This is a post-peer-review, pre-copyedit version of a paper published in Hussain A, Zhao H, Ren J, Zheng J, Liu C, Luo B & Zhao X (eds.) *Advances in Brain Inspired Cognitive Systems*. Lecture Notes in Computer Science, 10989. BICS 2018: 9th International Conference on Brain Inspired Cognitive Systems, Xi'an, China, 07.07.2018-08.07.2018. Cham, Switzerland: Springer International Publishing, pp. 589-596. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-00563-4_57

A Semi-Supervised Corpus Annotation for Saudi Sentiment Analysis using Twitter

Abdulrahman Alqarafi[1,2], Ahsan Adeel[1], Ahmed Hawalah[2], Kevin Swingler[1], and Amir Hussain[1]

[1] CogBID Lab, Dept. of Computing Science and Mathematics
University of Stirling, FK9 4LA Stirling, UK
[2] University of Taibah, Madina, Saudi Arabia

Abstract. In the literature, limited work has been conducted to develop sentiment resources for Saudi dialect. The lack of resources such as dialectical lexicons and corpora are some of the major bottlenecks to the successful development of Arabic sentiment analysis models. In this paper, a semi-supervised approach is presented to construct an annotated sentiment corpus for Saudi dialect using Twitter. The presented approach is primarily based on a list of lexicons built by using word embedding techniques such as word2vec. A huge corpus extracted from twitter is annotated and manually reviewed to exclude incorrect annotated tweets which is publicly available. For corpus validation, state-of-the-art classification algorithms (such as Logistic Regression, Support Vector Machine, and Naive Bayes) are applied and evaluated. Simulation results demonstrate that the Naive Bayes algorithm outperformed all other approaches and achieved accuracy up to 91%.

Keywords: Sentiment analysis, Saudi dialect, Word embedding

Introduction

Sentiment analysis has gained a lot more research attention due to the emergence of social media. The principal goal of sentiment analysis is to classify text as positive, negative or neutral [10]. Sentiment analysis for Arabic language possesses different challenges compared to other languages such as dealing with Modern Standard Language (MSA) and Dialects that significantly varies from one region to another [18, 5]. Sentiment analysis is based on three main approaches: 1) Supervised approach. 2) Unsupervised approach and 3) Hybrid approach [13]. The supervised approach is based on a set of annotated messages (that is usually constructed manually [15, 3]). The unsupervised approach is based on sentiment lexicon (which is often built automatically by exploiting English dictionaries such as Wordnet or Sentimentwordnet,etc) [8, 6]. The hybrid approach is a combination of these two [14, 19]. However, most of the existing related research works have concentrated more on dialects such as Egyptian, Levantine , Jordanian etc. In addition, the constructed resources are not publicly available. In this work a publicly available lexicons are developed based on word embedding

for Saudi dialect (Dialect which suffers from a subsequent lack of works and studies). In addition, the developed lexicons are evaluated on a semi-supervised constructed corpus. The developed corpus is collected from Twitter, automatically annotated, and reviewed by a native Saudi dialect speaker. The resulted corpus contains 4000 messages (2000 positive and 2000 negative sentiments)[1].

The paper is organized as follows: Section 2 presents the developed novel corpus. Simulation results are presented in Section 4. Finally, Section 5 concludes this work with some future directions.

Related works

In this paper, two kind of resources are considered: 1)sentiment lexicons and 2) An annotated sentiment corpus. This section is divided into two parts, the first part presents related work on sentiment lexicon construction, whereas the second part presents the works on building and annotating corpus.

Sentiment lexicon construction

Approaches for building lexicons include, manual approach, dictionary-based approach, and corpus-based approach. Building lexicons manually is time consuming and requires more time and resource. In the dictionary-based approach, number of seeds words are collected manually and then utilized the synonym and antonym of the list using common dictionary such as WordNet. In the corpus-based approach, a seed list is used to extract the similar words from the corpus. To build lexicons for Arabic language, a number of techniques have been utilized in the literature. However, most of them have focused on Modern Standard Arabic (MSA) and other dialects such as Egyptian and Levantine, while only few researches studied building lexicons for Saudi dialect. El-Beltagy and Ali [11] proposed a lexicon-based method that learns the weights of the lexicon words from a large corpus of tweets. They reported the achieved accuracy up to 70%. However, it was based on Egyptian dialect. Abdul-Majeed etl [1] built a large scale multi-genre multi dialect Arabic sentiment lexicon and contains only two dialects Egyptian and Levantine. Furthermore, a large-scale Arabic Sentiment Lexicon (ArSenL) was proposed by Badaro et al. [9]. The authors constructed ArSenLis using a combination of English SentiWordnet (ESWN), Arabic WordNet, and the Arabic Morphological Analyzer (AraMorph). However, it only covers MSA. Eskander elt in[12] followed the same approach in ArSenLis and built a Sentiment Lexicon for Standard Arabic SLSA. In the proposed approach, the authors linked an Arabic morphological analyzer Aramorph lexicon with SentiWordNet and included MSA only. For Saudi dialect, there are some proposed lexicons such as Adayel and Azmi [4] for Saudi dialect lexicon. This lexicon contains only around 1500 terms. AraSenTi [3] developed an Arabic sentiment lexicons based on tweets called AraSenTi-Trans and AraSenTi-PMI. Assiri [7]

[1]Please contact aaq@cs.stir.ac.uk or ahu@stir.ac.uk to access the dataset

built a large lexicon that contains 14,000 sentiment terms based on a pre-created lexicon developed by Badaro et al. [9] and encoded using the Buckwalter translation. It is to be noted that most of the aforementioned approaches focused on other dialects and ignored Saudi dialect. In addition, lexicons were built without using word-embedding.

Annotated corpus construction

The supervised approach depends essentially on the existence of annotated data. Most of the existing approaches adopt manual annotation; hence, develop a restricted corpus which lacks good generalization. Foe example, OCA corpus contains 500 Arabic comments (250 positive and 250 negative), manually preprocessed, then segmented, and root extracted with a tool dedicated to Arabic [17]. AWATIF is a multi-genre corpus containing 10 723 sentences in Arabic manually annotated in objective and subjective polarity and then annotated the subjective sentences in positive, negative or neutral [2]. ASTD used 10,000 Arab Tweets, annotated in objective and subjective polarities and mixed with annotators of Amazon Mechanical Turk [15]. AraSenTi-Tweet used 17 573 Saudi tweets, manually annotated into four classes (positive, negative, neutral and mixed) [3]. In contrast to aforementioned work, in this work word embedding has been used to extend a lexicon (composed from a set of seeds that we constructed manually). The sentiment corpus is constructed in a semi-supervised manner depending on the constructed lexicon. The validation of this corpus is performed using three different classifiers. The corpus is available online.

Data collection

Building Lexicon

We manually built the sentiment lexicon based on deep learning word embedding technique for Arabic language developed by [20]. It contains approximately 3000 words (1500 positive and 1500 negative). A group of seeds words were collected manually and annotated by an expert in the language. We searched for the similar words in the dictionary. Each word in the lexicon is assigned a similarity value from the word embedding. Any value below 60% similarity is excluded. To make this enrichment, we utilized AraVec which is a pre-trained Arabic word embedding model . It is trained using word2vec and includes data from Tweets, World Wide Web pages and Wikipedia articles. The total number of utlized tokens are approximately around 3 billion. The lexicon contains some words which were written incorrectly but they are very common. In dialect, people sometimes do not follow the writing standards or rules and use slang. This is more clear in negations. MSA contains some negations which could change the meaning from positive to negative and vice versa. However, users sometimes link the negation with the words. The existing Arabic stemmers find it difficult to deal with dialects. For example (مو حلو) they combine them into (موحلو) which could be considered as different words.

words	Translation	Negations	Translation
أحب	I love	لا أحب	I hate
حلو	Beautiful	مو حلو	Ugly
يعجبني	I like	ما يعجبني	Don't like
مزبوط	Good	مش مزبوط	Not good

Table 1. Example of negations

Building Corpus

Saudi people are one of the largest Twitter users. In Twitter, people express their opinions in few words and short sentences due to the restrictions in number of letters. Consequently, people try to create new ways to overcome this challenge. For example, they combine some phrase together such as "ما يعجبني" contains negation "ما" and word "يعجبني" which means like to become "مايعجبني" without space. This is challenging because most of the available Arabic NLP tools fail in stemming the dialects. In order to build the corpus, the approach presented in [16] is followed. In this case we collected approximately 15000 tweets and classified them into positive and negative polarities. The adopted approach first separated the tweets into two classes: positive and negative based on a list of strong emoticons such as heart, devil, etc. Figure 1 illustrates the different steps of building the corpus.

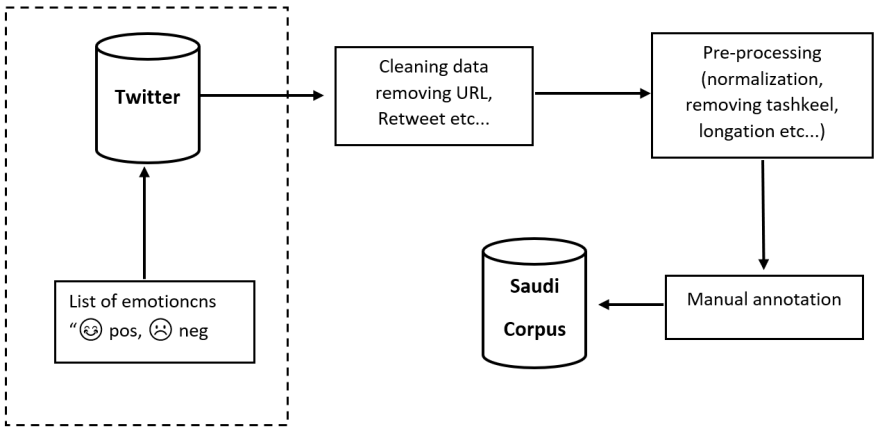

Fig. 1. Constructing twitter corpus framework

The preprocessing steps include (1) Normalization: Where we removed any punctuations in the corpus that includes 'Tashkeel'. (2) Remove mentions such

as names, retweets etc. Based on the constructed lexicon, we extracted tweets containing sentiments words. (3) Remove longation, in order to delete repeated letters. However, the suppression begin with more than two repeated letters (because we could find words with two repeated letters like "ممتنع" and "ممتاز").Finally, the collected data were annotated manually into two different classes positive and negative. For annotation, the annotators followed some guidelines for annotating the tweets such as,

- it should hold an opinion (حقيقة قرار مزبوط) "A really great decision"
- the speaker expresses emotion (انا مغرم باجهزة ابل) "I am in love with Apple devices"
- does not include news (وقد أفاد المصدر عن وقوع حادث شنيع في في طريق العقبة) "According to the reporter, a terrible accident took place in Alaqaba road" etc.

The resulted dataset contains 4000 tweets classified as 2000 positive and 2000 negative. Table 2. presents few examples from our dataset. In addition, Table 3 shows some statistics regarding the corpus.

Tweets	Annotation
انا عن نفسي اشوف تركيا من **افضل** الاماكن للسياحة "I would say Turkey is one of the best places to travel"	Positive
دعاية **خرافية** عن الايفون بعد ٣٠ سنة "A beautiful advertisement about iPhone after 30 years"	Positive
عشاء **مطنوخ** من أبو ريان الليلة "A delicious dinner from Abu Rayan tonight"	Positive
الفندق **خايس** وكله صراصير ولا خدمة ولا شي "The hotel is disgusting and there is no service as well"	Negative
مع نفسك يا رجال هذا **متغطرس ومهايطي** ولا يشوف احد "This man is arrogant and hypocritical"	Negative
فاشلين هم وبرنامجهم اللي مايشتغل ابد **كرهوني** اطلب **هالزلابة** "These are failures and their program is not working"	Negative

Table 2. Positive and negative tweets examples

Collected Tweets	35000 Tweets	
Using emoticons list	15000	
Cleaning data	8000	
Annotating data	6000	
Total	4000	
	2000 Positive	2000 Negative

Table 3. Some statistics regarding the corpus

Experiments

Experimental environment

For the experiment part, we divided our corpus into training and testing using 5 cross validation and Stratified K-Folds cross-validator. In addition, Term frequency-Inverse document frequency (TF-IDF) is used for feature extraction. For lexicon, only words having more that 60% similarity with initial our seed words are used.

Classification

Four different classifiers (Logistic-Regression, Support Vector Machine (SVM), Stochastic Gradient Descent Classifier (SGD), and Naive Bayes) were applied to the annotated corpus. For evaluation, f1 Score, precision and recall performance are used. The tf-idf for the annotated corpus are calculated and classified. Table 4 shows the achieved result of the first experiment. In the second experiment, constructed lexicon are considered and improved results are presented in Table 5.

Classifier	F1 Score	Precision	Recall
Logistic-Regression	0.88	0.83	**0.89**
SGD	0.87	0.86	0.88
SVM	0.74	0.61	0.87
Naïve Bayes	**0.89**	**0.90**	0.88

Table 4. The accuracy result before adding the lexicon features

It can be seen that Naive Bayes classifier achieved the best accuracy in both f1 score and precision 89% and 90% respectively, where Logistics Regression achieved the highest score in recall with 89%.

Classifier	F1 Score	Precision	Recall
Logistic-Regression	0.89	0.87	**0.90**
SGD	0.88	0.89	0.87
SVM	0.83	0.78	0.89
Naïve Bayes	**0.90**	**0.91**	0.89

Table 5. The overall accuracy

Adding lexicon features in the second experiment showed a slight improve in the accuracy. Naive Bayes achieved the best accuracy with an increase of 1% in f1-score and precision respectively where Logistic Regression had around 1 % increase in recall and achieving 90% .

Conclusion

In this paper, a semi-supervised approach is presented to construct an annotated sentiment corpus from Saudi tweets. In addition, word embedding techniques such as word2vec are exploited to build a lexicon and annotate the corpus. A set of experiments based on different classification algorithms are conducted. Simulation results demonstrate that the Naive Bayes classifier achieved the best precision of 91%, revealing the benefits of using word embedding technique for building the lexicon. There are still few challenges to be addressed such as processing for different dialects and limited resources (e.g. PoS tagging, stemming etc.). In addition, there are some linguistic features which require more investigation such as negations. Hence, the future work includes the investigation of linguistic features and development of algorithms to better deal with these features.

References

1. Abdul-Mageed, M., Diab, M.: Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). European Language Resources Association (ELRA) (2014), http://www.aclweb.org/anthology/L14-1702
2. Abdul-Mageed, M., Diab, M.T.: Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In: LREC. pp. 3907–3914. Citeseer (2012)
3. Al-Twairesh, N., Al-Khalifa, H.S., Al-Salman, A.S.: Arasenti: Large-scale twitter-specific arabic sentiment lexicons. In: ACL (2016)
4. Aldayel, H.K., Azmi, A.M.: Arabic tweets sentiment analysis – a hybrid scheme. Journal of Information Science **42**(6), 782–797 (2016)

5. Alqarafi, A.S., Adeel, A., Gogate, M., Dashitpour, K., Hussain, A., Durrani, T.: Toward's arabic multi-modal sentiment analysis. In: Liang, Q., Mu, J., Jia, M., Wang, W., Feng, X., Zhang, B. (eds.) Communications, Signal Processing, and Systems. pp. 2378–2386. Springer Singapore, Singapore (2019)
6. Altrabsheh, N., El-Masri, M., Mansour, H.: Combining sentiment lexicons of arabic terms (2017)
7. Assiri, A., Emam, A., Al-Dossari, H.: Towards enhancement of a lexicon-based approach for saudi dialect sentiment analysis. Journal of Information Science **44**(2), 184–202 (2018). https://doi.org/10.1177/0165551516688143
8. Badaro, G., Baly, R., Hajj, H., Habash, N., El-Hajj, W.: A large scale arabic sentiment lexicon for arabic opinion mining. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). pp. 165–173 (2014)
9. Badaro, G., Baly, R., Hajj, H.M., Habash, N., El-Hajj, W.: A large scale arabic sentiment lexicon for arabic opinion mining. In: ANLP@EMNLP (2014)
10. Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A.Y.A., Gelbukh, A., Zhou, Q.: Multilingual sentiment analysis: State of the art and independent comparison of techniques. Cognitive Computation **8**(4), 757–771 (Aug 2016). https://doi.org/10.1007/s12559-016-9415-7, https://doi.org/10.1007/s12559-016-9415-7
11. El-Beltagy, S.R., Ali, A.: Open issues in the sentiment analysis of arabic social media: A case study. 2013 9th International Conference on Innovations in Information Technology (IIT) pp. 215–220 (2013)
12. Eskander, R., Rambow, O.: Slsa: A sentiment lexicon for standard arabic. In: EMNLP (2015)
13. Guellil, I., Boukhalfa, K.: Social big data mining: A survey focused on opinion mining and sentiments analysis. In: Programming and Systems (ISPS), 2015 12th International Symposium on. pp. 1–10. IEEE (2015)
14. Khalifa, K., Omar, N.: A hybrid method using lexicon-based approach and naive bayes classifier for arabic opinion question answering. Journal of Computer Science **10**(10), 1961 (2014)
15. Nabil, M., Aly, M., Atiya, A.: Astd: Arabic sentiment tweets dataset. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2515–2519 (2015)
16. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
17. Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., Perea-Ortega, J.M.: Oca: Opinion corpus for arabic. Journal of the Association for Information Science and Technology **62**(10), 2045–2054 (2011)
18. Sadat, F., Kazemi, F., Farzindar, A.: Automatic identification of arabic dialects in social media. In: Proceedings of the first international workshop on Social media retrieval and analysis. pp. 35–40. ACM (2014)
19. Shoukry, A., Rafea, A.: A hybrid approach for sentiment classification of egyptian dialect tweets. In: Arabic Computational Linguistics (ACLing), 2015 First International Conference on. pp. 78–85. IEEE (2015)
20. Soliman, A.B., Eissa, K., El-Beltagy, S.R.: Aravec: A set of arabic word embedding models for use in arabic nlp. Procedia Computer Science **117**, 256–265 (2017)