Sentiment Analysis for colloquial Arabic Language

Mohamed A.Rahim[1], Mohamed Nagy[1], Mohamed Sayed[1], Mohamed ElBahy[1], Moataz Lotfy[1], Hesham Salama[1], and Ensaf H. Mohamed[1,*]
[1]Faculty of Computers & Artificial Intelligence, Helwan University
Email: (Ensaf_hussein@fci.helwan.edu.eg)

1. Abstract

Social media [1] is a huge source of information; it is increasingly being used by governments, companies, and marketers to understand how the crowd thinks. Sentiment analysis is a research field that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes Arabic Sentiment Ontology (ASO) that contains different words that express feelings and how strongly these words express these feelings.
In this paper We proposed a new lexicon-based model for Arabic sentiment analysis with support of Vader Module, the accuracy of the model was **86.6%.**

2. Introduction

Given the importance of Sentiment Analysis [2], many research works have been devoted to this research area. However, most of these studies have focused on English and other Indo-European languages. Very few studies have addressed Sentiment Analysis in morphologically rich languages such as Arabic. Nevertheless, given the increasing number of Arabic internet users and the exponential growth of Arabic online content, Sentiment Analysis in this language has gained the attention of many researchers in the last decade.

The objective of this paper to produce sentiment analysis lexicon-based module using an existing English module named Vader with some changes according to Arabic rules. There are three levels of granularity namely document level, sentence level, and aspect level. We will use the sentence level.

3. Related Work

In this section, we will present research related to Arabic sentiment Analysis field with focus on dialectal Arabic study cases. Arabic language is characterized by a wide number of dialects varieties. Besides Modern Standard Arabic used as a formal language, different Arabic dialects are used for nearly all everyday speaking situations. By the emergence of social media and the various electronic networks, enabling Arab users to express their opinions using different Arabic dialects, researchers have raised the need to consider this amount of generated content especially by the study of the peculiarities related to written forms of these different dialects

3.1 Sentiment analysis approaches for MSA and Arabic dialects

- Abdul-Mageed and Diab [3] constructed a large-scale multi-lingual lexicon based on both MSA and colloquial Arabic (Egyptian and Levantine) for sentiment analysis, called SANA. SANA lexicon is a

combination of many lexicons, such as, SIFAAT, HUDA and an automatic collected corpus (with both statistical method and machine translation).

- Diab et al. [4] developed an electronic lexicon that can be used in different NLP tasks, sentiment analysis in our case. Their lexicon consists of three parts: MSA, dialectal Arabic and English. Authors made Tharwa publicly available which can be used mainly for the Egyptian dialect sentiment analysis
- Itani et al. [5] conducted a comparison between the lexicon-based and corpus-based approach by using both MSA and Arabic dialects. The experimental results show that lexicon-based approach (83.4% of accuracy) outperforms the corpus-based approach
- Hossam S. Ibrahim et al. [6] presented a feature-based sentence level sentiment analysis approach for Arabic language. They used a lexicon consisting of Arabic phrases to improve the polarity detection of Arabic sentences. Also, many linguistic features have been used, such as, Intensifiers, Shifters and negation. The developed lexicon focuses on both MSA and Egyptian dialectal Arabic. Experimental results showed that the proposed approach obtained 95% of accuracy using SVM classifier.

4. Data Collection

4.1 First dataset

1. This dataset about Egyptian Arabic and Modern Standard Arabic sentiment words without weights.
2. Gave each word weight that weight positive or negative number according to type of word and we made that with support of Vader lexicon file but that was not enough cause lexicon need a lot of words and that`s problem of that module type

4.2 Second dataset

The dataset contains Arabic positive and negative words with weight 1, -1 for positive and negative words respectively.

4.3 Third dataset (Vader Lexicons)

After that we think about something to make dynamic dataset our dataset is updatable. it update itself with support of Vader dataset when module cannot find word in our dataset he try to Synonyms of that word in English and search in Vader lexicons if he find it try to calculate average score for that word and save it in our data set so this calculation won't happen again for same word.

So, after all this work our module is supported by large number of sentiment word lexicons.

Table 1: the used dataset description.

Dataset	Lexicon Size	Language
Dataset 1	5061 words	Arabic
Dataset 2	4055 words	Arabic
Vader	7517 words	English

5. Module Implementation

5.1 Data Preprocessing

Texts go through some preprocessing methods such as spell checking, tokenization, stop words removal, punctuations removal.

5.2 Word Score

After prepare all of that now we ready to work on sentence then check if word in our Arabic lexicon dataset to get score if not , try to get score based on English meaning of the word and synonyms by searching in Vader Lexicons then get average score and save it in Arabic lexicons.

5.3 Detect Special words such as (جدا, كثيرا, كافي, احيانا, فعلا…etc.) this words increases or decreases the score of the previous main word.

5.4 Detect Negation and Conjunction

- **First case** gets all negation keywords and detecting negation of any word by going back two words in sentence before this word if we find any of negation keywords that mean this word will be negated.
- **Second case** if we don`t find negation keyword. we try to search about special words such as : و ، او this words if word before them in negation type, check word after them if it negated word so it's okay if not we negate it such as : الكتاب ليس جيد وممل so جيد will negated and ممل because it negative word so للغاية won't negated.
- **Third case** now searches about special words such as: لكن، بس, ولكن because these words Increases score of the words that after them and decrease score of the words that before them.

5.5 Calculate Score based on some equations to get positive, negative and neutral score and this score will be normalized based on the following rule

Normalized score = score/math.sqrt ((score*score) + alpha), alpha = 15. (1)

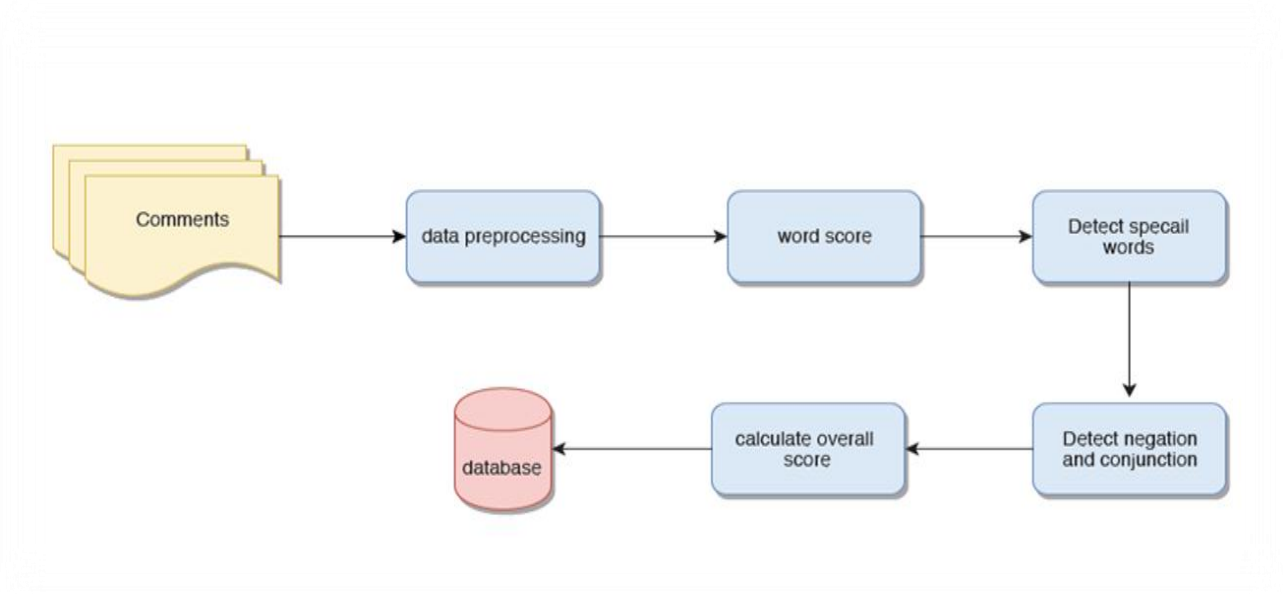

Figure 1: show the flow of sentiment module

6. Module Features

There is two type of features in our module first type **Vader features** that implemented for English language and converted into Arabic language and
Second type **extra features** that related to Arabic language and wasn't in Vader.

1- Vader Features

- Typical negations (e.g., "ليس جيد ").
- Use of contractions as negations (e.g., "لم يكن جيدا").
- Conventional use of punctuation to signal increased sentiment intensity (e.g., "جيد!!!!").
- Using **degree modifiers** to alter sentiment intensity (e.g., "جدا ").
- Understanding many sentiment-laden emoticons such as :) and :D.

2- Extra Features

- Blind Negation (e.g., "تحتاج للمزيد من العمل").
- Bidirectional flow to detect Negation and Conjunction and degree modifiers.
- Auto text correct (e.g., "جداااااااا" will converted to "جدا").

7. Results

After module finished the next step is testing phase so this work done on labeled Arabic book reviews dataset called **LABR**, we make three experiments each on 100 reviews we got accuracy for each we will take the average of them to get the overall accuracy the figure below shows the results accuracies.

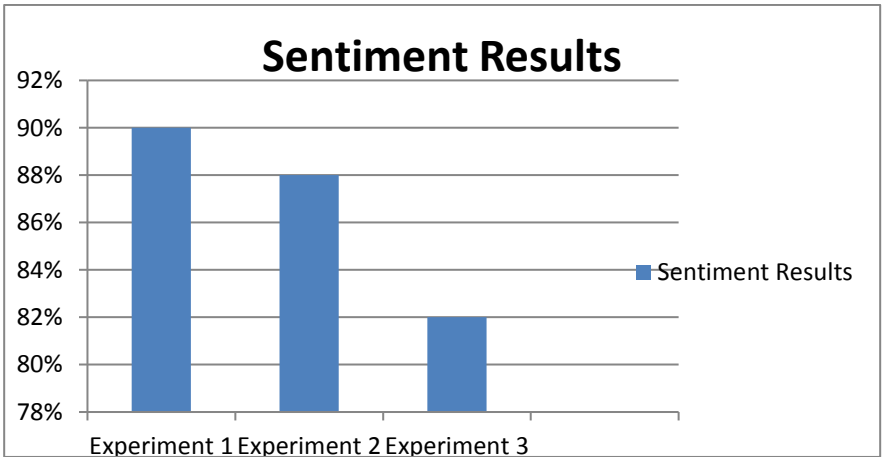

Figure 2: shows the sentiment module accuracy

8. Conclusion and Future work

We proposed in this paper a new lexicon-based approach for Arabic sentiment analysis with support of Vader Module, understanding

sentiment of a certain entity is crucial for decision makers to understand what their future actions need to be. Little work has been done so far to analyze sentiments we produced this module with accuracy is **86.6%.**

In the future we are planning to extend our work in the following directions.

- Extend lexicons to include more accuracy.
- Extend more in Arabic Rule and methods to include more accuracy.
- Enhance the implementation of our approach to report more accurate sentiments.
- Think about Hybrid sentiment module using Lexicon-based approach with machine learning in one module to get their benefits.

9. References

1- **Samir Tartir , Ibrahim Abdul-Nabi 2017** Semantic Sentiment Analysis in Arabic Social Media Journal of King Saud University - Computer and Information Sciences (29), (2), (4) (2017), pp. 229-233.

2- **Naaima Boudad , Rdouan Faizi , Rachid Oulad , Haj Thami Raddouane Chiheb** Sentiment analysis in Arabic: A review of the literature Ain Shams Engineering Journal **(21) (7)(2017).**

3- **Abdul-Mageed, M. and M.T. Diab. SANA:** A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis. in LREC. **2014.**

4- **Diab, M., et al. Tharwa:** A large scale dialectal arabic-standard arabic-english lexicon. in Proceedings of the Language Resources and Evaluation Conference (LREC). **2014.**

5- **Itani, M.M., et al.** Classifying sentiment in arabic social networks: Naïve search versus Naïve bayes. in Advances in Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on. **2012. IEEE.**

6- **Ibrahim, H.S., S.M. Abdou, M. Gheith,** Sentiment Analysis for Modern Standard Arabic And Colloquial. arXiv preprint arXiv:1505.03105, 2015.

7- **M'hamed Mataoui, Omar Zelmati, Madiba Boumechache** A Proposed Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic.

8- **Sanjeera Siddiqui , Azza Abdel Monem , and Khaled Shaalan** Sentiment Analysis in Arabic (2016).

9- **Synth. Lect. Hum. Lang. Technol** B. **Liu** Sentiment analysis and opinion mining., **5 (1) (2012), pp. 1-168.**

10- **Mahmoud Nabil, Mohamed Aly, Amir F. Atiya** LABR: A Large-Scale Arabic Sentiment Analysis Benchmark arXiv:1411.6718v2 [cs.CL] 3 May 2015.