**Individual In-depth Report**

**Member name:** [Krupaben Kothadia](mailto:kkothadi@asu.edu)

**Evaluated by:** [Gautham Vijayaraj](mailto:gvijaya6@asu.edu)

**Date:** 09/17/2023

**Tasks Assigned:**

* This report is guided by the writing center's input.
* This report summarizes the Data Preprocessing and NLP techniques utilized for Arabic language tweets by referring to the paper “[Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages](https://drive.google.com/file/d/1AA1mL9x8Nbc39Pn1_JSlQLd4bKAZ32Vd/view?usp=drive_link)” [1]. This paper [1] produces a dataset that can accurately classify suspicious Arabic tweets on twitter.

**Summary:**

The intricate structure of Arabic sentences can make them tough to analyze. The authors collected Arabic tweets from Twitter using an API. Saudi Arabia has a high number of Twitter users [2]. This study found that 72% of tweets in the Arab region are in Arabic. Arabic language has different sentence rules and word meanings than English. So, techniques used for English don't work directly for Arabic.

In this study, the authors faced two big challenges. First, they had to figure out the meaning of Arabic tweets that didn't have diacritics. Second, they needed to do stemming and lemmatization. Diacritics are important in Arabic because they change word meanings [3]. But in social media, people don't often use diacritics, making it harder to classify tweets.

* Dataset: They received the data in JSON format and stored it in Excel with Arabic support. The dataset has 1555 tweets: 826 suspicious and 729 not suspicious.
* Data Preprocessing:
  + Data Filtration: They removed things like punctuation, retweets, URLs, and media to save time.
  + Data Tokenization: They split tweets into words and removed short, unimportant words like "and."
  + Stemming and Lemmatization: They simplified words to their base form to make classification easier.
  + Manual Labeling: They manually sorted tweets into two groups: suspicious (label 1) and not suspicious (label 0).

**Outcome:**

The authors overcame challenges in analyzing Arabic tweets by collecting data from Twitter, focusing on Saudi Arabia. They simplified and prepared the data, improving the accuracy of their classification system. This work enhances our understanding of complex Arabic sentence structures in the context of social media.

**References** *(with citation)*

[1] M. S. Al-Ghamdi and M. A. Khan, “Intelligent analysis of Arabic tweets for detection of suspicious messages,” *Arabian Journal for Science and Engineering*, vol. 45, no. 8, pp. 6021–6032, Mar. 2020, doi: 10.1007/s13369-020-04447-0.

[2] S. Bahkali, A. Almaiman, A. Bahkali, S. Almaiman, M. Househ, and K. Al-Surimi, “The role of social media in promoting women’s health education in Saudi Arabia.,” *PubMed*, vol. 213, pp. 259–62, Jan. 2015, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26153009

[3] S. Abed, M. H. Alshayeji, and S. Sultan, “Diacritics effect on Arabic speech recognition,” *Arabian Journal for Science and Engineering*, vol. 44, no. 11, pp. 9043–9056, Jul. 2019, doi: 10.1007/s13369-019-04024-0.

**Evaluation of Report**

**Evaluation summary with justification.**

This paper outlines the methods by analyzing the challenges in analyzing Arabic tweets by collecting data from Twitter. The above is done using data preprocessing techniques like the Data Filtration and Data Tokenization. The authors simplified and prepared the data, improving the accuracy of their classification system..

**The quality of the major result(s) with justification.**  
  
Results from this paper include a discrete study of the analysis of Arabic Tweets in twitter with the use of data preprocessing techniques and NLP techniques. This research paper was helpful as these techniques can be used to further classify the tweets into suspicious and non-suspicious tweets.

**The usefulness of the paper to the overall project.**   
  
This paper shares insights about challenges in analyzing Arabic tweets by collecting data from Twitter. The different types of data preprocessing techniques would be helpful to analyze and extract data from other social networking sites too.

**Other comments**

No Comments

**Evaluation Approval  
  
Evaluation by:** [Gautham Vijayaraj](mailto:gvijaya6@asu.edu) **Date:** 09/17/2023

**Is the written report of the in-depth study complete with all the major result(s) of the paper(s)? If not, provide as many examples of the major result(s) missing in the written report as possible. (in bullet form). [Normally within 100 words]**

* Yes, the report summarizes all the major results of the paper.
* The paper summarizes analyzing Arabic tweets in the twitter social media platform.
* The tweets containing different sentences are processed by different data preprocessing techniques which are mentioned in the in-depth report.

**Is each section of the guidelines sufficiently completed? If not, point out what is missing. [Normally within 40 words].**

Yes, each section of the guidelines is sufficiently completed.

**Is the quality of this version of the written report satisfactory? If not, then why not? [Normally within 40 words]**

Yes, the quality of the report is satisfactory.

**Approval.  
  
Approved by:** [Gautham Vijayaraj](mailto:gvijaya6@asu.edu) **Date:** 09/17/2023 **Is the quality of this written in-depth study report and Evaluation report satisfactory? If not, then why not? (limit: 40 words)**

Yes, the quality of the written report and evaluation report is satisfactory and concise.