma) by considering its meaning and context.
- Unlike stemming, which blindly removes suffixes or prefixes, lemmatization uses morphological analysis and contextual information to return the actual root form of a word.

Lemmatization(Cont.)

- Why Lemmatization is Important in NLP:
  - **Contextual Understanding:** Lemmatization helps retain the meaning of a word by reducing it to its correct root form based on its part of speech (POS).
  - **Example:** Verb: "running" → "run“
  - **Example**: Noun: "runners" → "runner“
  - **Improved Accuracy in Text Processing:** Lemmatization improves accuracy in NLP tasks like sentiment analysis, text classification, and information retrieval by treating different inflections of a word as the same word.
  - **Example:** "connects", "connected", and "connecting" all reduce to "connect" when lemmatized, ensuring consistent treatment of word variations.
  - **Preprocessing for NLP Models:** It is used in the preprocessing phase for text normalization, enabling models to better handle the natural variations in language.

Lemmatization(Cont.)

- Benefits of Lemmatization:
  - **Accuracy:** Lemmatization provides more accurate results than stemming because it considers the grammatical structure and meaning of the word.
  - **Semantic Understanding:** By recognizing relationships between words, lemmatization ensures that the root form retains the original meaning, which is especially important in tasks like machine translation or summarization.
  - **Reduced Noise:** Lemmatization reduces noise in text by converting different forms of a word (e.g., "am", "are", "is" → "be") to a common root, helping the NLP model focus on the core meaning.

Lemmatization(Cont.)

- Challenges of Lemmatization:
  - **Slower Processing:** Lemmatization is more computationally expensive than stemming since it involves morphological analysis and requires the context or part of speech of the word.
  - **Need for Part of Speech Tagging:** Lemmatization often requires knowing the part of speech (POS) of a word to accurately return its base form.
  - **For instance**, "saw" as a noun and "saw" as a verb have different lemmatizations ("saw" and "see," respectively).

Lemmatization(Cont.)

- Use Cases of Lemmatization in NLP:
  - **Text Classification:** Lemmatization is used to normalize text for tasks like spam detection, where the variations of words like "run," "running," or "runs" should be treated as the same word.
  - **Search Engines:** Lemmatization helps search engines return better results by treating different forms of a word (e.g., "connected," "connects") as one.
  - **Sentiment Analysis**: By reducing inflections of words to their base forms, lemmatization helps in understanding sentiment better, especially when handling informal or diverse text inputs.