Mazajak: An Online Arabic Sentiment Analyser

Ibrahim Abu Farha
School of Informatics
University of Edinburgh
Edinburgh, UK
aibrahim@ed.ac.uk

Walid Magdy
School of Informatics
University of Edinburgh
Edinburgh, UK
wmagdy@inf.ed.ac.uk

Abstract

Sentiment analysis (SA) is one of the most useful natural language processing applications. Literature is flooding with many papers and systems addressing this task, but most of the work is focused on English. In this paper, we present "Mazajak", an online system for Arabic SA. The system is based on a deep learning model, which achieves state-of-the-art results on many Arabic dialect datasets including SemEval 2017 and ASTD. The availability of such system should assist various applications and research that rely on sentiment analysis as a tool.

1 Introduction

Sentiment analysis (SA) can be defined as the process of extracting and analysing the sentiment and polarity in a given piece of text (Liu, 2012). It is one of the tasks in the larger natural language processing (NLP) field. The rapid and wide increase in the use of social media platforms, and the reliance on online shopping and marketing resulted in a flood of information. Many researchers started analysing and mining data for the task of public opinion mining. Sentiment analysis is one of the vital approaches to extract public opinion from large corpora of text. Companies can benefit from understanding the feedback of their costumers and their opinions. Governments as well can use it to understand the reaction of people to their policies and actions.

Work on SA started in early 2000s, particularly with the work of (Pang et al., 2002), where they studied the sentiment of movies' reviews. The work has developed since then and it spanned different topics and fields such as social media. SA gained a lot of interest from researchers who recognised its importance and benefits. However, most of the work is focused on English whereas Arabic did not receive much attention until recently, but it still lacks behind due to the many challenges of the Arabic language; including the large variety in dialects (Habash, 2010; Darwish et al., 2014) and the complex morphology of the language (Abdul-Mageed et al., 2011).

Recently, the world witnessed a strong revolution in deep learning which was the driving force for many improvements in many fields. The work on English NLP started utilising deep learning models from an early stage, then followed by Arabic NLP. The utilisation of deep learning for Arabic SA started to receive more attention recently showing significant improvement in performance (Dahou et al., 2016; Al-Sallab et al., 2015; Alayba et al., 2018; Al-Smadi et al., 2018).

While there is a considerable amount of work that studies Arabic SA (Al-Ayyoub et al., 2019), to the best of our knowledge, there is no existing open-source tool for Arabic SA that could be used directly. The only work that we are aware of is SentiStrength[1] (Thelwall et al., 2010), which is mainly developed for English, but supports other languages including Arabic. However, it uses a basic dictionary-based approach that works with Arabic MSA and terribly fails with dialects which is the main language used in social media.

In this paper, we present Mazajak[2], an Online Arabic sentiment analysis system that utilises deep learning and massive Arabic word embeddings. The system is available as an online API that can be used by other researchers.

2 Related work

The literature of Arabic SA has many attempts to tackle the problem, however most of the work

[1] http://sentistrength.wlv.ac.uk/#Non-English

[2] http://mazajak.inf.ed.ac.uk:8000/

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 192–198
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

is based on conventional machine learning algorithms with few attempts to use deep learning. A recent publication (Al-Ayyoub et al., 2019) presents a comprehensive survey on Arabic SA.

In (Al-Smadi et al., 2017a), the authors proposed an aspect-based SA system for Arabic hotel reviews, in which they used SVM and recurrent neural networks (RNNs). In another work (Shoeb and Ahmed, 2017), the authors applied SA on tweets using Naive Bayes (NB) and KNN, they achieved relatively good results. Al-Ayyoub et al. (2015) also created a large lexicon of Arabic terms extracted from news articles. Based on their lexicon, they built an SA system and tested it on data collected from Twitter. In (Elmasry et al., 2014), the authors aimed to tackle the problem of dialects. They built a slang sentimental words and idioms lexicon (SSWIL) and conducted some experiments using SVM and the new lexicon.

In the realm of social media analysis, the work in (Abdulla et al., 2013) introduced a dataset of 2000 tweets, which the authors used to conduct an experiment with lexicon-based and ML-based systems. They found that combining both approaches would achieve better results. Abdul-Mageed et al. (2014) proposed an SA system for social media. In their work, they experimented and studied a large variety of features. They also studied the effect of the dialects and morphological richness of Arabic. Moreover, In (Abdul-Mageed, 2017a,b), the authors studied the different ways to handle the Arabic morphological richness for SA. They studied the effect of segmentation in representing the lexical input, also they tried to study the weight and importance of these segments for SA.

In SemEval 2017, a sentiment analysis task was presented that included Arabic (Rosenthal et al., 2017). El-Beltagy et al. (2017) were ranked first in SemEval 2017 task for Arabic SA. They used a set of hand-engineered and lexicon-based features, the classifier of choice was a complement NB classifier. The second rank in the same task was for the work of Jabreel and Moreno (2017), who introduced a rich set of features that are mostly based on bag of words (BoW) model in addition to some features extracted from word embeddings. They used SVM as their classification algorithm.

Dahou et al. (2016) proposed a set of word embeddings to be used for Arabic SA, which was built using a corpus of 3.4 billion words. They used a convolutional neural network (CNN) based system to evaluate their embeddings and the results were promising. Another use of word embeddings was in (Aziz Altowayan and Tao, 2016), where the authors created their own word embeddings and used them as the only features to be fed to the classifier without any engineered features, the results were comparable and slightly better than those of other systems.

In (Alayba et al., 2017), the authors presented their own SA dataset of opinions on health services. They built an SA system and it was tested on the new dataset. Their experiments included the use of many ML algorithms including CNNs. Al-Sallab et al. (2015) experimented with different deep learning models such as recursive auto-encoder (RAE), deep belief networks (DBN) and deep auto-encoder (DAE). They relied on the ArSenL lexicon (Badaro et al., 2014) to build the feature vectors. In (Al-Smadi et al., 2017b), the authors addressed the aspect-based sentiment analysis (ABSA). In their experiments, they used RNNs and SVM as classifiers, the results showed that SVM was superior. Alayba et al. (2018) built an SA system that is based on a combination of CNNs and LSTMs. They tested their model on two datasets, Ar-Twitter and Arabic Health Services datasets, where they achieved accuracies of 88.1% and 94.3% respectively. In (Al-Smadi et al., 2018), the authors proposed an aspect-based sentiment analysis system, their model is based on a Bi-LSTM and conditional random field (CRF). They tested their model on Arabic hotels' reviews dataset, they achieved an F-score of 70%. Elshakankery and Ahmed (2019) proposed a hybrid system for Arabic SA, that utilises lexicon-based and machine learning based approaches. In their work, they experimented with multiple dataset such as ASTD and ArTwitter. They used different classifiers for the task, which varied from using conventional machine learning, to deep learning models.

Among the previous mentioned work, we are not aware of any released open-source tool for Arabic SA, which is considered one of the largest limitations in Arabic NLP. While there are many Arabic NLP tools for various tasks, including segmentation, POS tagging, and diacritization (Pasha et al., 2014; Abdelali et al., 2016), the Arabic NLP research community still lack a tool for sentiment analysis. In this work, we offer the first open-source SA tool for Arabic social media .

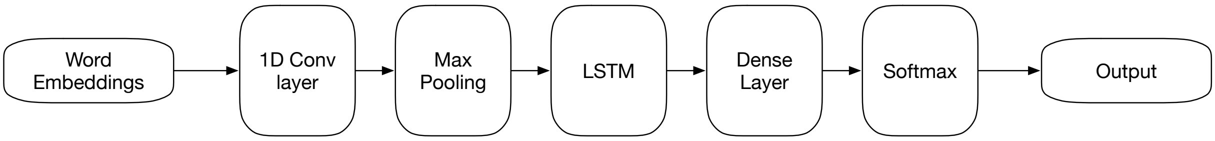


Figure 1: Model architecture.

3 Methodology

This section describes the different components and steps that are used by our system, Mazajak.

3.1 Data Preprocessing

In general, this step is an initial step that aims to reduce the inconsistencies and normalise the data into a coherent form so that it can be handled easily. The steps are mainly based on the work in (El-Beltagy et al., 2017). In our implementation, we used the following steps:

- Letter normalisation: unifying the letters that appear in different forms. We replace {، آ إ ، أ} with {ا}, {ة} with {ه} and {ى} with {ي} (Darwish et al., 2014).
- Elongation removal: removing the repeated letters which might appear specially in social media data (Darwish et al., 2012).
- Cleaning: removing unknown characters, diacritics, punctuation, URLs, etc.

3.2 Text representation

Sentences are represented as two dimensional matrix where each row represents a word, and each word is represented by its corresponding embedding. We set the size of the embedding D to 300. In our work, short sentences are padded to match the longest sentence in the training set.

Word embeddings were created using the word2vec (Mikolov et al., 2013), the skip-gram architecture was used. The embeddings were built using a corpus of 250M unique Arabic tweets; this makes it the largest Arabic word embeddings set when compared to the available AraVec (Soliman et al., 2017), which is currently the largest set, built using a corpus of 67M tweets. The tweets were collected over different time periods between 2013 and 2016 to ensure the coverage of different topics. The large and diverse corpus ensures that many dialects are covered which would help in reducing the effect of dialectal variation. When creating the embeddings, the same preprocessing steps utilised in the SA system were used.

Parameter	Value
#LSTM cells	128
Recurrent dropout	20%
Output dropout	20%
#Filters	300
Filter size	3
Pooling size	2
Optimizer	Adam
Learning rate	0.0001
Activation	ReLU

Table 1: CNN-LSTM model hyper-parameters.

3.3 Model Architecture

The model is built on a CNN followed by an LSTM. The CNN works as a feature extractor, where it learns the local patterns inside the sentence and provides representative features. The LSTM works on the extracted features where the context and word ordering would be taken into consideration. The model has been designed after extensive comparison to existing models in literature, and has been shown to be the most effective one among the state-of-the-art models, as demonstrated in next section. Figure 1 shows the architecture, the embeddings are fed into the CNN, after that they are fed to a max-pooling layer, the reason behind using max pooling is to have the most important features which conforms with the fact that sentiment is usually expressed in specific words. The extracted features are fed into an LSTM which is followed by a softmax layer that would give a probability distribution over the output classes. The hyper-parameters used in our architecture is shown in Table 1.

4 Model Performance

4.1 Experimental Setup

To examine the effectiveness of our model before offering it online for public use, we tested the model on three different datasets. The first is SemEval 2017 task 4-A benchmark dataset (Rosenthal et al., 2017), which consists of 6,100 testing tweets and 3,555 training ones. All tweets are labelled to one of three classes: positive, negative or neutral. The second dataset is ASTD benchmark dataset (Nabil et al., 2015), which con-

Dataset	*System*	*AvgRec*	F^{PN}	*Acc*
SemEval	(El-Beltagy et al., 2017)	0.58	0.61	0.58
	Mazajak	**0.61**	**0.63**	**0.62**
ASTD	(Heikal et al., 2018)	0.61	0.71	0.65
	Mazajak	**0.62**	**0.72**	**0.66**
ArSAS	Mazajak	**0.90**	**0.90**	**0.92**

Table 2: Mazajak performance in sentiment analysis in comparison to the state-of-the-art systems over three benchmark datasets

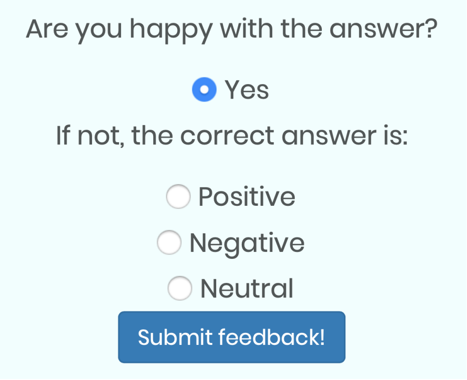
sists of 10,006 tweets, 6,691 of them are objective which means that they are not useful for SA. The rest are divided over three sentiment classes. The third dataset is ArSAS (Elmadany et al., 2018), the largest available dataset for Arabic SA which consists over 21K tweets labelled over four sentiment classes: positive, negative, neutral, and mixed. The mixed class has the smallest number of samples, thus we decided to ignore it. In addition, ArSAS has a confidence value for each label. We decided to keep only the tweets with confidence level over 50% and ignore the rest. After this step, we end up with 17,784 tweets in the ArSAS dataset labelled with three sentiment labels. Both ASTD and ArSAS datasets have no specific splitting of the data to test and train; thus, we applied random sampling to split both datasets to 80/20% for train/test respectively.

4.2 Baselines and Evaluation

To ensure having Mazajak achieving state-of-the-art performance, we compared its effectiveness to the existing best reported performance on each of the three datasets. For evaluation, we followed the same methodology adopted by SemEval 2017 task which uses average recall, F^{PN} and accuracy. F^{PN} is the macro-average F-score over the positive and negative classes only while neglecting the neutral class (Rosenthal et al., 2017). The best performing system in the SemEval 2017 task is the one described in (El-Beltagy et al., 2017) which achieved an F^{PN} of 0.61. For the ASTD, the best reported results are by (Heikal et al., 2018) who used an ensemble system combining output of CNN and Bi-LSTM architectures, which achieved an F^{PN} of 0.71. These two systems are used as our baselines. For the ArSAS dataset, we are not aware of any reported results on it yet.

4.3 Classification Performance

Table 2 reports the classification results of our system Mazajak and compares it to the state-of-the-art systems for the three benchmark datasets. As shown in the table, Mazajak model outperformed the current state-of-the-art models on the SemEval and ASTD datasets. In addition, it achieved a high performance on the ArSAS dataset. Our reported scores are higher than current top systems for all the evaluation scores, including average recall, F^{PN}, and accuracy. These results confirm that our model choice for our tool represents the current state-of-the-art for Arabic SA.


Are you happy with the answer?

● Yes

If not, the correct answer is:

○ Positive

○ Negative

○ Neutral

Figure 2: Sentiment feedback form on Mazajak.

5 Mazajak Online API

Our Arabic SA model is deployed as an online system, **Mazajak**[3], and can be accessed online at "`Mazajak.inf.ed.ac.uk:8000`".

The final model hosted online is trained on the SemEval and ASTD dataset combined[4].

Our online tool provides four modes of operation as follows:

- **Text Input:** where the user can input any piece of text into a text-box, and the system will display the polarity of the sentiment in the text. This mode allows the user to give

[3] the word "Mazajak" (مزاجك), is an Arabic word which means "your mood".

[4] this is different from the experimentation above when we were comparing the system to state-of-the-art.

Tweet	Sentiment
يعني لولا ما هالدكتور بدو يفصلني كان هسا انا بالبيت. اوفت.	negative
يعني خلاص الشتا خلص و هندخل في الحر و التلزيق و مش هلبس اللبس اللي مرمي في الدولاب وملبستوش ده	negative
مش كل اللي بيقرب منك عايز يخدعك في قلوب طيبة بدور علي اللي زيها	positive
وشلون اضيق بحضورك وانت كل الدروب اللي ليآ ضآق صدري سقت رجلي لها	positive
خروج ايطاليا من تصفيات كاس العالم اليوم علي يد السويد بالتعادل بعد ستين عام بدون انقطاع عن المونديال	neutral
غوغل تتحدى أبل وسامسونغ: أعلنت جوجل يوم الثلاثاء عن هاتف بيكسل جديد بكاميرا مميزة.	neutral

Table 3: Examples of some tweets classified using Mazajak.

feedback on the output sentiment using the form shown in Figure 2. This, in turn, would help to continuously collect more training data. The collected data is used periodically to update our model to improve the system performance.

- **Batch Mode:** where the user provides a file with multiple lines of text, and the system returns back an output file with the corresponding sentiment to each line in the input file.
- **Timeline mode:** where the user provides a Twitter account name, and the system will analyse the sentiment of the tweets in the account's timeline. The output is a graph showing the number of the tweets of each of the classes over time and an overall ratio of the percentages of the tweets corresponding to each class as shown in Figure 3.
- **Online API:** where an API could be downloaded to help other research that utilises sentiment analysis. The API provides two functions, either getting the sentiment of a sentence or a list of sentences. The API functions are provided in Python, but with a few lines of coding it can be accessed using other programming languages.

Table 3 shows some examples of classified tweets using the tool, these examples show that the model can handle the dialectal variations.

Our online system would be updated periodically with new training data and potentially better preforming models. We aim that Mazajak would serve the research community in analysing sentiment in Arabic text in a simple way, which, as we hope, would further promote the research in Arabic language.

6 Conclusion

In this paper, we presented Mazajak, the first online Arabic sentiment analysis tool. The system utilises the advancements in the NLP and deep learning fields. The model, on which the system relies, achieves state-of-the-art results on three of the benchmark datasets for Arabic SA including SemEval 2017 task, ASTD and ArSAS. The system is available as an online API that can be accessed easily, which would help and ease the work of other researchers in applications that make use of sentiment information.

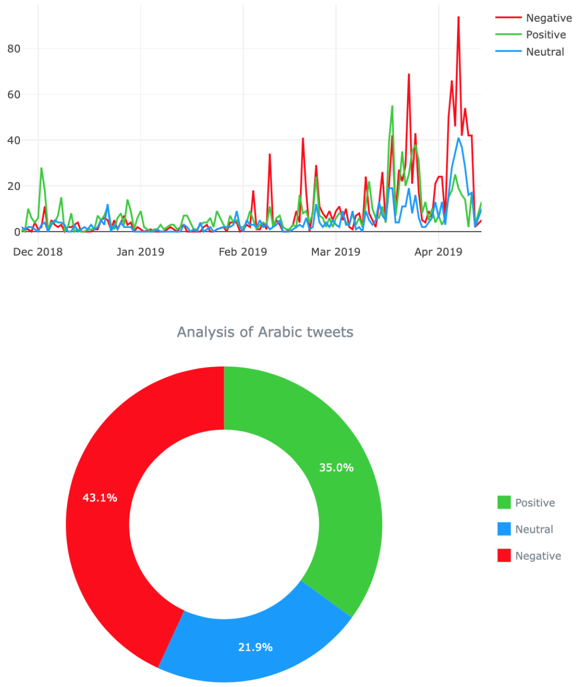


Figure 3: Twitter timeline analysis sample output.

Mazajak is offered for free use for research purposes. For commercial usage, please contact the authors.

In the future, we hope to improve the model so it would achieve better results. Also, we look forward to add more features such as the ability to handle Arabizi –Arabic written in English alphabet– and emojis.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16.

Muhammad Abdul-Mageed. 2017a. Modeling arabic subjectivity and sentiment in lexical space. *Information Processing & Management*.

Muhammad Abdul-Mageed. 2017b. Not all segments are created equal: Syntactically motivated sentiment analysis in lexical space. In *Proceedings of the third Arabic natural language processing workshop*, pages 147–156.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Muhammad Abdul-Mageed, Mona T. Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 587–591, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nawaf A. Abdulla, Nizar A. Ahmed, Mohammed A. Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*, pages 1–6. IEEE.

Mahmoud Al-Ayyoub, Safa Bani Essa, and Izzat Alsmadi. 2015. Lexicon-based sentiment analysis of arabic tweets. *International Journal of Social Network Mining*, 2(2):101–114.

Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N. Al-Kabi. 2019. A comprehensive survey of arabic sentiment analysis. *Information Processing & Management*, 56(2):320–342.

Ahmad Al-Sallab, Hazem Hajj, Gilbert Badaro, Ramy Baly, Wassim El-Hajj, and Khaled Bashir Shaban. 2015. Deep learning models for sentiment analysis in arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 9–17.

Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. 2017a. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. *Journal of Computational Science*.

Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. 2017b. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. *Journal of Computational Science*.

Mohammad Al-Smadi, Bashar Talafha, Mahmoud Al-Ayyoub, and Yaser Jararweh. 2018. Using long short-term memory deep neural networks for aspect-based sentiment analysis of arabic reviews. *International Journal of Machine Learning and Cybernetics*, pages 1–13.

Abdulaziz M. Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. 2017. Arabic language sentiment analysis on health services. *CoRR*, abs/1702.03197.

Abdulaziz M. Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. 2018. A combined cnn and lstm model for arabic sentiment analysis. *arXiv preprint arXiv:1807.02911*.

A. Aziz Altowayan and Lixin Tao. 2016. Word embeddings for arabic sentiment analysis. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3820–3825.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 165–173.

Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, Mohamed Houcine Haddoud, and Pengfei Duan. 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*, pages 2418–2427.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2427–2430. ACM.

Kareem Darwish, Walid Magdy, et al. 2014. Arabic information retrieval. *Foundations and Trends® in Information Retrieval*, 7(4):239–342.

Samhaa R. El-Beltagy, Mona El Kalamawy, and Abu Bakr Soliman. 2017. Niletmrg at semeval-2017 task 4: Arabic sentiment analysis. *arXiv preprint arXiv:1710.08458*.

AbdelRahim Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. Arsas: An arabic speech-act and sentiment corpus of tweets.

Mostafa Elmasry, Taysir Soliman, and Abdel-Rahman Hedar. 2014. Sentiment analysis of arabic slang comments on facebook. *International Journal of Computers & Technology*, 12(5):3470–3478.

Kariman Elshakankery and Mona F. Ahmed. 2019. Hilatsa: A hybrid incremental learning approach for arabic tweets sentiment analysis. *Egyptian Informatics Journal*.

Nizar Y. Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Maha Heikal, Marwan Torki, and Nagwa El-Makky. 2018. Sentiment analysis of arabic tweets using deep learning. *Procedia Computer Science*, 142:114–122.

Mohammed Jabreel and Antonio Moreno. 2017. Sitaka at semeval-2017 task 4: Sentiment analysis in twitter based on a rich set of features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 694–699.

Bing Liu. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T. Diab, Ahmed El-Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Md Shoeb and Jawed Ahmed. 2017. Sentiment analysis and classification of tweets using data mining. *International Research Journal of Engineering and Technology*, 4(12).

Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256 – 265. Arabic Computational Linguistics.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.