Brandon L Medina
166827
IST 332

NLP Final Project – Milestone Check

Google Colab Link –
https://colab.research.google.com/drive/1QdV6y5OYibCY03V157N0DwoDw9gvrLA0?usp=sharing

Part 1 – Introduction & Research Question

Problem Statement

While structured data collected during MLB games captures the statistical aspect of the game, fan reactions expressed through social media remain largely unstructured and unanalyzed. This gap limits an organization's ability to monitor fan engagement, identify moments of heightened interest, and assess sentiment in real-time. By applying natural language processing models that align social media comments with official play-by-play data, this research aims to transform unstructured social media discourse into structured insights, thereby bridging the gap between quantitative game data and qualitative audience reactions.

Research Question

How effectively can machine learning and natural language processing (NLP) models detect and classify Major League Baseball (MLB) in-game events in real time using social media comments that are temporally aligned with official play-by-play data?

Sub Questions

- How can temporal alignment be accurately used between Reddit comments and MLB play-by-play data to approximate actual game events?
- How do classic NLP methods like TF-IDF compare to transformer-based models like BERT or Longformer in predicting events from fan discourse?
- Can models trained on comments with known outcomes effectively generalize to unseen games or new social media platforms (e.g., X/Twitter)

Targeted Variables

- Relevance classification – determines whether a comment refers to an actual in-game event.
- Event classification – identifies the specific event that occurred in-game (e.g., homerun, strikeout, stolen base) based on the comment text.

Part 2 – Corpus Creation (Data Collection)

Corpus Creation

- Collect at least 10,000 comments from r/baseball "Game Thread" posts and associated social-media sources.
- Retrieve official MLB play-by-play data using the pybaseball API.

- Parse timestamps and align each comment with the nearest in-game event to produce weak labels for supervised learning.
- Save the corpus in a structured format (CSV) with fields for comment text, upvotes, timestamp, team/game ID, and assigned event label.

Data Collection

- The dataset is constructed by combining social-media commentary with official MLB game data. Reddit comments are collected from r/baseball "Game Thread" posts using the PRAW API, which provides authenticated access to comment text, timestamps, upvotes, and metadata needed for temporal alignment.
- To pair fan reactions with in-game events, official MLB play-by-play data is retrieved through the pybaseball library. This includes pitch-level and event-level records required for generating approximate labels.

Ethical Considerations

To ensure responsible and compliant use of user-generated content, the following ethical practices are applied to the dataset:

- Use of Public Data Only: All comments are collected from publicly available Reddit Game Thread posts
- Compliance With Platform Policies: Data access occurs exclusively through the official Reddit API
- Removal of Non-Human or Irrelevant Content: Comments from bots, AutoModerator, deleted accounts, or system-generated posts are removed to preserve the integrity of the corpus.
- Minimizing Identifiability: Personal identifiers are avoided wherever possible to ensure users cannot be directly traced.

To support anonymity, the comment_author column will be dropped, it is not required for any analytical steps. Uniqueness can be fully maintained through the comment_id, which is sufficient for linking, grouping, and identifying records during processing.

Summary Statistics

```
=== Summary Statistics ===
Total comments: 17,500
Total posts represented: 7

Average comment length: 12.33 tokens
Min comment length: 1
Max comment length: 347

Vocabulary size: 23,747

Comment score distribution:
count    17500.00000
mean        12.68360
std          8.93844
min          5.00000
25%          8.00000
50%         11.00000
75%         14.00000
max        284.00000
Name: comment_score, dtype: float64

Dataset date range: 2025-10-24 23:02:08+00:00  to  2025-11-02 04:21:47+00:00
```

Brandon L Medina
166827
IST 332

Part 3 – Text Preprocessing

1. Foundation

Filtered out NaN from descriptions to remove non-event-related comments, removed non-text comments like GIFs, and deleted short comments with three characters or less.

2. Normalization

Lowercased all comments to reduce vocabulary sparsity, which decreases duplicate lexical forms.

3. Remove URLs, Emojis, and non-linguistic noise

After lowercasing, I removed visual elements and non-linguistic noise like URLs, Reddit markdown, GIF placeholders, and emojis to prevent tokenization into meaningless tokens.

4. Remove punctuation and special characters

Removed any non-alphabetic characters and punctuation that could inflate vocabulary, as well as any potential leftover symbols from normalization.

5. Tokenization

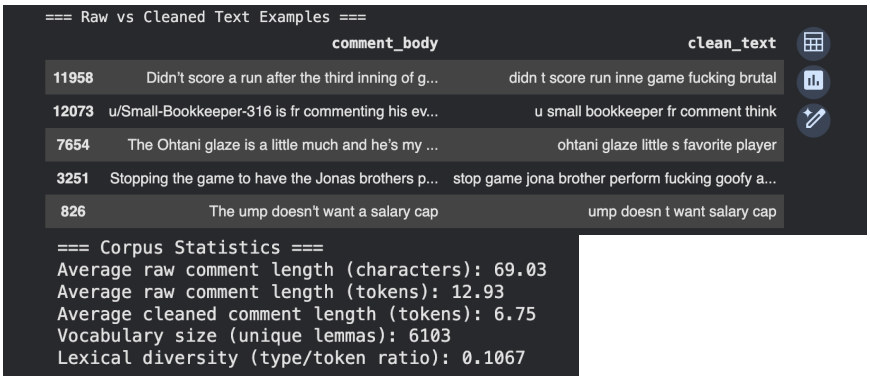
Tokenization occurs after converting text to lowercase, removing noise, and stripping punctuation. This process ensures tokens are pure alphabetic units, free of URLs, emojis, or punctuation. It helps prevent the tokenizer from producing many irrelevant tokens, reduces vocabulary size, and supplies clean data for lemmatization and stopword removal.

6. Removed stopwords

I performed stopword removal after tokenization because stopwords need to be evaluated individually. Punctuation and noise must be eliminated first to prevent false matches. Removing stopwords at this stage also reduces the number of tokens sent to lemmatization, thereby improving efficiency.

7. Lemmatization

I chose lemmatization over stemming because it transforms each token into its base or dictionary form (lemma). This method retains the core meaning of words by combining inflected or conjugated variations.



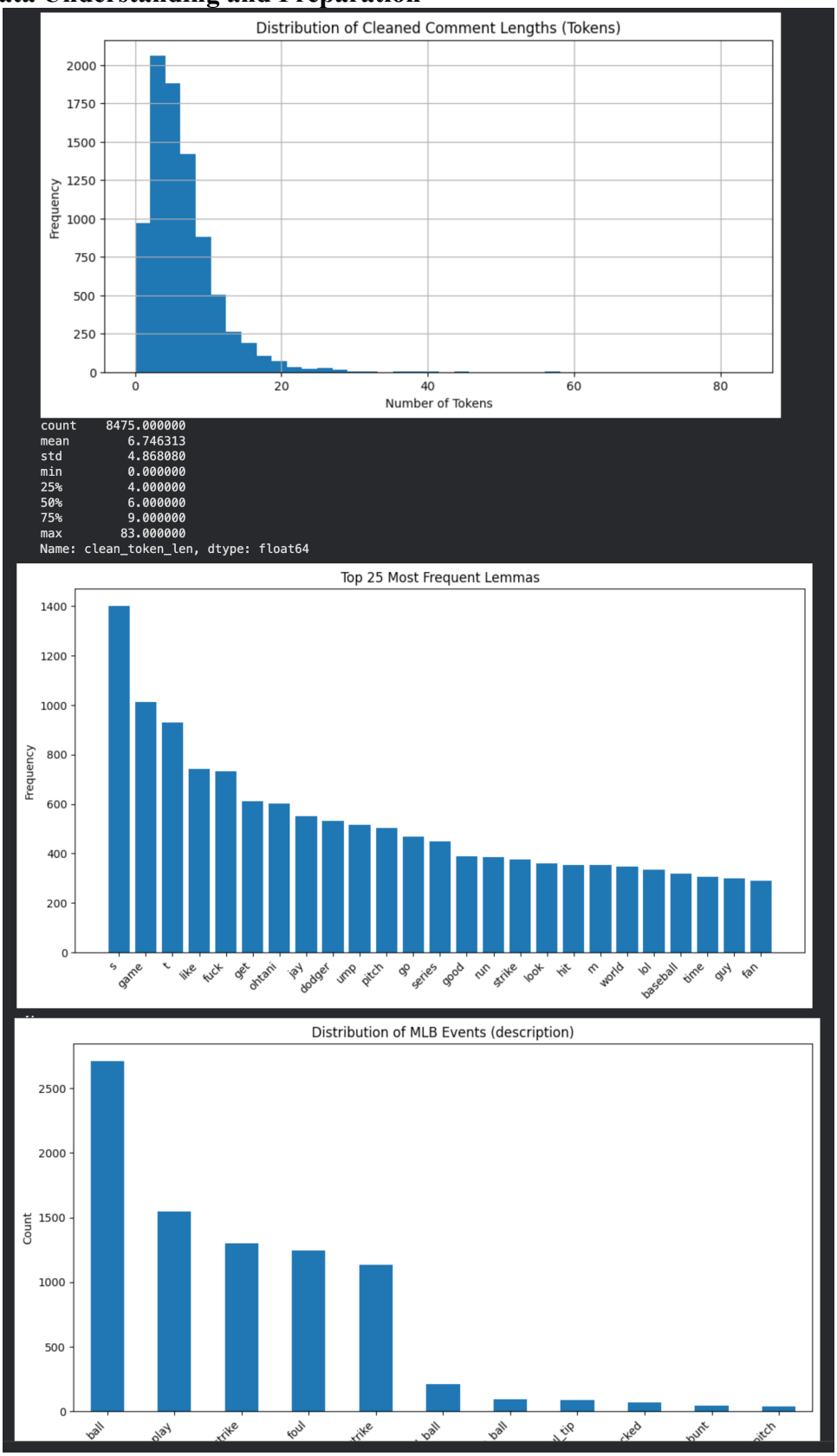
```
=== Raw vs Cleaned Text Examples ===
```

	comment_body	clean_text
11958	Didn't score a run after the third inning of g...	didn t score run inne game fucking brutal
12073	u/Small-Bookkeeper-316 is fr commenting his ev...	u small bookkeeper fr comment think
7654	The Ohtani glaze is a little much and he's my ...	ohtani glaze little s favorite player
3251	Stopping the game to have the Jonas brothers p...	stop game jona brother perform fucking goofy a...
826	The ump doesn't want a salary cap	ump doesn t want salary cap

```
=== Corpus Statistics ===
Average raw comment length (characters): 69.03
Average raw comment length (tokens): 12.93
Average cleaned comment length (tokens): 6.75
Vocabulary size (unique lemmas): 6103
Lexical diversity (type/token ratio): 0.1067
```

Brandon L Medina
166827
IST 332

Part 4 – Data Understanding and Preparation




```
count    8475.000000
mean        6.746313
std         4.868080
min         0.000000
25%         4.000000
50%         6.000000
75%         9.000000
max        83.000000
Name: clean_token_len, dtype: float64
```