Project Proposal: Building a sentimental analysis model for Arabic dataset

Rabab Alkhalifa – raalkhalifa@iau.edu.sa

1. BACKGROUND

Sentiment analysis is a sub-field of natural language processing and employs machine learning, computational linguistics and data mining. The goal of sentiment analysis is to automatically detect the polarity of a text and try to systematically identify, extract, quantify, and study affective states and subjective information within text.

2. PROBLEM STATEMENT

In research will build upon existed Arabic dataset from the web which were designed for detecting emotions for text, the goal is to optimise existing models toward designing a robust computational approach for analysing and detecting emotions from Arabic dataset. Additionally, conducting research on the state of art techniques for Arabic sentimental analysis. The outcome from this research will contribute knowledge towards recognise, understand and detect accurate emotion from Arabic dataset called Arabic Sentiment Tweets Dataset (ASTD).

3. SOLUTION STATEMENT

This research project aims to explore the most accurate model to classify Arabic sentimental dataset. This will be look at through the evaluation for different proposed models to contribute in designing a robust model for evaluating emotions.

4. PROJECT DESIGN

In order to achieve a solution for the problem statement, the main objectives are defined as follows:

- Review the literature to evaluate various experiments and datasets that was used to extract sentiments from Arabic datasets and discuss the outcomes of experimentation by replicating selected experiments model with ASTD dataset.
- Create an enhanced sentimental analysis model that is an upgraded version from the existing models that proved to gain better insights from sentimental dataset. This include a model to compare a set of reviewed experimented models and enhanced model developed after observation.
- Statistically evaluate, analyse and interpret the results obtained to assess the accuracy of sentimental analysis models using Arabic dataset.

5. KEY RESEARCH QUESTIONS

The proposed research will aim to address three major questions:

- Which machine learning models/techniques are most suited to the task of textual emotion recognition in Arabic sentiments?
- How to prepare Arabic data for semantic analysis using stop words and would that increase the accuracy?

- What are the best tools to evaluate, analyse and interpret the results obtained to assess the accuracy of sentimental analysis models using Arabic dataset?

6. METHODOLOGY

6.1 BENCHMARK MODEL

In this research, we will use ASDA designed and proposed in [1] as the benchmark model. We aim to replicate the experiments proposed in [2], [3], [4] papers.

6.2 EVALUATION MTRICS

We will describe the performance measurements by using numerical scalars and ROC curve to easily describe the accuracy of the selected models.

6.2.1 SCALARS

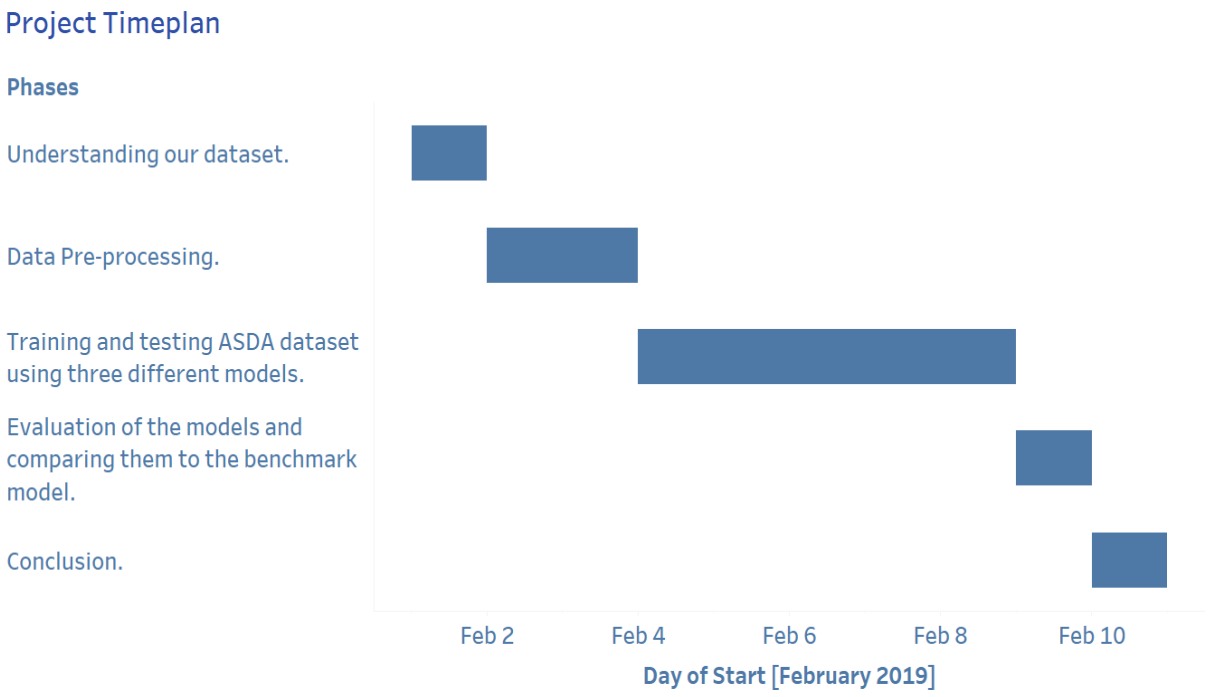
Contingency table will be constructed in the aim of measuring true positive, turn negative, false positive and false negative. Then perception, recall and F-measure will be accurately analysed to avoid selecting model that always predict negative.

6.2.2 ROC curve

ROC curve is a graphical illustration which is useful in comparing model results and efficient way to display the relationship between sensitivity and specificity.

7. TIMELINE

The project is divided into five phases. As follows,



8. EXPECTED OUTCOMES

The main outcome from this research project are:

1. Review the state of art techniques for Arabic sentimental analysis.
2. Develop an enhanced model for Arabic Sentimental Analysis.

9. Dataset References

1. Mahmoud Nabil, ***Mohamed Aly***, and Amir Atiya. **ASTD: Arabic Sentiment Tweets Dataset,** *EMNLP, 2015*. [pdf]
2. Arabic stopwords: http://www.ranks.nl/stopwords/arabic
3. Taghva, K., Elkoury, R., and Coombs, J. 2005. Arabic Stemming without a root dictionary. Information Science Research Institute. University of Nevada, Las Vegas, USA.

10. BIBILOGRAPHY

Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., & Abdullah, S. (2018). Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis. Journal of Information Science, 44(3), 345–362. https://doi.org/10.1177/0165551516683908

Alwakid, G., Osman, T., & Hughes-Roberts, T. (2017). Challenges in Sentiment Analysis for Arabic Social Networks. *Procedia Computer Science*, *117*, 89-100.

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. *Information Processing & Management*, *56*(2), 320-342.

11. References

[1] Nabil, M., Aly, M., & Atiya, A. (2015). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2515-2519).

[2] Mohammad, A. S., Al-Ayyoub, M., Al-Sarhan, H., & Jararweh, Y. (2015, December). Using aspect-based sentiment analysis to evaluate arabic news affect on readers. In *Utility and Cloud Computing (UCC), 2015 IEEE/ACM 8th International Conference on* (pp. 436-441). IEEE.

[3] Dahou, A., Xiong, S., Zhou, J., Haddoud, M. H., & Duan, P. (2016). Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2418-2427).

[4] Vo, D. T., & Zhang, Y. (2016). Don't Count, Predict! An Automatic Approach to Learning Sentiment Lexicons for Short Text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 219-224).