

Lab 2 Report - NLTK

Text Preprocessing and NLP Analysis

1. Objective

To apply fundamental Natural Language Processing (NLP) techniques on the constructed news corpus using the NLTK library and perform basic sentiment analysis and classification.

2. Dataset

The dataset used was `news_corpus.jsonl`, created in Lab 1 from the Fake and Real News dataset.

The corpus contains labeled news articles (0 = Fake, 1 = Real).

3. Methodology

The following NLP techniques were implemented:

a) Tokenization

- Word-level and sentence-level tokenization using NLTK.

```
Tokens: ['berlin', '(', 'reuters', ')', '-', 'germany', 's', 'social',
'democrats', '(', 'spd', ')', 'faced', 'pressure', 'on']
```

b) Stopword Removal

- Removed common English stopwords to reduce noise.

```
===== STOPWORD REMOVAL =====
Original token count: 712
After stopword removal: 391
First 20 filtered tokens: ['berlin', 'reuters', 'germany', 'social',
'democrats', 'spd', 'faced', 'pressure', 'wednesday', 'consider', 'offering',
'coalition', 'talks', 'chancellor', 'angela', 'merkel', 'conservatives',
'settle', 'worst', 'political']
```

c) Stemming

- Applied Porter Stemmer to reduce words to root form.

```
===== STEMMING =====
First 20 stemmed tokens: ['berlin', 'reuter', 'germani', 'social', 'democrat',
'spd', 'face', 'pressur', 'wednesday', 'consid', 'offer', 'coalit', 'talk',
'chancellor', 'angela', 'merkel', 'conserv', 'settl', 'worst', 'polit']
```

d) Lemmatization

- Used WordNet Lemmatizer to obtain meaningful base forms.

```
===== LEMMATIZATION =====
First 20 lemmatized tokens: ['berlin', 'reuters', 'germany', 'social',
'democrat', 'spd', 'faced', 'pressure', 'wednesday', 'consider', 'offering',
'coalition', 'talk', 'chancellor', 'angela', 'merkel', 'conservative',
'settle', 'worst', 'political']
```

===== STEM vs LEMMA COMPARISON =====

berlin		berlin		berlin
reuters		reuter		reuters
germany		germani		germany
social		social		social
democrats		democrat		democrat
spd		spd		spd
faced		face		faced
pressure		pressur		pressure
wednesday		wednesday		wednesday
consider		consid		consider

e) POS Tagging

- Assigned grammatical tags (noun, verb, adjective, etc.) to tokens.

```
===== POS TAGGING =====
[('berlin', 'NN'), ('(', '('), ('reuters', 'NNS'), (')', ')'), ('-', ':'),
('germany', 'NN'), ('s', 'VBP'), ('social', 'JJ'), ('democrats', 'NNS'), ('(', '('),
('spd', 'NN'), (')', ')'), ('faced', 'VBD'), ('pressure', 'NN'), ('on', 'IN'),
('wednesday', 'NN'), ('to', 'TO'), ('consider', 'VB'), ('offering', 'VBG'),
('coalition', 'NN')]
```

f) Named Entity Recognition (NER)

- Identified entities such as persons, locations, and organizations.

```
NER: (S
    berlin/NN
    (/(
        reuters/NNS
    )/)
    -/:
    germany/NN
    s/VBP
    social/JJ
    democrats/NNS
    (/(
        spd/NN
    )/
    faced/VBD
    pressure/NN
    on/IN
    wednesday/NN
    to/T0
    consider/VB
    offering/VBG
    coalition/NN)
```

g) Sentiment Analysis (VADER)

- Computed compound sentiment scores for each article.

Average sentiment: -0.02376073999999995

h) Naive Bayes Classification

- Built a probabilistic classifier using word frequency features.
 - Evaluated model performance using accuracy.
- i) Word Cloud Visualization**
- Generated word clouds to visualize frequent terms in the corpus.

4. Results

- Successfully implemented a complete NLP preprocessing pipeline.
- Sentiment polarity scores were generated for all articles.
- Naive Bayes classifier achieved reasonable performance.
- Word cloud visualizations highlighted common vocabulary patterns.



5. Conclusion

This lab demonstrated practical implementation of core NLP preprocessing techniques and basic text classification using NLTK. The processed data and extracted features provide a strong foundation for advanced machine learning models in the next lab.

6. Git Link

(Lab 2 Git Link:-)[https://github.com/MYTH-il/NLP/blob/main/lab2_nltk.ipynb]