EMA at SemEval-2018 Task 1: Emotion Mining for Arabic

Gilbert Badaro, Obeida El Jundi, Alaa Khaddaj, Alaa Maarouf, Raslan Kain
Hazem Hajj, Wassim El-Hajj†
Department of Electrical and Computer Engineering, American University of Beirut
† Department of Computer Science, American University of Beirut
Beirut, Lebanon
{ggb05;oae15;awk11;aim20;rhk44;hh63;we07}@aub.edu.lb

Abstract

While significant progress has been achieved for **O**pinion **M**ining in **A**rabic (**OMA**), very limited efforts have been put towards the task of Emotion mining in Arabic. In fact, businesses are interested in learning a fine-grained representation of how users are feeling towards their products or services. In this work, we describe the methods used by the team **E**motion **M**ining in **A**rabic (**EMA**), as part of the SemEval-2018 Task 1 for Affect Mining for Arabic tweets. EMA participated in all 5 subtasks. For the five tasks, several preprocessing steps were evaluated and eventually the best system included diacritics removal, elongation adjustment, replacement of emojis by the corresponding Arabic word, character normalization and light stemming. Moreover, several features were evaluated along with different classification and regression techniques. For the 5 subtasks, word embeddings feature turned out to perform best along with Ensemble technique. EMA achieved the 1st place in subtask 5, and 3rd place in subtasks 1 and 3.

1 Introduction

Emotion recognition has captured the interest of researchers for many years. Different models have been used to detect people's emotions such as human computer interaction (HCI) (Hibbeln et al., 2017; Patwardhan and Knapp, 2017; Constantine et al., 2016) and their facial expressions (Trad et al., 2012; Wegrzyn et al., 2017). Recently, with Web 2.0, the size of textual data charged with opinions and emotions on the web has tremendously increased. Thus, researchers have been looking at automatically performing sentiment and emotion analysis from textual data. In fact, learning emotions of users is critical for different applications such as shaping marketing strategies (Bougie et al., 2003), providing customers with better personalized recommendations for advertisements and products (Mohammad and Yang, 2011), improving recommendation of typical recommender systems (Badaro et al., 2013, 2014c,d), tracking emotions of users towards politicians, movies, music, products, etc, (Pang et al., 2008), or accurately predicting stock market prices (Bollen et al., 2011).

Some efforts have already been placed in developing emotion classification models from text (Shaheen et al., 2014; Houjeij et al., 2012; Abdul-Mageed and Ungar, 2017). Since sentiment lexicons helped in improving the accuracy of sentiment classification models (Liu and Zhang, 2012; Taboada et al., 2011), several researchers are working on developing emotion lexicons for different languages such as English, French, Chinese (Mohammad, 2017; Bandhakavi et al., 2017; Yang et al., 2007; Poria et al., 2012; Das et al., 2012; Mohammad et al., 2013; Abdaoui et al., 2017; Staiano and Guerini, 2014; Badaro et al., 2018a). There were also couple of attempts for developing Arabic emotion lexicons (Mohammad and Turney, 2013; Mohammad et al., 2013; El Gohary et al., 2013; Badaro et al., 2018b).

Building on our previous work on opinion mining which involved development of sentiment lexicons (ArSenL (Badaro et al., 2014a)), opinion mining models (Baly et al., 2014; Al Sallab et al., 2015; Al-Sallab et al., 2017; Baly et al., 2017b) and applications (Badaro et al., 2014b, 2015), and building on our analysis and characterization for Twitter Data (Baly et al., 2017a,c), we participate in SemEval 2018 Task 1 (Mohammad et al., 2018): Affect in Arabic Tweets. In fact, analyzing sentiment and emotions from dialectal Arabic such as text data from Twitter is of great importance given the tremendous increase of Arabic speaking users

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 236–244
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

on Twitter.[1]

In this paper, we describe our approaches to SemEval 2018 Task 1 (Mohammad et al., 2018): Affect in Arabic Tweets, along with the achieved results for each of the subtasks where we employed preprocessing steps, features and classification models based on our prior work on sentiment analysis. In section 2, we present a brief overview of related work to emotion classification for English and Arabic. In section 3, we describe the five subtasks that are part of Affect in Tweet task. In section 4, we present our proposed approach and finally, we conclude in section 5.

2 Related Work

There have been extensive efforts for extracting emotions from different modalities including HCI (Constantine et al., 2016; Hibbeln et al., 2017; Patwardhan and Knapp, 2017), facial expressions (Trad et al., 2012; Wegrzyn et al., 2017) and speech (Houjeij et al., 2012). The related work for text emotion classification can be categorized into approaches for Emotion classification in English, that are leading the advances, versus research progress in Emotion in Arabic texts.

Emotion detection task from text is usually defined as a categorical classification task, where given a text, the classifier needs to predict the emotion label corresponding to the input text. Two typical categorical representations for emotions exist: Ekman representation (Ekman, 1992) which includes anger, happiness, surprise, disgust, sadness and fear and Plutchik model (Plutchik, 1980, 1994) which includes Ekman's six emotions in addition to two labels: trust and anticipation.

2.1 English Emotion Analysis

In general, there are three different approaches for emotion classification: keyword-based detection, learning-based detection, and hybrid detection (Avetisyan et al., 2016).

Keyword-based techniques, also known as lexicon-based, depend on identifying emotional keywords in the input sentence (Strapparava et al., 2004; Mohammad and Turney, 2010, 2013). These models rely on the existence of large scale emotion lexicons and their accuracy is correlated with the accuracy of the emotion lexicon that is being utilized. On the other hand, they do not require the existence of training data.

Learning-based approaches or feature-based approaches depend on the existence of annotated training data that are processed in order to extract several features such as syntactic, stylistic and semantic features (Ho and Cao, 2012; Bandhakavi et al., 2017). Additionally, in hybrid methods, emotions are detected by using a combination of emotional keywords and learning patterns collected from training datasets.

Due to the notable lack of resources related to emotion (annotated data and lexicons), progress on automatic affect intensity is still lagging. Mohammad and Bravo-Marquez (2017) created not only the first datasets of tweets annotated with emotion intensities, but also developed an emotion regression system with benchmark results. Abdul-Mageed and Ungar (2017) developed a large scale English dataset with fine grained emotion labels and trained deep learning models on top of it achieving an average accuracy of 87.58%.

2.2 Arabic Emotion Analysis

Emotion recognition for Arabic text has been gaining more attention recently. El Gohary et al. (2013) applied a knowledge-based approach to achieve 65% accuracy on the six basic Ekman emotions. Rabie and Sturm (2014) extracted a sample Arabic emotion lexicon and demonstrated how it enhanced the emotion detection results. Sayed et al. (2016) utilized Conditional Random Fields (CRF) and AdaBoost classifiers for classifying emotions of tweets and expression levels in which CRF achieved the best results. Alsharif et al. (2013) used Naive Bayes and SVM to classify Arabic poems into four emotion classes.

While some attempts were performed for Emotion recognition from Arabic text, there is still a lot of area for improvement as for example, developing large scale emotion lexicon for more accurate emotion recognition model, developing highly accurate emotion mining models for MSA as well as dialectal Arabic whether through the use of feature based approaches or deep learning.

3 SemEval 2018 Task 1: Affect in Arabic Tweets

We describe in this section the subtasks of SemEval 2018 task 1.

[1] https://weedoo.tech/twitter-arab-world-statistics-feb-2017/

3.1 Subtasks' Descriptions

SemEval 2018 Task 1 Affect in Tweets (Mohammad et al., 2018) included five subtasks each with annotated dataset for English, Arabic and Spanish. The tasks were as follows:

1. EI-reg (Emotion Intensity Regression Task): Given a tweet and an emotion E (anger, fear, joy or sadness), determine the intensity of E that best represents the emotion intensity of the tweeter by predicting a real-valued score between 0 (least E) and 1 (most E).

2. EI-oc (Emotion Intensity Ordinal Classification): Given a tweet and an emotion E, classify the tweet into one of four ordinal classes of intensity of E, from 0 (low amount) to 3 (high amount), that best represents the mental state of the tweeter.

3. V-reg (a sentiment intensity regression task): Given a tweet, determine the valence (V) that best represents the mental state of the tweeter by predicting a real-valued score between 0 (most negative) and 1 (most positive).

4. V-oc (a sentiment analysis, ordinal classification, task): Given a tweet, classify it into one of seven ordinal classes, from -3 (very negative) to +3 (very positive), corresponding to various levels of positive and negative sentiment intensity, that best represents the sentiment of the tweeter.

5. E-c (an emotion classification task): Given a tweet, classify it as *neutral (no emotion)* or as one, or more, of eleven given emotions that best represent the tweeter.

3.2 Datasets

For each of the 5 tasks, 3 sets of datasets were released, each set corresponding to a language (English, Arabic and Spanish). For each language, 3 datasets were released (training, development and test). For subtasks 1 and 2 Arabic, each emotion of the four emotions had a training set of around 800 tweets on average and a development set of around 200 tweets. Subtasks 3 and 4 Arabic had a dataset consisting of 932 tweets for training and 138 tweets for development. For subtask 5 Arabic, 2278 tweets were used for training and 585 tweets for development.

4 Explored Models for Competition

We present a description of EMA system covering preprocessing steps, features used, machine learning models employed and results achieved. An overview of the system is show in Figure 1.

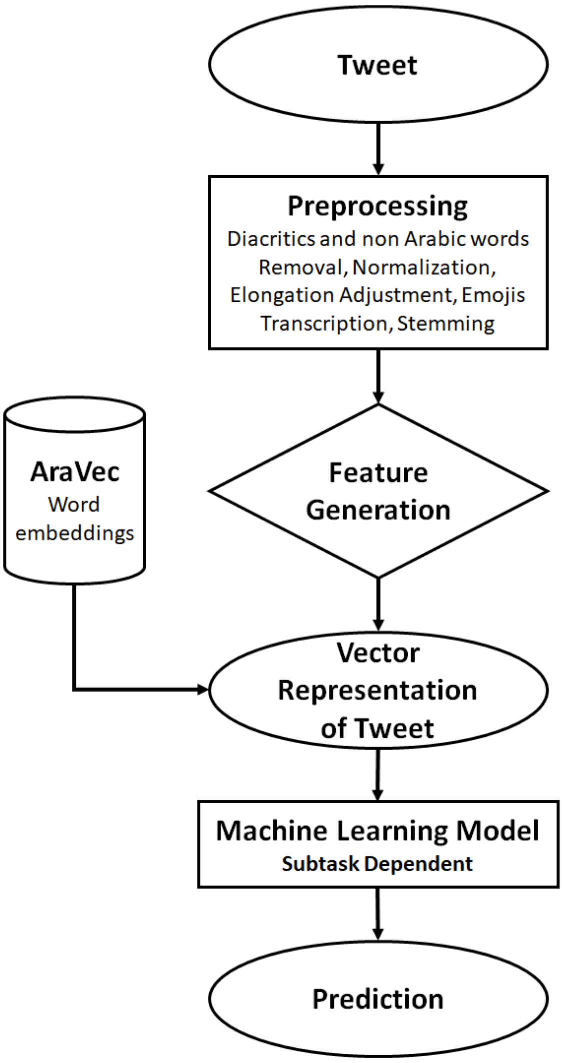
Tweet
Preprocessing
Diacritics and non Arabic words Removal, Normalization, Elongation Adjustment, Emojis Transcription, Stemming
AraVec
Word embeddings
Feature Generation
Vector Representation of Tweet
Machine Learning Model
Subtask Dependent
Prediction

Figure 1: Overview of EMA System.

4.1 Preprocessing

The provided datasets contained raw tweets that included different properties used in Twitter such as hashtags, user mentions, urls, images, Arabizi and emojis. Thus, preprocessing steps were needed to enhance the analysis of the tweet. We experimented with different preprocessing configurations that led to mixed results. For example, using stems instead of lemmas proved to be better. One justification is that tweets are mostly in dialectal Arabic while most Arabic morphological analyzers are trained on MSA data. We present next the steps that led to the best performance.

We first applied the normalization rules followed by Shoukry and Rafea (2012): Diacritics were removed, the "hamza" on characters was normalized in addition to normalizing some word ending characters such as the “t marbouta” and “ya’ maqsoura”. We then removed elongations as well as non Arabic letters. We manually created a lexicon containing the most frequent emojis in tweets and transcribed each emoji to its corresponding Arabic word. The lexicon consisted of 100 emojis. The tweets were finally stemmed using A Robust Arabic Light Stemmer (ARLSTEM) (Abainia et al., 2017).

4.2 Features

We have tried different features separately including unigrams, bigrams, trigrams, scores from emotion lexicon, ArSEL (Badaro et al., 2018b), sentiment lexicon, ArSenL ((Badaro et al., 2014a) and word embeddings from AraVec (Soliman et al., 2017) and FastText by Facebook (Bojanowski et al., 2016). AraVec was trained on three different datasets (Wikipedia, Text data from Web and Twitter) while FastText was trained on Wikipedia. Using word embeddings from AraVec outperformed significantly all other features including word embeddings trained on Wikipedia provided by Facebook. This is likely due to the fact that AraVec is a large scale dataset (around 205,000 words) trained on the same data domain (twitter), and includes several Arabic dialects. Word embeddings overcome the problem of sparsity present with n-grams and also reduce semantic complexity by providing similar representations to words that can appear in the same context. Each word was represented by a vector of real numbers of dimension 300. The sentence embeddings were computed by taking the average of its word embeddings. If a word did not have a vector representation, we tried using its stem’s representation. If neither the word nor its stem had a vector representation in AraVec, the average of the embeddings of closest words was utilized. By closest words, we mean words that had the smallest minimum edit distance (Levenshtein distance) with the target term. Eventually, each tweet was represented by a vector consisting of 300 real valued numbers. The same feature is used for all subtasks. For feature extraction, we used Python with NLTK, gensim and Numpy libraries.

4.3 Classification and Regression Models

Overall, we tried different learning models including Ridge regression, support vector machines, random forests, ensemble methods and deep neural networks such as convolutional neural networks with long short term memory layer. Deep neural networks performed poorly compared to other models. One possible explanation was that the training data size was very small and deep neural networks perform best when trained on a large scale data to ensure a well representation of the data (Beleites et al., 2013).

For regression subtasks 1 and 3, we tried different machine learning models including Ridge, Elastic Net, Decision Trees, random forest, xgboost and support vector regressor with (rbf kernel). The best was an Ensemble of Ridge regression (RR), Support Vector Regressor (SVR), and Random Forests (RF). In fact, the 3 models performed reasonably well on their own. For classification subtasks 2 and 4, we also tried different classification models including Ridge, Elastic Net, Decision Trees, Random Forest, Support Vector Classifier (SVC) with linear and non linear kernels and convolutional neural nets. For subtask 2, SVC performed best. As for subtask 4, an ensemble of SVC and Ridge Classifier performed best. Ridge Classifier allows defining a linear mapping without allowing weights to be large thanks to regularization effect for generalization while SVM tries to find the best classification margins. Adding ElasticNet did not help much since L1 and L2 errors were already covered by optimized using the ensemble of Ridge and SVM. Moreover, Zhou et al. (2015) shows that ElasticNet can be reduced to SVM. Random Forest with its large number of estimators had a better generalization than regular decision trees. Combining all these models in an ensemble model ensured a better generalization and accuracy on the test data.

For subtask 5, we tested SVC (with both penalties L1 and L2), RC, RF and Ensemble. SVC with L1 performed best. While Pearson correlation measure was used for evaluating subtasks 1 to 4, Accuracy was used to evaluate subtask 5.

For all subtasks, we utilized the training data for training the different models and the development set was treated as unseen data in order to make sure that comparison across the different models is fair. The best model was selected based on its performance on the development set. Our focus was on feature extraction and preprocessing, so most feature-based models performed well. One main problem faced in all problems was sparsity, since most tweets were in Dialectical Arabic.

4.4 Experimental Results

All experiments were conducted using Python with scikit-learn and Keras libraries. A grid search mechanism was utilized to optimize the hyperparameters of the different learning models used and whose performances are reported in below tables: alpha parameter for Ridge, penalty C, kernel and gamma for Support Vectors, and, number trees, maximum tree depth and number of features per tree for Random Forests. Rows 2 to 5 in tables 1 and 2 show the results (Pearson Score) of the different regression techniques used for subtasks 1 and 3 respectively on the corresponding development sets for each of the four emotions (Joy, Sadness, Poor and Anger). Average performance is also reported in the last column. The last two rows in table 1 show the result on the test set of our Ensemble model on average and per each emotion and the performance of the best team for subtask1 respectively. The last two rows in table 2 show the performance of Ridge Regression on the test set and the performance of the best team respectively. In both subtasks, EMA ranked 3rd among participants. By examining the results of the different participants in subtask 1, we can observe that the proposed systems perform best for the Joy emotion. Tables 3 and 4 show the hyperparameters for each technique employed. For Random Forest, the number of estimators was set to 1000.

In Tables 5 and 6, we show the results of subtasks 2 and 4 respectively. SVC was the best performing model on the development set in subtask 2 and Ensemble methods performed best in subtask 4. The last column in table 5 shows the performance of SVC on the test set on average and per each of the four emotions. The last row in table 6 represents the Pearson score achieved by the Ensemble of RC and SVC on the test set. EMA was ranked 8th and 5th in subtasks 2 and 4 respectively. Tables 7 and 8 show the best hyperparameters of the classification models used.

Regression Model	Joy	Sadness	Fear	Anger	Avg
RR	0.610	0.635	0.481	0.566	0.573
SVR	0.615	0.628	0.484	0.567	0.574
RF	0.578	0.547	0.413	0.458	0.499
Ensemble	0.624	0.630	0.488	0.563	**0.576**
Ensemble on Test	0.709	0.656	0.593	0.615	**0.643**
Best (AffecThor)	0.756	0.694	0.642	0.647	0.685

Table 1: Subtask 1 Pearson Correlation Results on Dev and Test Sets. RR = Ridge Regression; SVR = Support Vector Regressor; RF = Random Forest.

Regression Model	**Pearson Correlation**
RR	**0.746**
SVR	0.744
RF	0.609
Ensemble	0.737
Ensemble on Test	**0.804**
Best (EiTAKA)	0.8284

Table 2: Subtask 3 Pearson Correlation Results on Dev and Test Sets. RR = Ridge Regression; SVR = Support Vector Regressor; RF = Random Forest.

Regression Model	Joy	Sadness	Fear	Anger
Ridge (alpha)	7.1	5.9	3.7	4.9
SVR (C)	4.4	4.7	10	4.9
RF (depth)	10	10	10	10

Table 3: Subtask 1 Regression Models' Hyperparameters.

Regression Model	**Parameter Value**
Ridge (alpha)	**3.9**
SVR (C)	5.6
RF (depth)	10

Table 4: Subtask 3 Regression Models' Hyperparameters.

Model	**RC**	**SVC**	**Ens**	**SVC on Test**	**Best (AffecThor)**
Joy	0.502	0.484	0.480	0.215	0.631
Sadness	0.587	0.594	0.589	0.535	0.618
Fear	0.373	0.431	0.390	0.242	0.551
Anger	0.472	0.518	0.497	0.077	0.551
Average	0.484	**0.507**	0.489	**0.267**	0.587

Table 5: Subtask 2 Pearson Correlation Results on Dev and Test Sets. RC = Ridge Classification; SVC = Support Vector Classifier; Ens = Ensemble.

Classification Model	**Pearson Correlation**
RC	0.611
SVC	0.623
Ensemble	**0.625**
Ensemble on Test	**0.643**
Best (EiTAKA)	0.809

Table 6: Subtask 4 Pearson Correlation Results on Dev and Test Sets. RC = Ridge Classification; SVC = Support Vector Classifier.

Model	**RC (alpha)**	**SVC (C)**
Joy	18.2	19.5
Sadness	3.3	29.4
Fear	20.6	17.1
Anger	15.4	19.5

Table 7: Subtask 2 Classification Models' Hyperparameters.

Finally, Table 9 shows the results of subtask 5 on the development and the test sets where for a given tweet, the tweet is classified either as neutral

Classification Model	**Parameter Value**
RC (alpha)	27.2
SVC (C)	10.7

Table 8: Subtask 4 Classification Models' Hyperparameters.

or as one or more of 11 emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust). Linear SVC performed best among all classifiers. EMA ranked 1st in subtask 5. Table 10 shows the best hyperparameters for each classification model used. The number of estimators for Random Forest was set to 1000.

Classification Model	**Accuracy**
SVC L1	**0.488**
SVC L2	0.484
RC	0.443
RF	0.370
Ensemble	0.401
SVC L1 on Test	**0.489**

Table 9: Subtask 5 Accuracy Results on Dev and Test Sets. RC = Ridge Classification; SVC = Support Vector Classifier; RF = Random Forest.

Classification Model	**Parameter Value**
SVC L1 (C)	1.98
SVC L2 (C)	0.3
RC (alpha)	7.9
RF (depth)	14

Table 10: Subtask 5 Classification Models' Hyperparameters.

5 Conclusion and Future Work

In this paper, we presented EMA (Emotion Mining in Arabic) at SemEval 2018 Task 1 Affect in Tweets to perform Arabic Emotion and Sentiment mining. Several methods were tested for deciding on features, regression and classification techniques. Word embeddings provided the best feature while the choice of the predictor was task dependent. EMA ranked 1st in subtask 5 and 3rd in subtasks 1 and 3. As future work, we suggest finding the best combination of the different features that were employed in separate models. Other future work includes dealing with sparsity caused by dialectal Arabic.

References

Kheireddine Abainia, Siham Ouamour, and Halim Sayoud. 2017. A novel robust arabic light stemmer. *Journal of Experimental & Theoretical Artificial Intelligence*, 29:557–573.

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine–grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 718–728.

Ahmad Al-Sallab, Ramy Baly, Hazem Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro. 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):25.

Ahmad A Al Sallab, Ramy Baly, Gilbert Badaro, Hazem Hajj, Wassim El Hajj, and Khaled B Shaban. 2015. Deep learning models for sentiment analysis in arabic. In *ANLP Workshop*, volume 9.

Ouais Alsharif, Deema Alshamaa, and Nada Ghneim. 2013. Emotion classification in arabic poetry using machine learning. *International Journal of Computer Applications*, 65(16).

H Avetisyan, O Bruna, and J Holub. 2016. Overview of existing algorithms for emotion classification. uncertainties in evaluations of accuracies. In *Journal of Physics: Conference Series*, volume 772. IOP Publishing.

Gilbert Badaro, Ramy Baly, Rana Akel, Linda Fayad, Jeffrey Khairallah, Hazem Hajj, Khaled Shaban, and Wassim El-Hajj. 2015. A light lexicon-based mobile application for sentiment mining of arabic tweets. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 18–25.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014a. A large scale Arabic sentiment lexicon for Arabic opinion mining. *ANLP 2014*, 165.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, Wassim El-hajj, and Khaled Shaban. 2014b. An efficient model for sentiment classification of Arabic tweets on mobiles. In *Qatar*

Foundation Annual Research Conference, 1, page ITPP0631.

Gilbert Badaro, Hazem Hajj, Wassim El-Hajj, and Lama Nachman. 2013. A hybrid approach with collaborative filtering for recommender systems. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*, pages 349–354. IEEE.

Gilbert Badaro, Hazem Hajj, Ali Haddad, Wassim El-Hajj, and Khaled Bashir Shaban. 2014c. A multiresolution approach to recommender systems. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, page 9. ACM.

Gilbert Badaro, Hazem Hajj, Ali Haddad, Wassim El-Hajj, and Khaled Bashir Shaban. 2014d. Recommender systems using harmonic analysis. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 1004–1011. IEEE.

Gilbert Badaro, Hussein Jundi, Hazem Hajj, and Wassim El-Hajj. 2018a. Emowordnet: Automatic expansion of emotion lexicon using english wordnet. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM2018) co-located with NAACL2018.*

Gilbert Badaro, Hussein Jundi, Hazem Hajj, Wassim El-Hajj, and Nizar Habash. 2018b. Arsel: A large scale arabic sentiment and emotion lexicon. In *Proceedings of the 3rd Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT3) co-located with LREC2018.*

Ramy Baly, Gilbert Badaro, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Wassim El-Hajj, Nizar Habash, and Khaled Shaban. 2017a. A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 110–118.

Ramy Baly, Gilbert Badaro, Hazem Hajj, Nizar Habash, Wassim El Hajj, and Khaled Shaban. 2014. Semantic model representation for human's pre-conceived notions in arabic text with applications to sentiment mining. In *Qatar Foundation Annual Research Conference*, 1, page ITPP1075.

Ramy Baly, Gilbert Badaro, Ali Hamdi, Rawan Moukalled, Rita Aoun, Georges El-Khoury, Ahmad Al Sallab, Hazem Hajj, Nizar Habash, Khaled Shaban, et al. 2017b. Omam at semeval-2017 task 4: Evaluation of english state-of-the-art sentiment analysis models for arabic and a new topic-based model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 603–610.

Ramy Baly, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Khaled Bashir Shaban, and Wassim El-Hajj. 2017c. Comparative evaluation of sentiment analysis methods across arabic dialects. *Procedia Computer Science*, 117:266–273.

Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, and Deepak Padmanabhan. 2017. Lexicon generation for emotion detection from text. *IEEE intelligent systems*, 32(1):102–108.

Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. 2013. Sample size planning for classification models. *Analytica chimica acta*, 760:25–33.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Roger Bougie, Rik Pieters, and Marcel Zeelenberg. 2003. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the Academy of Marketing Science*, 31(4):377–393.

Layale Constantine, Gilbert Badaro, Hazem Hajj, Wassim El-Hajj, Lama Nachman, Mohamed BenSaleh, and Abdulfattah Obeid. 2016. A framework for emotion recognition from human computer interaction in natural setting. *22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016), Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM 2016).*

Dipankar Das, Soujanya Poria, and Sivaji Bandyopadhyay. 2012. A classifier based approach to emotion lexicon construction. In *NLDB*, pages 320–326. Springer.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Amira F El Gohary, Torky I Sultan, Maha A Hana, and Mohamed M El Dosoky. 2013. A computational approach for analyzing and detecting emotions in arabic text. *International Journal of Engineering Research and Applications (IJERA)*, 3:100–107.

Martin Hibbeln, Jeffrey L Jenkins, Christoph Schneider, Joseph S Valacich, and Markus Weinmann. 2017. How is your user feeling? inferring emotion through human–computer interaction devices. *MIS Quarterly*, 41(1).

Dung T Ho and Tru H Cao. 2012. A high-order hidden markov model for emotion detection from textual data. In *Pacific Rim Knowledge Acquisition Workshop*, pages 94–105. Springer.

Ali Houjeij, Layla Hamieh, Nader Mehdi, and Hazem Hajj. 2012. A novel approach for emotion classification based on fusion of text and speech. In *Telecommunications (ICT), 2012 19th International Conference on*, pages 1–6. IEEE.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Saif M Mohammad. 2017. Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. *arXiv preprint arXiv:1708.03696*.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M Mohammad and Tony Wenda Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 70–79. Association for Computational Linguistics.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Amol S Patwardhan and Gerald M Knapp. 2017. Multimodal affect analysis for product feedback assessment. *arXiv preprint arXiv:1705.02694*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1(3-31):4.

Robert Plutchik. 1994. *The psychology and biology of emotion.* HarperCollins College Publishers.

Soujanya Poria, Alexander Gelbukh, Dipankar Das, and Sivaji Bandyopadhyay. 2012. Fuzzy clustering for semi-supervised learning–case study: Construction of an emotion lexicon. In *Mexican International Conference on Artificial Intelligence*, pages 73–86. Springer.

Omneya Rabie and Christian Sturm. 2014. Feel the heat: Emotion detection in arabic social media content. In *The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014)*, pages 37–49. The Society of Digital Information and Wireless Communication.

Amr M Sayed, Samir AbdelRahman, Reem Bahgat, and Aly Fahmy. 2016. Time emotional analysis of arabic tweets at multiple levels. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 7(10):336–342.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 383–392. IEEE.

Amira Shoukry and Ahmed Rafea. 2012. Preprocessing egyptian dialect tweets for sentiment mining. In *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, page 47.

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.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: A lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Chadi Trad, Hazem M Hajj, Wassim El-Hajj, and Fatima Al-Jamil. 2012. Facial action unit and emotion recognition with head pose variations. In *ADMA*, pages 383–394. Springer.

Martin Wegrzyn, Maria Vogt, Berna Kireclioglu, Julia Schneider, and Johanna Kissler. 2017. Mapping the emotional face. how individual face parts contribute to successful emotion recognition. *PloS one*, 12(5):e0177239.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136. Association for Computational Linguistics.

Quan Zhou, Wenlin Chen, Shiji Song, Jacob R Gardner, Kilian Q Weinberger, and Yixin Chen. 2015. A reduction of the elastic net to support vector machines with an application to gpu computing. In *AAAI*, pages 3210–3216.