Dataset: BBC

- Consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005.
- Class Labels: 5 (business, entertainment, politics, sport, tech)

Link: http://mlg.ucd.ie/datasets/bbc.html

Methodology

1. Named Entity Recognition (NER)

Tools Used: Utilized Python with the spaCy library for NER.

Process:

- Loaded the dataset consisting of text files categorized into five topics.
- Applied the spaCy NLP model to extract named entities (persons, organizations, locations, etc.) from each document.
- Stored the results in a DataFrame (entities_df) with columns for the entity, its category, and the associated topic.

2. Sentiment Analysis

Tools Used: Employed TextBlob for sentiment analysis.

Process:

- Created a new DataFrame (documents_df) containing the full text of each document and its topic.
- Performed sentiment analysis on each document to obtain sentiment scores.
- Added these scores as a new column in documents_df.

3. Topic Modeling

Tools Used: Used scikit-learn for LDA (Latent Dirichlet Allocation) topic modeling.

Process:

- Preprocessed the text data (lowercasing, removing punctuation/numbers, tokenization, removing stopwords, lemmatization).
- Vectorized the text using CountVectorizer.
- Applied LDA to discover underlying topics in the corpus.
- Assigned the dominant topic to each document and analyzed the distribution.

Challenges Faced

1. NER Challenges

Data Cleaning: Required thorough preprocessing of the text data to ensure accurate entity recognition.

Computational Resources: Processing a large volume of documents was resource-intensive.

2. Sentiment Analysis Challenges

Contextual Nuances: Capturing the correct sentiment of the text was challenging due to linguistic nuances, sarcasm, and contextual meanings.

Alignment with Topics: Correlating sentiments with topics required careful consideration, especially in cases where topics overlapped.

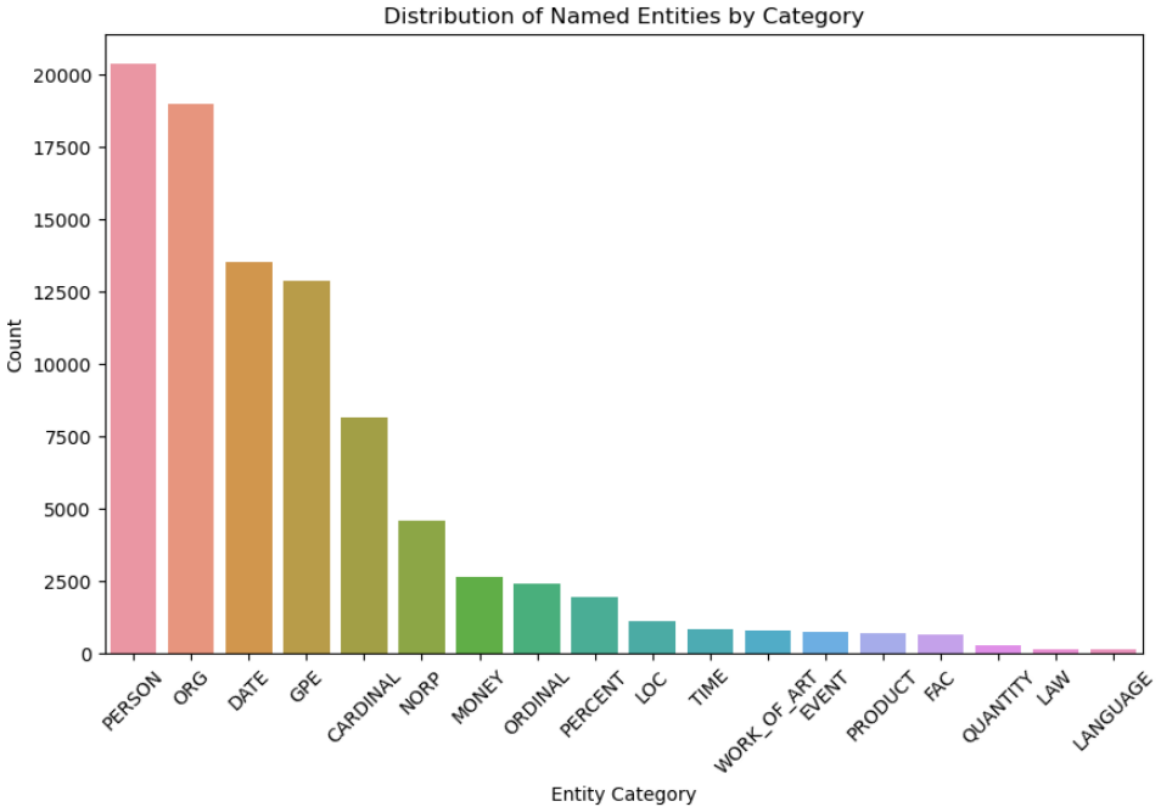
3. Topic Modeling Challenges

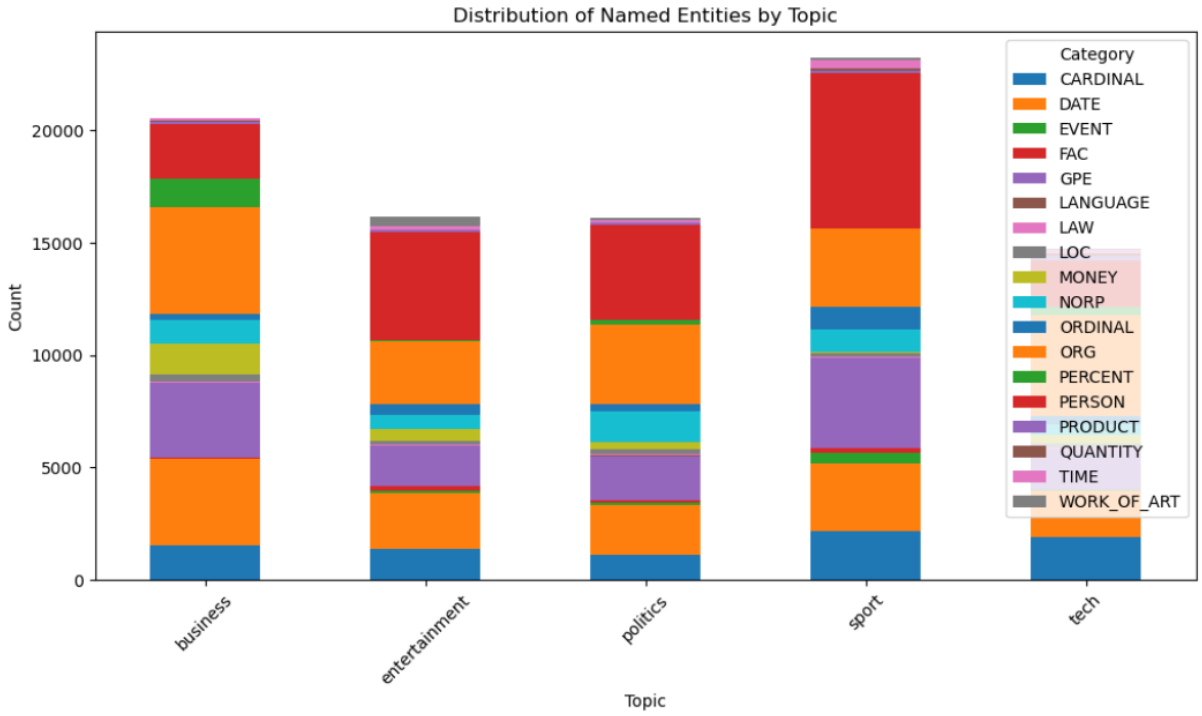
Choosing the Number of Topics: Determining the optimal number of topics (n_components in LDA) required experimentation.

Interpreting Topics: Some topics were not distinctly separable, and interpreting the meaning of each topic was not always straightforward.

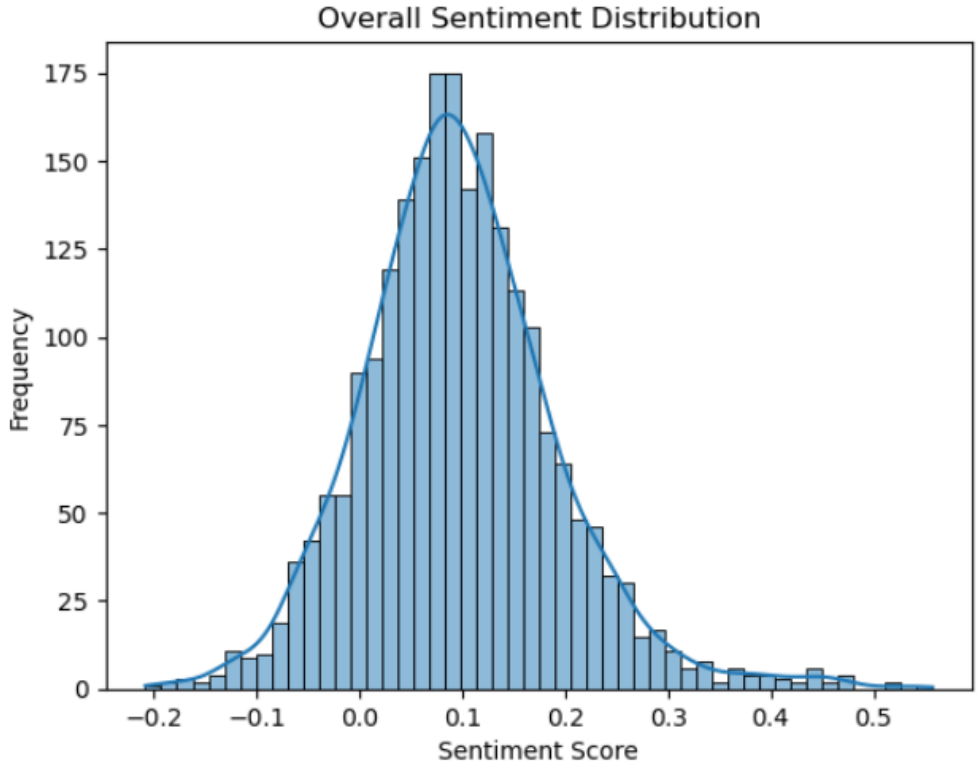
Findings

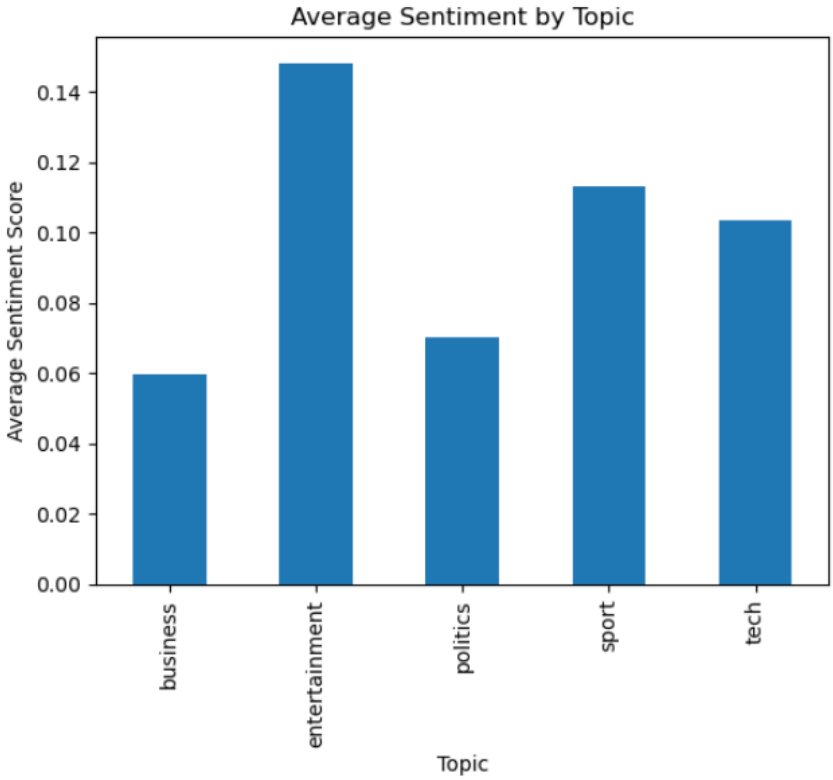
Named Entity Recognition



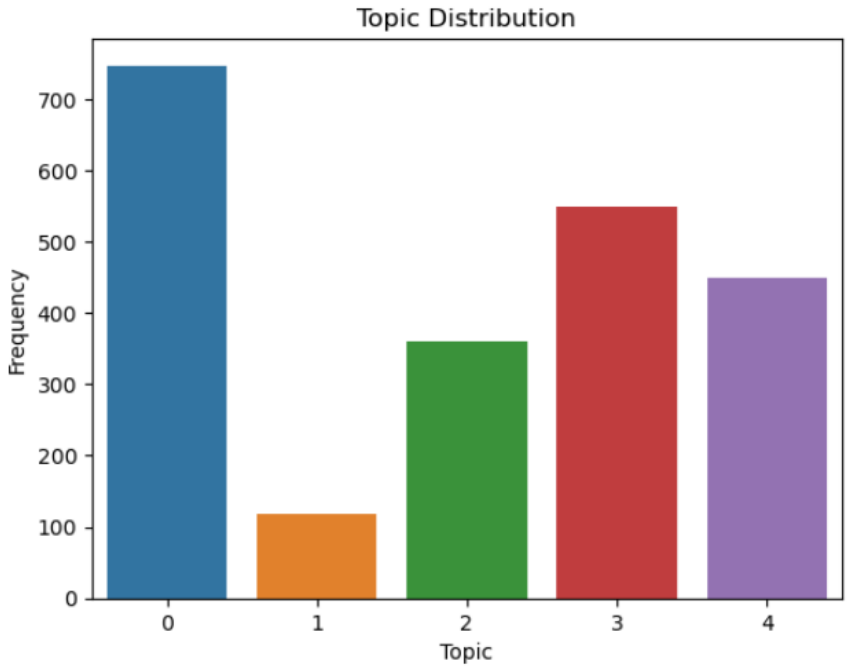


Sentiment Analysis





Topic Modeling



Note: topic modeling by LDA mostly corresponds to original category made by bbc