

NLP Milestoen1

Sohaila Ibrahim
ID: 52-21225

Nada Elbehery
ID: 52-8973

March 6, 2025

Contents

1	Overview	2
2	Dataset Description	2
3	Methodology	2
4	Preprocessing Flow	2
4.1	Segmentation	3
4.2	Stemming	3
4.3	Part-of-Speech Tagging	3
4.4	Named Entity Recognition (NER)	3
4.5	Stopword Removal	4
4.6	Removing Non-Arabic Characters	4
4.7	Sample text	4
5	TF-IDF Weighting	5
6	Vector Space Model	5
7	LDA Topic Modeling	5
7.1	Processing Steps	6
8	Conclusion	6

1 Overview

In Milestone 1, we focused on conducting a thorough data analysis and review of the dataset. This process involved exploring the structure and characteristics of the data, identifying patterns, trends, and any potential inconsistencies or irregularities. Additionally, we applied preprocessing and cleaning techniques to transform the raw dataset into a structured and well-organized format suitable for machine learning (ML) tasks.

2 Dataset Description

Our dataset consists of multiple text files containing transcripts in the Egyptian dialect. These texts are primarily sourced from various episodes of *El Mokhbir El Iqtisadi* (The Economic Informant), a popular Arabic-language YouTube channel that discusses global economic and political issues. The dataset covers a diverse range of topics, including:

- International trade and financial crises.
- Energy policies and resource control.
- Geopolitical conflicts and their economic implications.
- Technological advancements and digital economy dynamics.
- The economic rise and fall of corporations and industries.
- Social and economic trends in different regions.

Some files discuss historical events and economic strategies of major powers such as the United States, China, and Russia, while others focus on inflation, supply chain disruptions, and corporate financial maneuvers.

3 Methodology

Natural Language Processing (NLP) for Arabic text involves several preprocessing steps to enhance text representation for further analysis. This report details the sequential flow of preprocessing techniques applied to Arabic text, followed by TF-IDF weighting, Vector Space Model representation, and LDA topic modeling.

4 Preprocessing Flow

Farasa is a library designed for Arabic Natural Language Processing (NLP). It provides various tools for tasks such as tokenization, part-of-speech tagging, named entity recognition, stemming, diacritization, and segmentation. The preprocessing pipeline consists of multiple steps that leverage Farasa's modules to prepare Arabic text for analysis.

4.1 Segmentation

Segmentation is the process of breaking down words into smaller meaningful units, such as prefixes, stems, and suffixes. The `FarasaSegmenter` module is responsible for segmenting Arabic words into their morphemes.

```
from farasa.segmenter import FarasaSegmenter

def segment(text):
    segmenter = FarasaSegmenter()
    segmented = segmenter.segment(text)
    return segmented
```

4.2 Stemming

The `FarasaStemmer` module reduces words to their root forms.

```
from farasa.stemmer import FarasaStemmer

def stem(text):
    stemmer = FarasaStemmer()
    stemmed = stemmer.stem(text)
    return stemmed
```

4.3 Part-of-Speech Tagging

The `FarasaPOSTagger` module assigns part-of-speech (POS) tags to words in Arabic text.

```
from farasa.pos import FarasaPOSTagger

def tag(text):
    tagger = FarasaPOSTagger()
    tagged = tagger.tag(text)
    return tagged
```

4.4 Named Entity Recognition (NER)

The `FarasaNamedEntityRecognizer` module detects named entities such as persons, organizations, and locations in Arabic text.

```
from farasa.ner import FarasaNamedEntityRecognizer

def recognize(text):
    named_entity_recognizer = FarasaNamedEntityRecognizer()
    named_entity_recognized = named_entity_recognizer.recognize(text)
    return named_entity_recognized
```

4.5 Stopword Removal

Eliminating frequently occurring words with little semantic meaning. Predefined Arabic stop words were collected from the library `arabicstopwords` and extended with additional stop words to handle the Egyptian dialect.

4.6 Removing Non-Arabic Characters

Filtering out non-Arabic symbols to clean the text.

4.7 Sample text

For example, given the input text:

إزايك عامل إيه أنا النهاردة كان عندي شغل كثير بس الحمد لله خلصت كل حاجة بعد كده رحت الكافيه شوية عشان أريح دماغى إنت عامل إيه فى جديد

- After preprocessing, the text is transformed into:
 إيزايك عامل أي أنا نهارد ي شغل كتير حمد ل الله خلص كل حاج كده رح كافي
 شوي أريح دماغ إن عامل أي جديد
- After tagging, the text is transformed into:

[illegible]

- After entity recognition, the text is transformed into:

C/إيه/O عامل/O إنت/O بماغي/O أربع/O عثمان/O شوية/O الكيفية/O راحت/O كده/O بعد/O حاجة/O كل/O خلصت/O لله/O الحمد/O يس/O بكتير/O شغل/O عندي/O كان/O الشهر ده/O إنا/O إيه/O عامل/O از/O إلك/O حديد/O

Another example, input text:

سافر محمد بن سلمان إلى الرياض يوم الإثنين لحضور اجتماع مع ممثلي شركة أرامكو. ناقش الاجتماع مستقبل الطاقة في المملكة العربية السعودية وتأثير الأسعار العالمية. كما التقى محمد بالرئيس التنفيذي لشركة سابك وأكد على أهمية الاستثمارات المستقبلية. في اليوم التالي، زار جامعة الملك سعود والتقى بالطلاب والأساتذة لمناقشة أحدث الأبحاث العلمية.

perform_lda(dic) handles the creation of the model. It takes as an input a dictionary of word counts per document, structured as:

$$\text{dic} = \{\text{file_name}_i : \{\text{word}_j : \text{count}_{ij}\}\} \quad (1)$$

where each document (file) contains words mapped to their frequency.

7.1 Processing Steps

1. Extracts unique words from each document:

$$\text{processed_docs} = [\text{list of words from each document}] \quad (2)$$

2. Creates a Gensim dictionary mapping words to unique IDs.
3. Converts documents into a Bag-of-Words (BoW) representation:

$$\text{corpus} = [(\text{word_id}, \text{frequency}) \text{ tuples}] \quad (3)$$

4. Trains an LDA model with:

- **num_topics** = 6 (number of topics)
- **passes** = 10 (iterations for better topic assignment)
- **random_state** = 46 (for reproducibility)

8 Conclusion

This NLP pipeline successfully processes Arabic text, extracts key information using TF-IDF and NER, and represents documents numerically for further analysis using the Vector Space Model and LDA topic modeling.