Sentiment Analysis in Arabic

Sanjeera Siddiqui[1(✉)], Azza Abdel Monem[2], and Khaled Shaalan[1,3]

[1] British University in Dubai, Block 11, 1st and 2nd Floor,
Dubai International Academic City, Dubai, UAE
faizan.sanjeera@gmail.com, khaled.shaalan@buid.ac.ae
[2] Faculty of Computer and Information Sciences, Ain Shams University,
Abbassia, 11566 Cairo, Egypt
azza_monem@hotmail.com
[3] School of Informatics, University of Edinburgh, Edinburgh, UK

Abstract. The tasks that falls under the errands that takes after Natural Language Processing approaches includes Named Entity Recognition, Information Retrieval, Machine Translation, and so on. Wherein Sentiment Analysis utilizes Natural Language Processing as one of the way to locate the subjective content showing negative, positive or impartial (neutral) extremity (polarity). Due to the expanded utilization of online networking sites like Facebook, Instagram, Twitter, Sentiment Analysis has increased colossal statures. Examination of sentiments helps organizations, government and other association to extemporize their items and administration in view of the audits or remarks. This paper introduces an Innovative methodology that investigates the part of lexicalization for Arabic Sentiment examination. The system was put in place with two principles rules– "equivalent to" and "within the text" rules. The outcomes subsequently accomplished with these rules methodology gave 89.6 % accuracy when tried on baseline dataset, and 50.1 % exactness on OCA, the second dataset. A further examination shows 19.5 % in system1 increase in accuracy when compared with baseline dataset.

Keywords: Sentiment analysis · Opinion mining · Rule-based approach · Arabic natural language processing

1 Introduction

Web, additionally termed as World Wide Web, contains heaps of data. Web furnishes individuals with an open space to impart their insights or assumptions, their encounters, and their inclinations on a substance or product. The aim of Sentiment Analysis is to perceive the content with assessments and mastermind them in a way adjusting to the extremity (polarity), which incorporates: negative, positive or unbiased (neutral). Sentiments takes the organizations to tremendous statures [6, 7]. Dialects talked by individuals identifies with their way of life and what they talk, thus distinctive dialects are talked or learnt in better places, which contrast in components also in qualities. Arabic Natural dialect handling is moving a large portion of the scientist's outlook to Arabic, because of the expansion utilization of Arabic dialect by people and the expanded web Arabic clients. Arabic dialect holds one of the main ten position in the

E. Métais et al. (Eds.): NLDB 2016, LNCS 9612, pp. 409–414, 2016.
DOI: 10.1007/978-3-319-41754-7_41

overall utilized dialects. Arabic Natural dialect handling in sentiment investigation is taking gigantic consideration because of the inaccessibility of assets. This requires a need to develop the work in Arabic Sentiment Analysis.

The rest of this paper is organized as follows. Related work is covered in Sect. 2, Data collection is covered in Sect. 3 followed by system implementation in Sect. 4. Section 5 covers results and lastly Sect. 6 depicts conclusion.

2 Related Work

Shoukry and Refea [7] took after a corpus-based methodology, accomplished an accuracy of 72.6 %. Positive, negative and unbiased polarity characterization done by Abdullah et al. [1] displayed a dictionary and sentiment examination tool with an accuracy of 70.05 % on tweeter dataset and 63.75 % on Yahoo Maktoob dataset.

In order to take a shot at Sentiment Analysis, the key parameter is the dataset. Late endeavors by Shaalan [5] outlined the significance of crowdsourcing as an extremely effective system for clarifying dataset. SVM classifier accomplished 72.6 % accuracy on twitter dataset of 1000 tweets [2, 7].

Feldman [4] worked at record level opinion investigation utilizing a consolidated methodology comprising of a vocabulary and Machine Learning approach with K-Nearest Neighbors and Maximum Entropy on a blended area corpus including instruction, legislative issues and games achieved an F-measure of 80.29 %.

Aldayel and Azmi [2] utilizing a half and half approach that is Lexical and Support Vector Machine classifier created 84.01 % precision. El-Halees [3] accomplished 79.90 % precision with Hybrid methodology containing lexical, entropy and K-closest neighbor.

3 Data Collection

Collection of data is vital to perform sentiment analysis. In the field of sentiment analysis, the key underlying data used is the opinion or sentiment data for checking the polarity and lexicon for building the rules or for machine learning classifiers. Lexicons used in this paper is an extension of lexicon shared by [1]. These lexicons contains named entities, adjectives, randomly paced and most importantly words which appeared to be common in both positive and negative reviews. New addition to the lexicon created in this paper includes one word tweet or review in the appropriate lexicon list.

Hence we add Abdullah et al.'s [1] dictionaries with the expansion of words from the dataset which were found to seem more than once. Taking into account the attentiveness of reiteration of these words and their situation, they were incorporated into both the rundown that is sure and negative records. For instance, "كاتب" (Writer) was rehashed in both negative and positive surveys. Henceforth, it was incorporated into positive and negative dictionary.

4 Implementation of Arabic Sentiment Analysis

The spreadsheet guideline (rule) based framework proposed in this section included three key stages to show the outcomes in the wake of handling. The principal stage is to enter the tweet. The second stage incorporates rules centered check wherein the entered subjective statement is gone through an arrangement of standards rules which includes "equal to" and "within the text rules", examining for the tweet extremity either negative or positive. The third stage is to transform the content into "Red" color demonstrating content is positive or "Green" showing content is negative. Figure 1 delineates the review of the framework proposed in this paper.

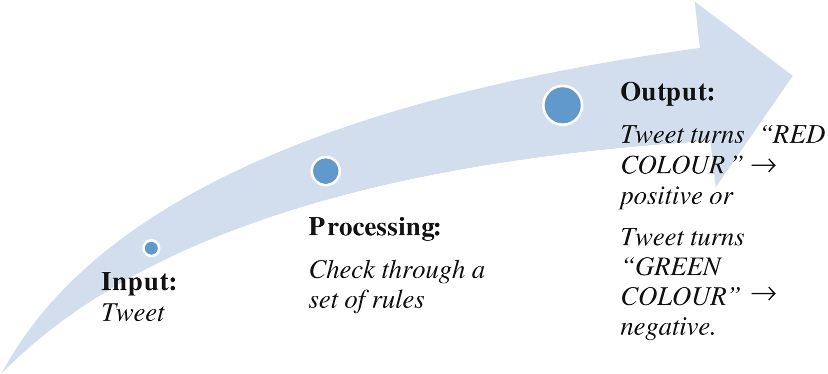
Input:
Tweet
Processing:
Check through a set of rules
Output:
Tweet turns "RED COLOUR" → positive or
Tweet turns "GREEN COLOUR" → negative.

Fig. 1. Overview of the system proposed

The methodology followed in this paper is guideline (rule) based, we allude (Shaalan 2010) for an audit about the importance of the principle based methodology and how it is utilized to apply distinctive NLP assignments or tasks. Parcel of examination till date included pre-handling or pre-processing as the real steps. In this exposition, with the very utilization of fitting standards the pre-preparing step was completely wiped out for the dataset, there were no changes done to the dataset. The tenet (rule) based methodology incorporate examples to gaze the entered tweet in the dictionary which in this case is the lexicon.

Methodology for the Formation of Rules. The very presence of the strategy that we have proposed in this paper, has enlivened a completely new methods, in order to address the basic and undealt issues with the lexicon based methodology. Both the extremity (polarity) dictionaries and extremity (polarity) opinions are scrutinized. This paper only focusses on positive and negative polarity examination.

The methodology in this research covers many crucial areas found to be missing in the literature review that we conducted. Firstly, the rules are not confined to search within text. Secondly, rules are not excluding common words found in both positive and negative lexicon and lastly, the proposed rule-based system covers all the directions where the word has been found to have a huge impact.

The rules incorporates two key types one rule checks for the word in the sentence, we framed it as “within the text” and for only one word tweet phrased as “equivalent to the text”. The key hidden base ground components which offered us some assistance with formulating fitting standards incorporates examination of the tweets and the augmentation of positive and negative dictionaries. The examination of tweets brought about distinguishing relations relating to words which were either disjoint or coincided.

The words which were disjoint that is totally showing either positive or negative extremity were incorporated into their separate vocabularies. The words which coincided that is the ones which were observed to be normal in both negative and positive reviews were incorporated into positive and also negative dictionaries.

The system introduced in this paper encompasses two sorts of key principles which were composed taking into account “equivalent to” or “within” the text passages. For instance, if the entered tweet contains “أصدق” (I trust), which is recorded in the positive assumption vocabulary (lexicon), then the extremity is positive. All in all, if the tweet content contains the words or is equivalent to a word from the positive rundown then the content transformed into “red” showed the tweet is positive. On the off chance that the tweet content “contain words” or “is equivalent to” from the negative rundown, then the content transformed into “Green” showing the tweet is negative.

Be that as it may, the real turnover was in the tenets (the rules) subsequently made. The words observed to be basic in light of the examination of the tweets were incorporated into both positive and negative dictionaries which were legitimized amid the rules creation stage. One key expansion to the vocabulary utilized as a part of this exposition was the expansion of words in the rundown which were single words that is the tweet which included one and only word. In view of the extremity the words were consequently set in the fitting rundown.

5 Results

The outcomes incorporate the examination of the considerable number of investigations led in this paper. In order to do the examination the accuracy of the considerable number of investigations are utilized. The system was tried on [1] and OCA dataset. Table 1 delineates the trials results with respect to System proposed in this paper.

Table 1. Results of applying system on datasets

Datasets/Accuracy	System tested on [1] dataset	System tested on OCA dataset
Precision	87.4	50.4
Recall	93.3	97.6
Accuracy	89.6	50.1

Table 1 Unmistakably follows the outperformance of standards made in System with enormous accuracy for [1] when contrasted with the outcomes on OCA dataset. The system proposed in this paper gave to be fruitful 39.5 % more accuracy than [1] than OCA dataset. The outcome variety in both the datasets requires an extension to the current rules proposed in this paper, which could effectively enhance the exactness.

This inquiry is exceptionally very much replied with a basic correlation of our system with guideline (rule) fastening approach results with Abdullah et al.'s [1] lexicon based methodology results. As the dataset utilized as a part of this paper is taken from Abdullah et al. [1], we have looked at our outcome with this dataset. Our proposed framework (rule based) beat their outcomes with 17.35 % increment in exactness when tried on System 1.

On contrasting our technique and Abdullah et al. [1] presented some key increments. Abdullah et al. [1] concentrated on augmenting the dictionaries which was only expansion of new words to the rundown even irrelevant to the test set, has abandoned them with no change in accuracy. Consequently, the analyses outflanked when contrasted with the outcomes reported in [1]. Table 2 delineates the correlation of the tests directed in this paper with [1].

This section exhibited the system developed. System incorporated the rules which secured "within the text" and "equivalent to" standards.

Table 2. Results correlation with [1]

Rule-based vs lexicon-based approaches	Accuracy
System proposed	89.6 %
Abdullah et al. [1]	70.05 %

6 Conclusion

This paper delineated how Sentiment Analysis has discovered its presence with the extremely propelled developments in online data. The key fundamental parameter seen is the sharing of audits on any setting and how this effect the clients to take choices on numerous things right from purchasing a motion picture ticket to purchasing a property to numerous propelled operations.

Revealing insight into how a word in one assessment represents positive extremity and how the same word could bring about to make the sentence negative when utilized as a part of an alternate setting. Sentiment Analysis is observed to be exceptionally helpful in measuring the effect of an item or administration, through the surveys or reviews that the general population have shared on it. This paper beats the vocabulary (lexicon) building process through the proper arrangement of words too not barring the basic words found in both the tweets for the dictionaries. Sentiment Analysis, through the organized set guidelines (rules) and through the right utilization of various rules including "contains content" and "equivalent to".

The reasonable noteworthiness in results in a manner acquired through the standards made makes the rules based methodology the most alluring methodology. The greater part of the scientists have concentrated just on Lexicon based yield. Our yield is new to the sentiment investigation period. The yield was exhibited through the adjustment in shade of entered tweet to "Red" for positive tweets and "Green" for negative tweets with the use of two rules, which obtained 19.5 % increase in results when compared to Abdullah et al. [1] but resulted in only 50 % accuracy for OCA dataset. This calls for a need to extend this rule based system to be improvised and enhanced to be fitted into the context of any dataset. In spite of the fact that this framework (system) was just cantered around positive and negative tweets.

As a key errand for further creating and upgrading this proposed framework, we would be anticipating enhance the current proposed framework to cover more rules, subjectivity order and in this way show impartial(neutral) extremity also.

References

1. Abdulla, N.A., Ahmed, N.A., Shehab, M.A., Al-Ayyoub, M., Al-Kabi, M.N., Al-rifai, S.: Towards improving the lexicon-based approach for arabic sentiment analysis. Int. J. Inf. Technol. Web Eng. (IJITWE) **9**(3), 55–71 (2014)
2. Aldayel, H.K., Azmi, A.M.: Arabic tweets sentiment analysis – a hybrid scheme. J. Inf. Sci. 0165551515610513 (2015)
3. El-Halees, A.: Arabic opinion mining using combined classification approach (2011)
4. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (2013)
5. Shaalan, K.: A survey of arabic named entity recognition and classification. Comput. Linguist. **40**(2), 469–510 (2014). MIT Press, USA
6. Shoukry, A., Rafea, A.: Preprocessing Egyptian dialect tweets for sentiment mining. In: 4th Workshop on Computational Approaches to Arabic Script-Based Languages, pp. 47–56 (2012)
7. Shoukry, A., Rafea, A.: Sentence-level Arabic sentiment analysis. In: 2012 International Conference on Collaboration Technologies and Systems (CTS), pp. 546–550. IEEE, May 2012