Nile University
جامعة النيل

BIAS DIRECTION DETECTION

- Amr Hossam 221000832
- Hossam Nasr 221000770
- Fares Ahmed 221000570
- Mohamed Nashaat 221001565
- Shorouk Sherif 221000645
- Amenah Medhat 221001792

DATA ANALYSIS

1. Initial Dataset Inspection:

Action: Loaded the dataset and examined its structure.

Insights:

1. The dataset contains 88,500 records with columns for Arabic text and bias labels.
2. No sinificant missing values in key columns (arabic_mt, label).
3. All columns have appropriate data types for further analysis.

2. Summary Statistics:

Action: Calculated sentence length (in characters) and word count.

Insights:

1. Average Sentence Length: The average sentence contains 50.84 words and 319.17 characters.
2. Sentence Length Distribution: Most sentences have a word count between 17 to 48 words (from the 25th to the 75th percentile)

3. Bias Label Distribution:

Action: Analyzed the distribution of bias labels across the dataset.

Insights:

1. Imbalance Observed: The Unbiased label is the most frequent, followed by Biased against Palestine and Biased against Israel.
2. This class imbalance may require special handling techniques during model training to ensure fair classification.

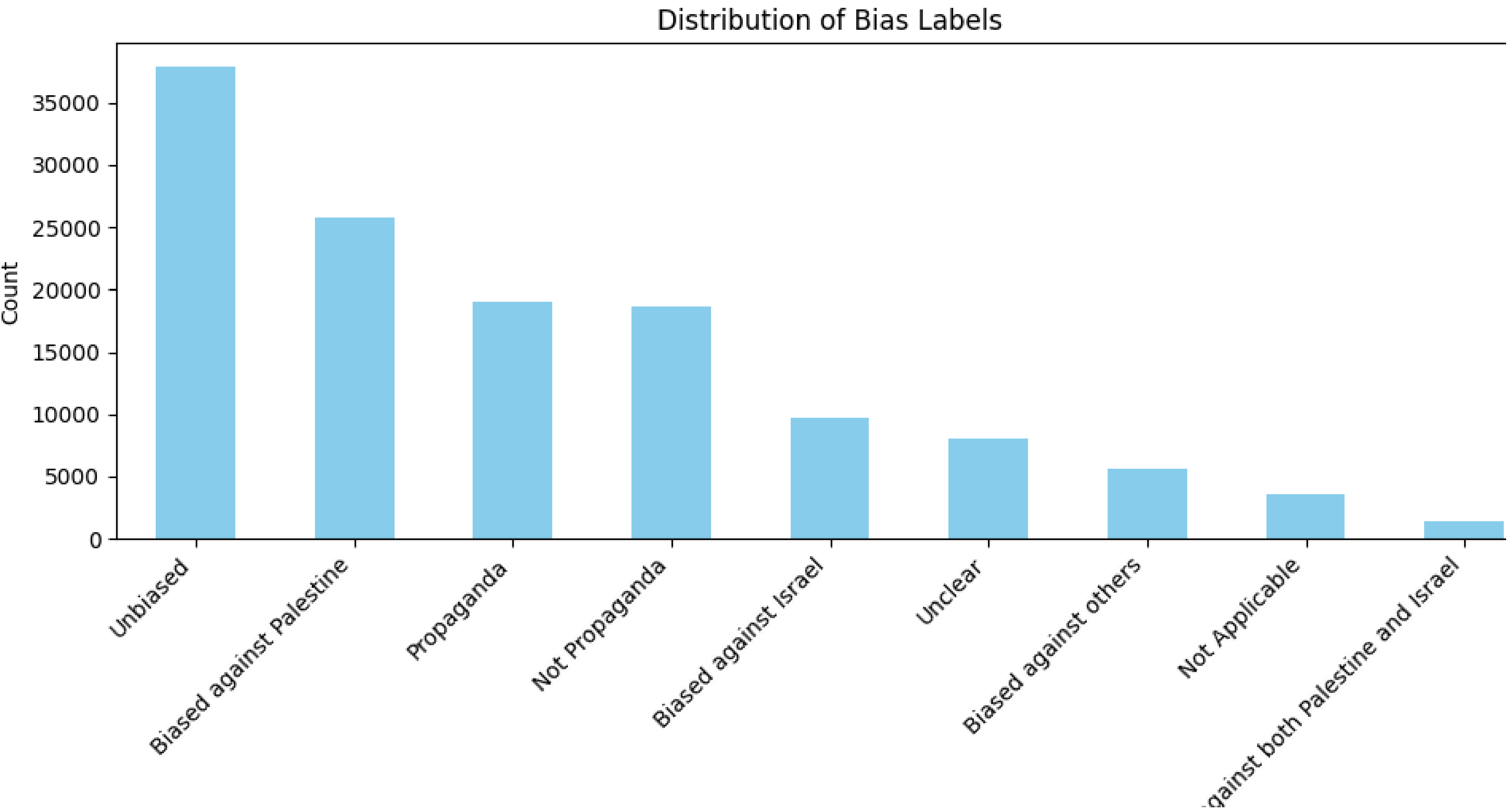
Nile University
جامعة النيل

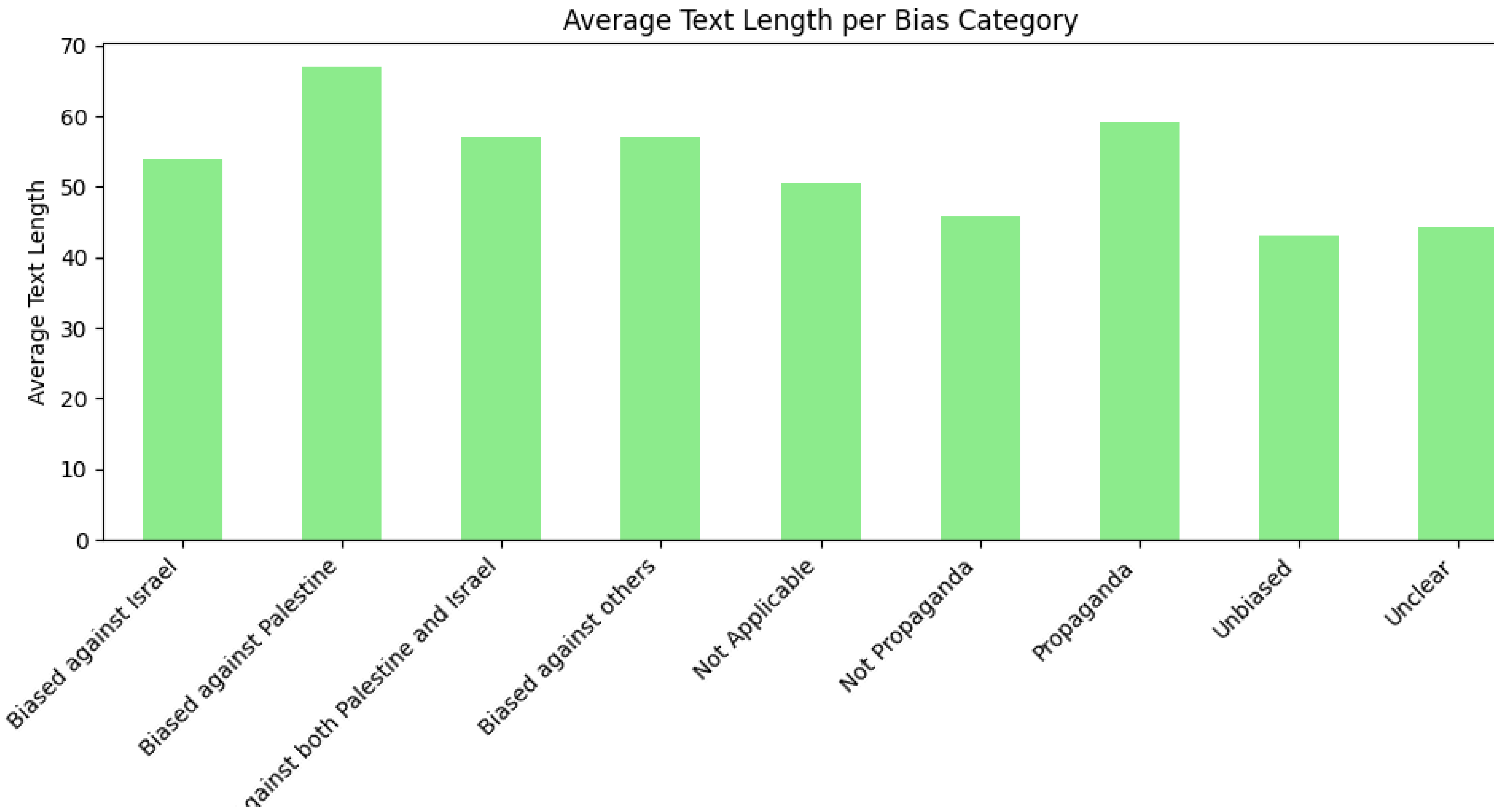
DATA VISUALIZATION

Nile University
جامعة النيل

DATA VISUALIZATION

1 Visualizing Class Imbalance using Bar Charts


Distribution of Bias Labels

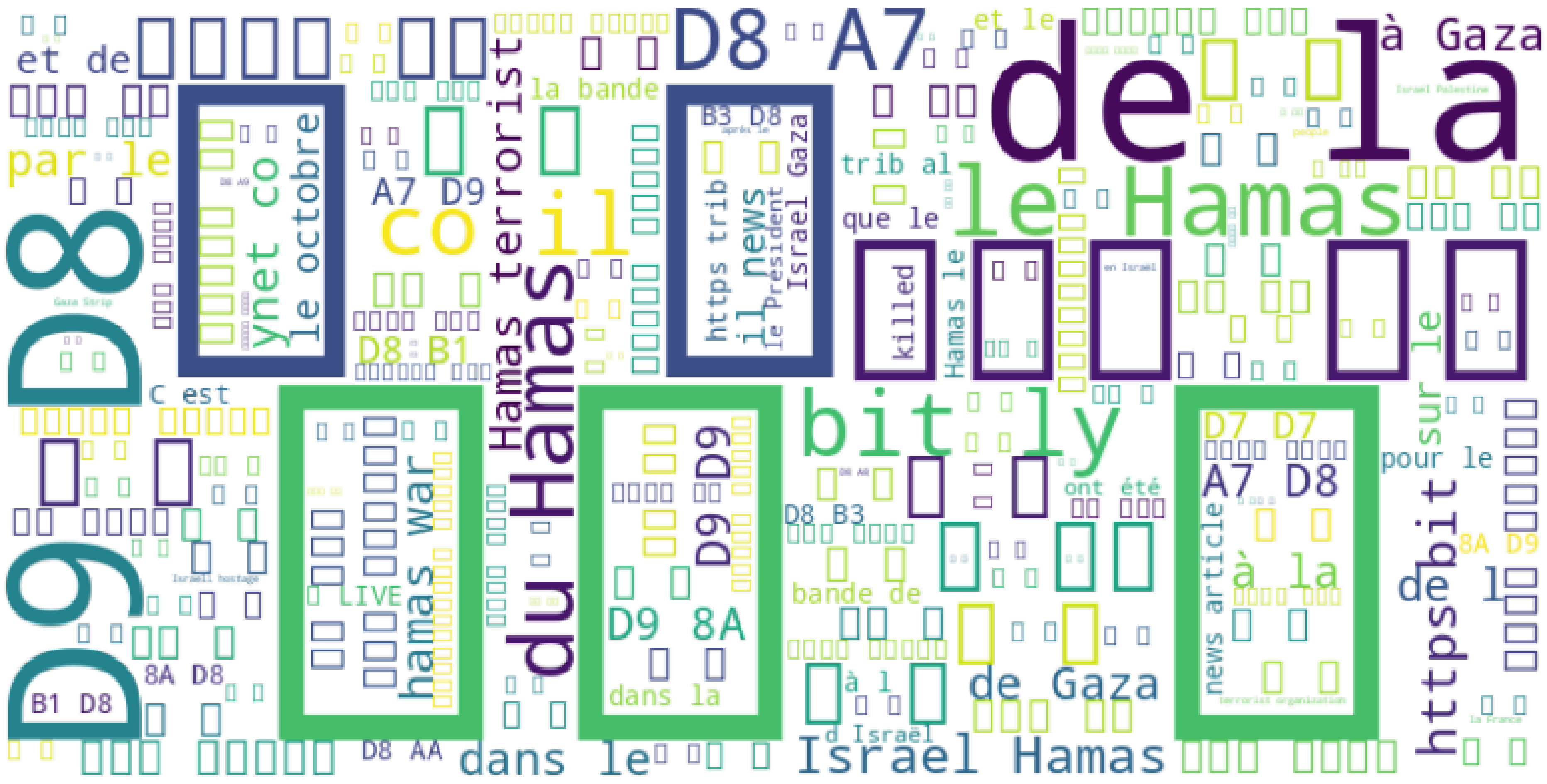

Average Text Length per Bias Category

Nile University
جامعة النيل

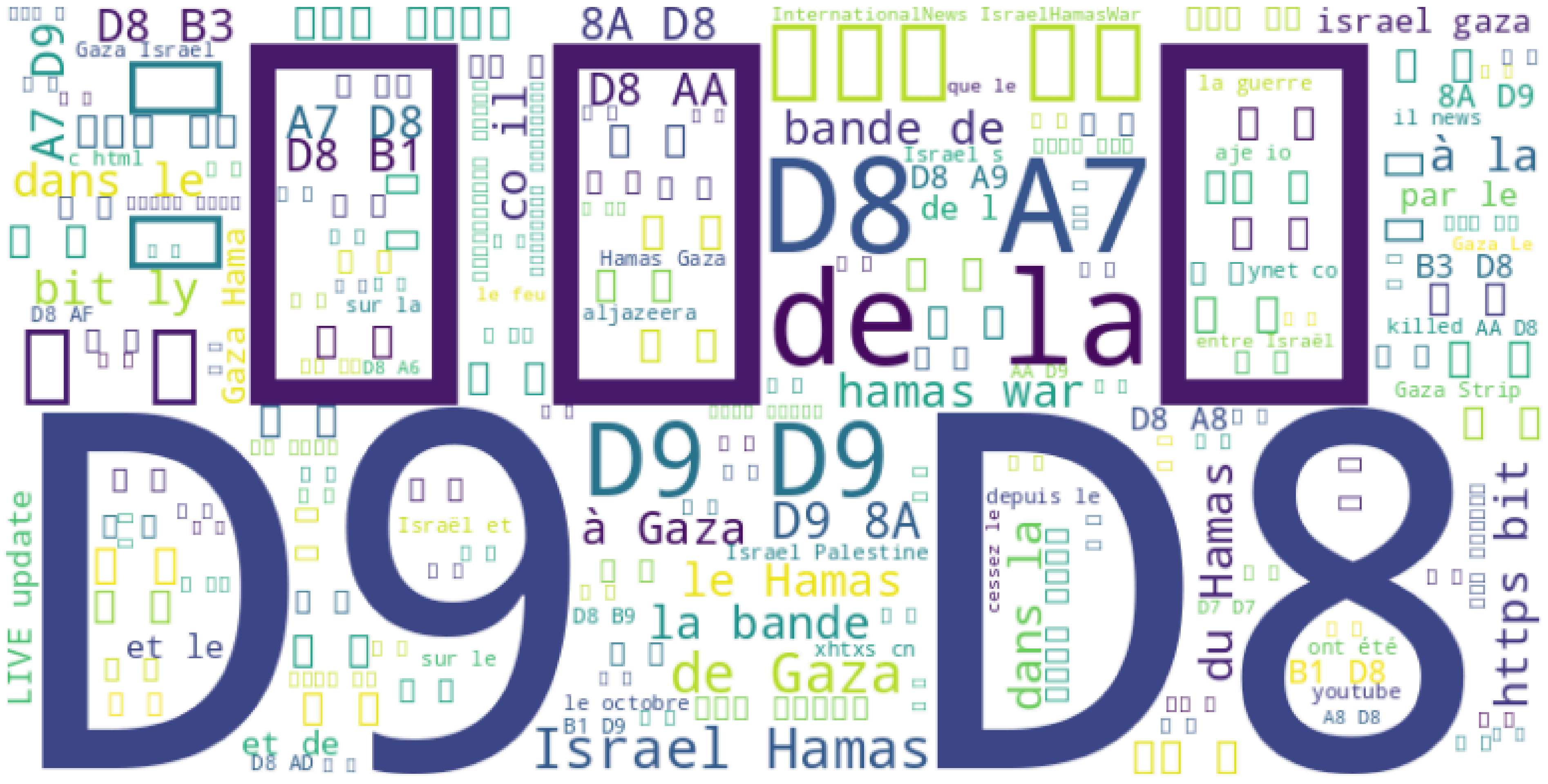
DATA VISUALIZATION

2 Visualizing Word Frequency using Word Clouds

Word Cloud for Bias: Biased against Palestine



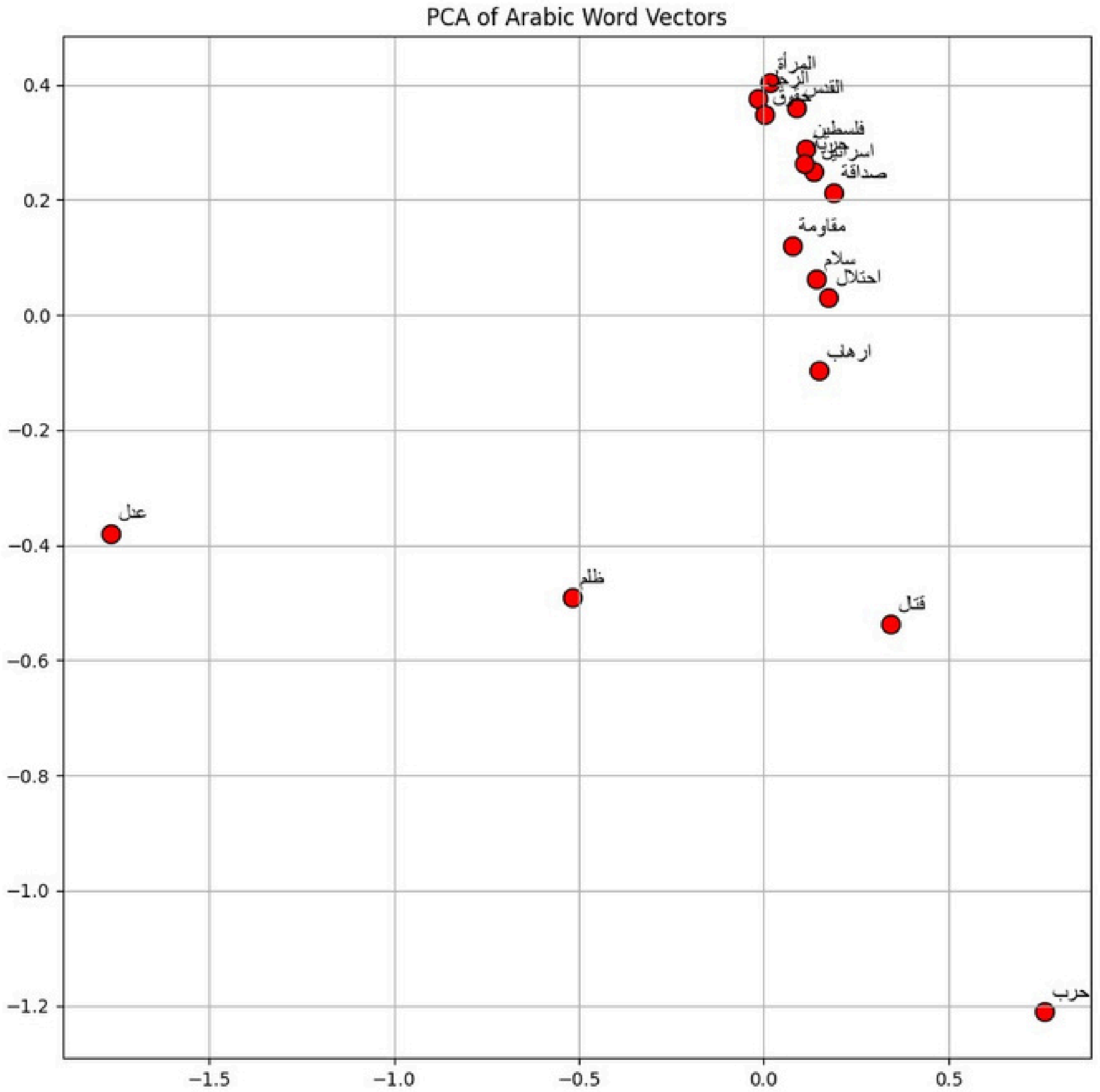
Word Cloud for Bias: Unbiased

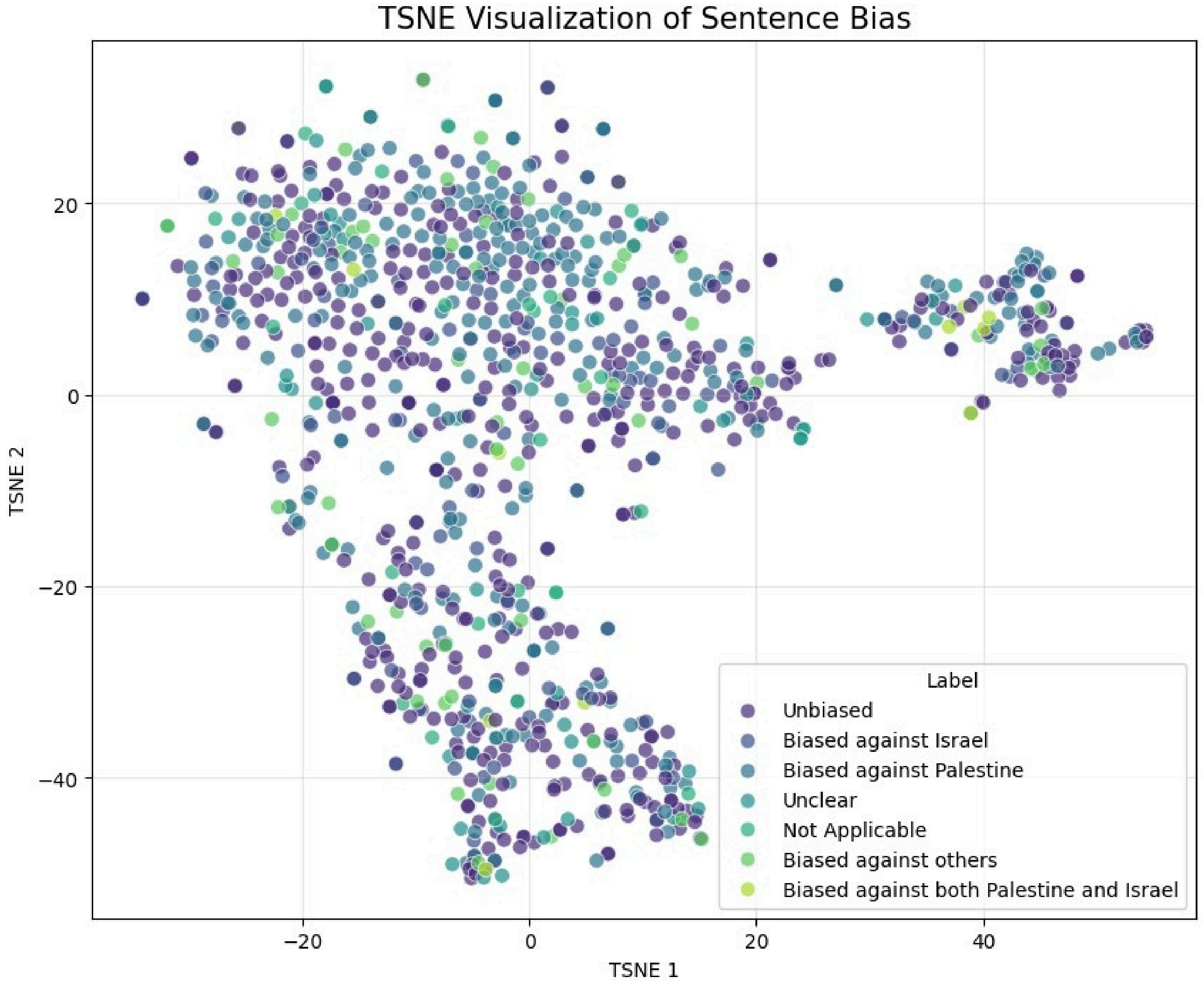


Nile University
جامعة النيل

DATA VISUALIZATION

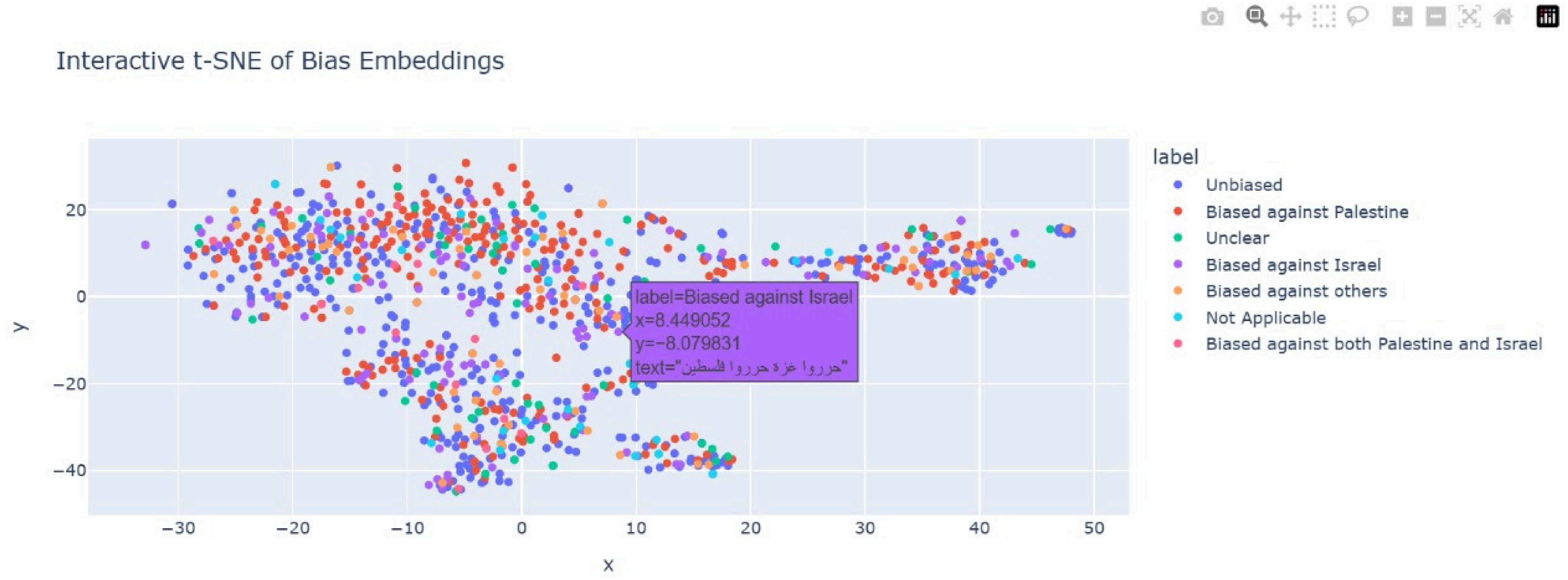
3 PCA & t-SNE Visualizations of Bias Labels





DATA VISUALIZATION

4 Interactive t-SNE map



Nile University
جامعة النيل

DATA PREPROCESSING

1. Tokenization:

The Arabic sentences were tokenized into individual words to break down the text into smaller units for further analysis.

2. Stopword Removal:

Common stopwords (words that do not carry significant meaning) were removed from the tokenized sentences to reduce noise and improve model performance.

3. Lemmatization:

Words were reduced to their base forms using lemmatization to ensure that different forms of the same word were treated as the same token.

4. Embedding Generation:

Word embeddings were generated for both the tokenized text and the stopword-free text using pretrained FastText embeddings. These embeddings represent each word as a vector in a high-dimensional space, capturing semantic meaning.

Nile University
جامعة النيل

THANK YOU!!