GERMAN INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Nouran Khaled Department of Informatics and Computer Science

Teaching Assistant: Amal Yassin

Instructor: Alia El Bolock

Arabic CyberBullying Detection Using Arabic Sentiment Analysis Summary

Nouran Khaled

Due Date: September 17, 2022

1 Summary

For reference: [1] is the main paper, while [2] is the related paper to be summarized

Research Questions

- What are the challenges and how to resolve them for the processing of the Arabic language?
- Different preprocessing steps affect the classification accuracy of the model?
- Which data mining tool provides better results?
- Which data mining tools are more time efficient for model classification?

Both papers are trying to find the most efficient way to detect the CyberBullying in Arabic. So While [1] focused on using uses methods of Classification such as PMI, Chi-square and Entropy. [2] Focuses more on using SVM classification and comparing the performance of efficiency of WEKA using Light Stemmer, WEKA using ArabicStemmerKhoja and Python efficiencies and comparing all three. [2] is targeted towards mainly finding the most accurate method and not only the speed of detecting.

The Datasets used by [2] is obtained by random collection and data from Twitter by using tools like AraBully words. And this data then enters a preprocessing phase where Almoshatheb Alarabi and Microsoft® Excel were used to clean the data. Following the cleaning, the normalising technique is used on the various words in the uniform version. Additionally, it gets rid of the diacritical marks, punctuation, and lengthening. The procedure of removing stop words from our dataset follows the use of this tokenization based on white spaces. WEKA and Python were both employed as data mining tools, while for

Light Stemmer and ArabicStemmerKhoja were both used as stemmers.

The classification process then differs as [2] uses the different reprocessed data using different classification methods which are SVM classifier tool, WEKA, and Python. While in [1] the data were classified to bullying and non-bullying. It was classified by three people and use an odd number of people to be the last classification after the majority opinion. And after processing the data, They contrasted the PMI approach with the Chi-square and Entropy approaches.

Finally the results resulting from [2] showed Using Light Stemmer and ArabicStemmerKhoja, the WEKA does the tweet classification more accurately than Python, with a little difference in accuracy. However, it is also noted that Python takes less time than WEKA to construct the classification model.

References

[1] Bedoor Y AlHarbi, Mashael S AlHarbi, Nouf J AlZahrani, Meshaiel M Alsheail, Jowharah F Alshobaili, and Dina M Ibrahim. Automatic cyber bullying detection in arabic social media. *Int. J. Eng. Res. Technol*, 12(12):2330–2335, 2019.

[2] Samar Almutiry and Mohamed Abdel Fattah. Arabic cyberbullying detection using arabic sentiment analysis. *The Egyptian Journal of Language Engineering*, 8(1):39–50, 2021.