Using Tweets and Emojis to build an Arabic Dataset for Sentiment Analysis (TEAD)

Houssem Abdellaoui[1] and Mounir Zrigui[2]

[1] hsm.abdellaoui@gmail.com,
Research Laboratory LaTICE
B. P. : 56, Bab Menara, 1008 Tunis, TUNISIA
http://www.latice.rnu.tn
[2] mounir.zrigui@fsm.rnu.tn
Université de Monastir, Faculté des Sciences de Monastir

Abstract. Our paper presents a distant supervision algorithm for automatically collecting and labeling '**TEAD**' ,a dataset for Arabic Sentiment Analysis (SA), using emojis and sentiment lexicons. The data was gathered from Twitter during the period between the 1^{st} of June and the 30^{th} of November 2017. Although the idea of using emojis to collect and label training data for SA, is not novel, but getting this approach to work for Arabic dialect was very challenging. We ended up with more than 6 million tweets labeled as Positive or Negative. We present the algorithm used to deal with mixed-content tweets (Modern Standard Arabic MSA and Dialect Arabic DA). We also provide properties and statistics of the dataset alongside experiments results. Our tryouts covered a wide range of standard classifiers proved to be efficient for sentiment classification problem.

Keywords: Sentiment Analysis, Opinion mining, Modern Standard Arabic, Arabic Dialect, Sentiment Dataset, Emojis, Sentiment Lexicon

1 Introduction

Sentiment analysis(SA) is the process of determining the sentiment or the opinion of a text. Obviously, we, as human beings, are good at this. We can look at a given text and immediately know what sentiment it holds (positive or negative). Companies and academic researchers across the world are trying to make machines able to do that. It is super useful for gaining insight into consumer's opinions. Once you understand how your customers feel, after checking out their comments or reviews, you can identify what they like and what they don't, and build things for them such as, recommendation systems or more targeted marketing companies. The same logic can be applied on other fields for instance: economy, business intelligence, politics, sports, education and so on.
It all stated in the 20^{th} century with Pang and Lee [25] and Turney [30]. However, we still can trace some much earlier work related to SA such the research of Jaime [2] in 1979 that tackled the problem of subjectivity understanding

and Ellen Spertus [23] who proposed a paper on automatic recognition of hostile messages in 1997.

Nowadays, SA is still gaining large attention. As shown in Figure 1, the trend of SA did not stop increasing since 2004. This is due to many facts.

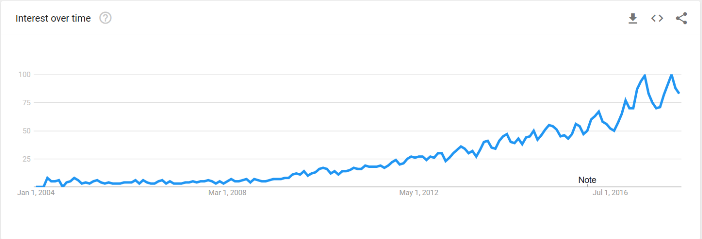

Fig. 1. Interest to SA from google trends (2004 - 2017).

First, the evolution of Natural Language Processing (NLP) is making a huge step towards more understanding the language computation and reasoning.

Second, we have now much more computational power easily accessible than what we used to.

And last but not least, the abundance data on the web 2.0 especially on social networks like Twitter, Facebook , Instagram, etc.

The work on SA is based on two main aspects.

The first one focuses on creating algorithms and techniques (machine learning, lexicon based and linguistic).

The second one, is when researchers are trying to build linguistic resources such as datasets and lexicons for SA.

The work that we provide in this paper , follows the second aspect . So we are going to present the process that has been done to obtain a dataset for Arabic SA. We will discuss our approach to collecting and labeling the dataset using emojis and sentiment lexicon. Also, we will highlight the problem of Arabic Dialect and how we managed to deal with it. Then, we will give details and statistics about the final **TEAD** dataset.

And finally, we will conclude with benchmark experiments and comparison with ASTD[27].

2 Related work

The main goal of SA is detecting the polarity of a review. But it should be preceded with identifying the subjectivity to make sure that the expressed view is opinionated. For the polarity classification task, many datasets were suggested in literature.

OCA[15] is one of the first sentiment dataset for Arabic language. It was manually collected from Arabic movies reviews. It contains 500 instances divided

Table 1. Arabic Sentiment Datasets.

Authors	Dataset	Size	Source
[16]	LABR	63257	Goodreads
[17]	AWATIF	2855	Wikipedia, Forums
[20]	HAAD	1513	LABR
[15]	OCA	500	Arabic movies reviews
[21]	AraSenti-Tweets	17573	Twitter
[27]	ASTD	10006	Twitter
[22]	SemEval	8366	Twitter
[19]	–	6894	Twitter

into 250 positive and 250 negative. It served as benchmark for many studies. In the same aspect **LABR** was proposed by Aly et.al[16] as the largest corpus for SA in that time. It holds more than 60 K review on books. The authors used a scale from one to five to rate them. Scale 1 and 2 for positive, 3 for neutral, 4 and 5 for negative. In 2012 Abdulmageed et al.[17] came up with **AWATIF** , a multi-genre corpus gathered from Wikipedia talk pages, web forms and Penn Arabic Tree Bank. **AWATIF** is not released online for free reuse or test. ElSahar and El-Beltagy [18] collected a multidomain Arabic review dataset. The scope of the reviews included hotels, movies, poducts and restaurants.
The role of social media is a key factor in the world where each part (corporations, brands, political figures, etc.) tries to have the most influence on users. The reasons behind this wave are simple. The first one, social media provides a huge amount of data easaly accessible from users from all arround the world. Second, this contents is always there ready to be used freely. We just need to know how to mine it. Twitter is a micro-blogging website where users can share and send short text messages called tweets, limited to 140[3] characters. It is thriving on the throne of social media with more than 6000 tweets per second. For Arabic language; many researches provide SA dataset collected from Twitter. Rafaee et al.[19] proposed a corpus for subjectivity analysis and SA. It comprises 6894 tweets (833 positive, 1848 negative, 3685 neutral and 528 mixed).Nabil et al. [27] used an automatic approach to construct their sentiment dataset. They called it **ASTD**; it consists of 10006 Arabic tweets divided into four classes (positive 793, negative 1684, mixed 832 and neutral 6691). Al-samadi [20] filtered **LABR** and selected 113 review. The selected ones were labeled for aspect-based SA. The annotation was made according to the **SemEval**2014-task4 guidelines. In 2017, Nora [21] proposed **AraSenti-Tweets** [21] dataset of Saudi dialect with 17573 tweets manually labeled to four classes (positive negative neutral and mixed). Also, in the same year the International Workshop on Semantic Evaluation proposed a new corpus for SA [22]. The data was gathered automatically

[3] The length of a tweet was expanded to 280 character starting from 26 Novembre, 2017

from Twitter and manually labeled. The dataset was provided to SemEval participants to accomplish 5 tasks.

- *Subtask A.*: Message Polarity Classification: Given a message, classify whether the message is of positive, negative, or neutral sentiment.
- *Subtasks B-C.* : Topic-Based Message Polarity Classification: Given a message and a topic, classify the message on B) two-point scale: positive or negative sentiment towards that topic C) five-point scale: sentiment conveyed by that tweet towards the topic on a five-point scale.
- *Subtasks D-E.* : Tweet quantification: Given a set of tweets about a given topic, estimate the distribution of the tweets across D) two-point scale: the ”Positive” and ”Negative” classes E) five-point scale: the five classes of a five-point scale.

3 The need for data and the use of emoji

3.1 Why do we need more data ?

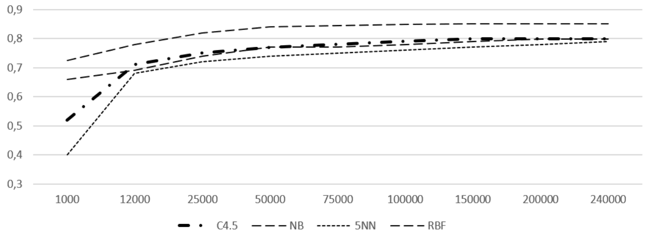

Fig. 2. The Effect of Dataset Size on Training Tweet Sentiment Classifiers (RBF: Radial basis function network,5NN: 5k-nearest neighbors, NB: Naive Bayes, C4.5 (algorithm for decision trees) for English language [26].

How many instances do we need to train a sentiment classifier?
The answer is not quite simple! No one can tell! This is an intractable problem that should be discovered through empirical investigation. The size of data required depends on many factors such as the complexity of the problem and the complexity of the learning algorithm. For examples, if a linear algorithm achieves good performance with hundreds of examples per class, we may need thousands of examples per class for a nonlinear algorithm, like random forest or deep neural networks.
Some studies tackled this problem like Pursa[26] and Kharde[6]. They concluded that dataset size significantly impacts classification performance like shown in figure 2. These two studies were carried out on English language. Arabic is more complex compared to Latin languages due to its agglutinate nature. Each word consists from combination of prefixes, stems and suffixes that results in very

complex morphology [1]. In fact, SA task for Arabic needs much more data. Meanwhile, literature review shows that freely available SA dataset for it are quite limited in size and number. In the reminder of this paper, we will present our tryouts to fill this gap by presenting **TEAD** an Arabic Sentiment Dataset collected from Twitter.

3.2 From emoticons to emoji to sentiment analysis

Emoticon is a stenography from facial expression. It eases the expression of feeling, mood and emotion. It enhances written messages with some nonverbal elements that attracts the attention of the reader and improves the overall understanding of the message. On the 19th of September 1982, Prof. Scott Fahlman of Carnegie Mellon University, proposed the first emoticons. He used ":-)" to distinguish jokes posts and ":-(" for serious ones. Since that, the use of emoticons had spread and new ones were created to express hugs, winks, kisses, etc.[28].
An emoji (Picture character in Japanese) is a step further. It appeared in Japan on the late 20th century. It is used on modern communications technologies. It facilitates the expression of emotions, sentiments, moods and even activities. As a new ideogram, it represents more than facial expressions, but also ideas, concepts, activities, building cars, animals, etc.
Several studies analyzed the use and effect of emojis on social networks like Twitter. They showed that tweets, with emojis included, are more likely to express emotions [7][4][5]. Some other researches created an emojis lexicon for SA [3].

4 Data collection and pre-processing

4.1 Collecting data from Twitter

The process of gathering data for the training task was performed during the period between the 1st of June and the 30th of November 2017. Using Twitter API and an online server from OVH[4], we were able to collect thousands of tweets each day. We followed these steps:

- Select the top 20 most used emojis on Twitter according to emjoitracker[5] on the 31st of May 2017 .
- Use Sentiment Emoji Ranking [3] to choose the ones that are the most subjective (we ended up with 10 emojis presented on Table 2).
- Use Twitter Stream API V1.1[6] for tweets live streaming with 3 filters:
  - Language = Ar (for Arabic letters)
  - Contains = Filtred list of emojis
  - Retweeted = No (to eliminate retweeted tweets)

[4] `https://www.ovh.com/`

[5] `http://emojitracker.com/`

[6] `https://blog.twitter.com/developer/en_us/a/2012/current-status-api-v1-1.html`

Table 2. List of the 10 most used Emojis on Twitter.

Unified id	Emoji	Sentiment	Description
1F602	😂	Positive	FACE WITH TEARS OF JOY
2764	❤	Positive	HEAVY BLACK HEART
1F60D	😍	Positive	SMILING FACE WITH HEART-SHAPED EYES
267B	♻	Positive	BLACK UNIVERSAL RECYCLING SYMBOL
2665	♥	Positive	BLACK HEART SUIT
1F62D	😭	Negative	LOUDLY CRYING FACE
1F60A	😊	Positive	SMILING FACE WITH SMILING EYES
1F612	😒	Negative	UNAMUSED FACE
1F629	😩	Negative	WEARY FACE
1F614	😔	Negative	PENSIVE FACE

The process yields to a dataset of **6 million Arabic tweets** with a vocabulary of 602721 distinct entities.

4.2 Arabic script for non Arabic languages

The Arabic script are not only used for writing Arabic, they are used in several other languages of Asia and Africa, such as Persian, Urdu, Azerbaijani and others. Unfortunately, the Twetter stream-api is not able to detect whether the language of a tweet is Arabic or not. It was interesting to find how many tweets in other language are there in our dataset. Sadly, we could not automate this process. We randomly extracted 2000 tweets and manually filtered them to find just one non-Arabic tweet. By this rate we can assume that the exitance of such type of noise in our dataset **TEAD** is rare.

4.3 Translation from Arabic Dialect to MSA

Modern Standard Arabic (MSA), which is the official language of the Arab world, is not used as frequently as Arabic dialects in Web. Indeed, it is more used in newspaper articles, TV news, education or on official occasions, such as conferences and seminars. On social network like Twitter the use of dialect is very common[14]. However, since all Arabic dialects have been using the same character set, and additionally plenty of the terminology are shared among diverse varieties, it is not a minor matter to differentiate and discrete the dialects from each other. Although some studies [13] proposed machine learning methods to do that. We use the Twitter API to locate the origin of the tweet using geographic localization system. We divide the dataset on six groups. This partition is proposed by Sadat[13]. The results are in figure 3 .

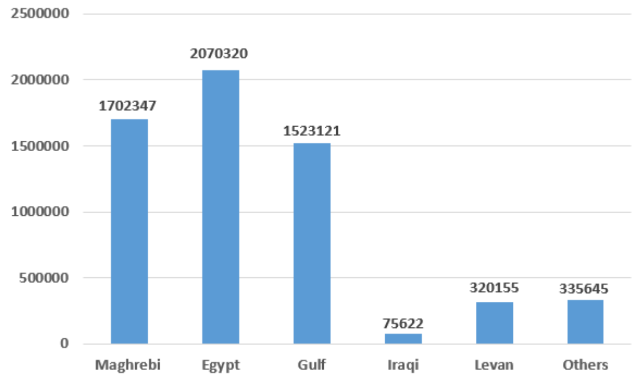

Fig. 3. Tweets repartition by dialect

Table 3. Lexicons used for translation of the Arabic dialect

Source	Dialect
[9]	Egypt
[10]	Levan
[11][14]	Maghrebi
[12]	Gulf

We used a simple and intuitive algorithm yet effective to replace dialect words with their respective synonyms in the MSA. The used dialect lexicons are presented in Table 3 . Unfortunately, we were not able to find any lexicon for Iraqui dialect so we were forced to omit all the Iraqui tweets from our dataset. We also deleted the tweets from the class ’Others’.

4.4 Preprocessing

Preprocessing is an essential step in almost any NLP tasks. It aims to eliminate the incomplete, noisy and inconsistent data. We followed this steps:

- **Removing URLs:** Tweets can contains links , so we need to remove them because they don’t contribute to sentiment classification.
- **Removing usernames:** Usersnames (@user) are also removed from the tweets.
- **Remove duplicated letters:** We replaced any letter that appears consecutively more than two times in a word by one letter . For example the word جمييييييلٌ *ǧmyyyyyylun* becames جميلٌ *ǧmylun*(beautiful) .
- **Remove punctuation and non Arabic symbols:** we also removed punctuation and others symbols that can be found in some tweets.
- **Tokenization and normalization:** we used stanford segmentor to preform the tokenization and normalization of the tweets.

4.5 Lexicon based approach for the dataset annotation

Facing the magnitude of the collected data, human labeling becomes expensive and takes a lot of time. We had to automate the process so we can keep up with the stream of tweets coming each minute. Our method to label the data is grounded on a lexicon-based approach for S A. We used the algorithm proposed in figure**??**. It is worth mentioning that this algorithm can handle negation.

We used Ar-SeLn [29] the first publicly available large scale Standard Arabic sentiment lexicon. We added the list of emojis used for gathering the tweets to the lexicons with their respective polarity according to the Sentiment Emoji Polarity Lexicon of Novak [3]. The result of the automatic annotation process is in the Table 4.

4.6 Manual validation of the automatic annotation

To validate our automatic approach of the data annotation, we randomly extracted 1000 tweet from each class.We performed a manual labeling on these portions of data by 2 native Arabic speaking annotators. The classification error rate was satisfactory as shown in table 5. The highest value was on the neutral set (11%). This is due to the complexity of capturing the actual subjectivity of a tweet when the number of positive token is equal to the negative ones.

5 Evaluation and results

5.1 Technical details

The training process aims to reveal hidden dependencies and patterns in the data that will be analyzed. Therefore, the training and test data set must be a representative sample of the target data. We conducted a set of benchmark experiments on **TEAD** and ASTD.Both datasets were randomly partitioned into training (70%) and test (30%). We used TF-IdF (token frequency inverse document frequency) and CBOW (continuous bag of words) as word representation features fol classical ML algorithms. For, deep learning models, we used Word2vec [24] for word embedding. We trained Word2vec(Skip-gram) with optimal parameters (vector size= 300, min-count = 5, window =3). Experiments were coded in Python3.6 using Scikt-Learn[7] and Keras[8] with Google Tensorflow[9] as backend. We used a machine with AMD FX 6-Core (3.5 GHz) and 16 GB of RAM. Tensorflow used CUDANN v5.1 with Geforce GTX 940.

5.2 Experimental Results and discussion

From the experimental results we can make the following observations:

[7] www.scikit-learn.org

[8] www.keras.io

[9] www.tensorflow.org

Algorithm 1: Sentiment Annotation Algorithm

```
Input : List of positive tokens from Ars-SeLn $Lp$
Input : List of negative tokens from Ars-SeLn $Ln$
Input : List of all the tweets $TEAD$
Input : List of all Arabic Negation words $NegList$
Output: List of positive tweets PosTweets
Output: List of negative tweets NegTweets
Output: List of objective tweets ObjTweets
begin
    foreach $tweet\ tj\ of\ TEAD$ do
        $SumPos/SumNeg$ accumulate the polarity of positive/negative tokens ;
        $SumPos \leftarrow SumNeg \leftarrow 0$;
        foreach $word\ ti\ in\ tj$ do
            if $ti\ in\ Lp\ and\ ti - 1\ not\ in\ NegList$ then
                $SumPos \leftarrow SumPos + 1$;
            else if $ti\ in\ Lp\ and\ ti - 1\ in\ NegList$ then
                $SumNeg \leftarrow SumNeg + 1$
            else
                /* The word has no positive sentiment score in
                   Ars-SeLn                                        */
            end
            if $ti\ in\ Ln\ and\ ti - 1\ not\ in\ NegList$ then
                $SumNeg \leftarrow SumNeg + 1$;
            else if $ti\ in\ Ln\ and\ ti - 1\ in\ NegList$ then
                $SumPos \leftarrow SumPos + 1$
            else
                /* The word has no negative sentiment score in
                   Ars-SeLn                                        */
            end
        end
        if $SumPos > SumNeg$ then
            PosTweets $\longleftarrow$ PosTweets $+\{tj\}$;
        end
        if $SumNeg > SumPos$ then
            NegTweets $\longleftarrow$ NegTweets $+\{tj\}$;
        end
        if $SumNeg == SumPos$ then
            ObjTweets $\longleftarrow$ ObjTweets $+\{tj\}$;
        end
    end
    return PosTweets, NegTweets,ObjTweets
end
```

Table 4. TEAD dataset statistics

	Number of tweets	Average tokens per tweet	Max tokens per tweet
Positive tweets	3,122,615	9,42	45
Negative tweets	2,115,325	9,25	34
Neutral tweets	378,003	11,36	39

Table 5. Manual validation of the automatic annotation

	Number of tweets	Classification error rate
Positive tweets	1000	5,6%
Negative tweets	1000	4.2%
Neutral tweets	1000	11,3%

- The hypothesis that we based our work on is tweets with emojis are more likely to be subjective. As a matter of fact, the results of annotation algorithm in table 4 confirm the assumption. The tweets labeled as objective were much less than the subjective ones.
- The results of the classification task using traditional ML algorithms on our dataset (**TEAD**) outperformed the ones obtained using ASTD dataset.
- We observe interesting patterns of correlation between training dataset size and learning results.
- SVM had the best experiment results and confirmed the previous work [27] assumption which is the suggested choice for SA.
- We used LSTM and CNN as deep learning (DL) classifiers. The less convenient results on ASTD proved that DL models needs a huge amount of training data to achieve better results.
- LSTM trained on **TEAD** shows encouraging results and open the doors to further investigation for the use of such a model in Arabic SA task.

Table 6. Classification Experimental Results Using TF-Idf as Text feature extraction (SVM:Support vector machine , LR: Logistic regression, M-NB :Multinomial naive Bayes , B-NB : Bernoulli naive Bayes ,DT: Decision tree, RF: Random Forest).

	SVM		LR		M-NB		B-NB		DT		RF	
	ASTD	TEAD	ASTD	TEAD	ASTD	TEAD	ASTD	TEAD	ASTD	TEAD	ASTD	TEAD
Precision	76 %	81 %	76%	77 %	72%	76%	81%	65%	78%	65%	84%	84%
Recall	75 %	83%	74 %	72%	72%	82%	74%	83 %	73%	73%	73%	69 %
F1-score	75,50 %	81,99 %	74,99%	74,42 %	74,42 %	76,68%	74,99%	81,99%	68,77%	75,42%	68,77%	75,76 %

From the experimental results we can make the following observations:

- The hypothesis that we based our work on is tweets with emojis are more likely to be subjective. As a matter of fact, the results of annotation algorithm in table 4 confirm the assumption. The tweets labeled as objective were much less than the subjective ones.

Table 7. Classification Experimental Results Using CBOW as Text feature extraction (SVM:Support vector machine , LR: Logistic regression, M-NB :Multinomial naive Bayes , B-NB : Bernoulli naive Bayes ,DT: Decision tree, RF: Random Forest).

	SVM		LR		M-NB		B-NB		DT		RF	
	ASTD	TEAD	ASTD	TEAD	ASTD	TEAD	ASTD	TEAD	ASTD	TEAD	ASTD	TEAD
Precision	70 %	**88 %**	71%	75 %	73%	70%	86%	66%	73%	66%	85%	85%
Recall	79 %	82%	80%	81%	72%	83%	79%	82 %	75%	70%	75%	47 %
F1-score	74,22 %	84,89 %	75,23%	83,89 %	73,46 %	77,67%	74,22%	83,95%	70,21%	71,46%	70,21%	60,53 %

Table 8. Exerimental results precision using deep learning models.

Dataset	**LSTM**	**CNN**
ASTD	81%	79%
TEAD	**87.5%**	86%

- In the classification experiment, **TEAD** outperformed ASTD on all the classical ML algorithms.
- We observe interesting patterns of correlation between training dataset size and learning results.
- SVM had the best experiment results and confirmed the previous work [27] assumption which is the suggested choice for SA.
- We used LSTM and CNN as deep learning (DL) classifiers. The less convenient results on the ASTD dataset proved that DL models needs a huge amount of training data to achieve better results.
- LSTM trained on the **TEAD** dataset shows encouraging results and open the doors to further investigation for the use of such a model in Arabic SA task.

6 Conclusion

In this paper we presented **TEAD** a large-scale Arabic tweets dataset. We provided details about the data collected. We used an emojis lexicon as key words for data gathering and tried to overcame the problem of using dialect instead of MSA. Some of benchmark experiments were established to compare **TEAD** to ASTD. Our dataset achieved a state of art performance with both calssical ML and deep learning classifiers. It outperformed existing literature results. In future work we intend to:

- Increase the size of the dataset.
- Try to find a better approach to deal with Arabic dialect.
- Build a specific deep learning model for Arabic SA and train it on **TEAD**.

Acknowledgments. The **TEAD** dataset will be freely available online for academic and research purposes.

References

1. Samir A., Mohamed E., Marwa M., and Aly F.. 2010 *Integrated machine learning techniques for arabic named entity recognition. IJCSI*, 7:27–36.
2. Carbonell, J. G. (1979). Subjective understanding: Computer models of belief systems. Technical report, YALE UNIV NEW HAVEN CONN DEPT OF COMPUTER SCIENCE.
3. Novak, P. K., Smailović, J., Sluban, B., and Mozetič, I. (2015). Sentiment of emojis. *PloS one*, 10(12):e0144296.
4. Schuff, H., Barnes, J., Mohme, J., Padó, S., and Klinger, R. (2017). Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.
5. Dhaoui, C., Dhaoui, C., Webster, C. M., Webster, C. M., Tan, L. P., and Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*, 34(6):480–488.
6. Kharde, V., Sonawane, P., et al. (2016). Sentiment analysis of twitter data: A survey of techniques. *arXiv preprint arXiv:1601.06971*.
7. Kramer, A. D. (2012). The spread of emotion via facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 767–770. ACM.
8. Mikolov, Tomas and Le, Quoc V and Sutskever, Ilya 2013 *arXiv preprint arXiv:1309.4168*
9. Diab, M. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014). Tharwa: A large scale dialectal arabic-standard arabic-english lexicon. In *LREC*, pages 3782–3789.
10. Graff, D. and Maamouri, M. (2012). Developing lmf-xml bilingual dictionaries for colloquial arabic dialects. In *LREC*, pages 269–274.
11. Harrat, S., Meftouh, K., Abbas, M., and Smaili, K. (2014). Building resources for algerian arabic dialects. In *Fifteenth Annual Conference of the International Speech Communication Association.*
12. Assiri, A., Emam, A., and Al-Dossari, H. (2017). Towards enhancement of a lexicon-based approach for saudi dialect sentiment analysis. *Journal of Information Science*, page 0165551516688143.
13. Sadat, F., Kazemi, F., and Farzindar, A. (2014). Automatic identification of arabic dialects in social media. In *Proceedings of the first international workshop on Social media retrieval and analysis*, pages 35–40. ACM.
14. Mohamed Ali Sghaier ,Mounir Zrigui 2017. *Tunisian Dialect-Modern Standard Arabic Bilingual Lexicon.* 14th ACS/IEEE International Conference on Computer Systems and Applications Hammamet Tunisia.
15. Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., and Perea-Ortega, J. M. (2011). Oca: Opinion corpus for arabic. *Journal of the Association for Information Science and Technology*, 62(10):2045–2054.
16. Aly, M. A. and Atiya, A. F. (2013). Labr: A large scale arabic book reviews dataset. In *ACL (2)*, pages 494–498.
17. Abdul-Mageed, M. and Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914.
18. ElSahar, H. and El-Beltagy, S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. In *CICLing (2)*, pages 23–34.

19. Refaee, E. and Rieser, V. (2014). An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273.
20. Al-Smadi, M., Qawasmeh, O., Talafha, B., and Quwaider, M. (2015). Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*, pages 726–730. IEEE.
21. Abdulla, N. (2014). *Towards building a sentiment analysis tool for colloquial and modern standard arabic reviews.* PhD thesis, MasterâĂŹs thesis. Computer Science Department, Jordan University of Science and Technology, Irbid, Jordan.
22. Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.
23. Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*, pages 1058–1065.
24. Tomas M., Ilya S., Kai C. and all. Efficient Estimation of Word Representations in Vector Space. 2013. *Distributed representations of words and phrases and their compositionality.* In Advances in neural information processing systems, pages 3111–3119.
25. Bo P. and Lillian L., and Shivakumar V. 2002. *Thumbs up?: sentiment classification using machine learning techniques.* In In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics.
26. Prusa, J., Khoshgoftaar, T. M., and Seliya, N. (2015). The effect of dataset size on training tweet sentiment classifiers. In *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*, pages 96–102. IEEE.
27. Nabil, M. A. Aly, and A. F. Atiya 2015. *Astd : Arabic sentiment tweets dataset.* EMNLP volume 57 pages 2515–2519
28. Hogenboom, A., Bal, D., Frasincar, F., Bal, M., De Jong, F., and Kaymak, U. (2015). Exploiting emoticons in polarity classification of text. *J. Web Eng.*, 14(1&2):22–40.
29. Badaro, G., Baly, R., Hajj, H., Habash, N., and El-Hajj, W. (2014). A large scale arabic sentiment lexicon for arabic opinion mining. *ANLP 2014*, 165.
30. Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.
31. Mohamed Ali Sghaier , Houssem Abdellaoui , Rami Ayadi ,Mounir Zrigui 2014. *Analyse de sentiments et extraction des opinions pour les sites e-commerce : application sur la langue arabe.* CITALA 5th International Conference on Arabic Language Processing , Oujda, Morocco.