Sentiment Classifier: Logistic Regression for Arabic Services' Reviews in Lebanon

Marwan Al Omari
Centre for Language Sciences and Communication
Lebanese University
Beirut, Lebanon
marwanalomari@yahoo.com

Moustafa Al-Hajj
Centre for Language Sciences and Communication
Lebanese University
Beirut, Lebanon
moustafa.alhajj@ul.edu.lb

Nacereddine Hammami
Faculty of Computer Science and Information
Jouf University
Al-Jawf, KSA
n.hammami@ju.edu.sa

Amani Sabra
Centre for Language Sciences and Communication
Lebanese University
Beirut, Lebanon
amani.sabra@ul.edu.lb

***Abstract*— This paper proposes a logistic regression approach paired with term and inverse document frequency (TF*IDF) for Arabic sentiment classification on services' reviews in Lebanon country. Reviews are about public services, including hotels, restaurants, shops, and others. We collected manually from Google reviews and Zomato, which have reached to 3916 reviews. Experiments show three core findings: 1) The classifier is confident when used to predict positive reviews. 2) The model is biased on predicting reviews with negative sentiment. Finally, the low percentage of negative reviews in the corpus contributes to the diffidence of logistic regression model.**

Keywords*— *Arabic natural language, Services' reviews, Logistic regression, Supervised machine learning, Arabic sentiment analysis.

I. INTRODUCTION

The content amount available on the World Wide Web (WWW) has significant increase over the past few decades. Particularly, social media has given an equal chance for all users to post whatever they want. One of those things is posting their thoughts and feelings about many diverse things [21] such as a service or a product provided by either the public sector as like government institutions or the private sector as companies ([6], [20]). According to the results obtained from [16], 30% of people in the Arab region uses Arabic language on social media in all aspects of their lives, one of these aspects is to review services they got from restaurants, institutions, cafés, shopping centers, etc. These reviews enfold many expressions and feelings.

Due to the fast growth of data on the web and the need to identify feelings and impressions related to vast amount of data, data must go into analysis to output a sentiment predication of what the user feels and reckons in the virtual world. Sentiment Analysis (SA) predicts users' opinions on product, service, food, education, hotel, etcetera [1]. SA, as one of natural language processing (NLP) applications, extracts sentiment information out of a given text input through class assignment, for example, positive or negative [17]. This classification could be of great help for media offices, government centers, research facilities, and businesses.

The purpose of this research is to use Logistic Regression model to classify reviews in Arabic language about restaurant, hotels, etc., in binary classification task: positive or negative sentiments. We will first build a dataset that contains Arabic reviews, then we will build the model to be used as classifier that classify reviews, finally results will be evaluated.

The paper organized as follows: Section II presents background and related works of sentiment analysis for the Arabic language. Section III details the process of corpus construction. Section IV characterizes the language used in the corpus. Section V describes the methodology of approaching the sentiment classification. Section VI presents results of experiments. Section VII concludes the ideas of the whole research and give insight about the future works.

II. LITERATURE REVIEW

A. Sentiment Analysis

Sentiment analysis has multiple analytic levels: document level, sentence level, phrase level, and aspect level [13]. At document level, sentiment predicated for the whole document, dealing with the input as a single entity, with an output of one class. At sentence level, sentiment predicated for each sentence in a document, with an output class. At aspect level, SA considers a specific category of input entry. For example, an input contains a user experience of sport, so the predication would be on that experience whether it is positive or negative.

Sentiment class predication could be based on two levels of classifications: polarity classification (positive or negative) or multiclass task (5 stars classification) [7].

B. Machine Learning

Supervised learning technique is one of the most practical use of machine learning (ML) that has an input variable (X) and an output variable (Y). Thus, this could be used to teach the algorithm mapping function from the input to the output by:

$$Y = f(X) \quad (1)$$

This technique requires labelled data with correct answers, surveying as a teacher for the learning process [19].

To conduct a supervised learning for sentiment analysis, there is a sequence of steps to follow as converting the text to numerical data presented in vector space, which eases the data mapping with labels, also performing feature extraction and selection to train the ML model.

C. Sentiment Analysis for Arabic Language

In [4], researchers worked on informal Arabic classification by a look-up dictionary on a corpus of 68 negative, 40 positive, and 35 neutral tweets. Overall, the approach achieved 73% accuracy (2012).

Authors of [9] used two corpora of 348 negative, 403 positive, and 57 neutral tweets and comments. The approach based on 380 lexicon seeds, polarized with two methods:

sum polarity (single polarity) and double polarity (each word has positive and negative weight). Research experiments yielded the best accuracy with double polarity method by 83.8% on first dataset and 63% on the second (2013).

In [18], researchers worked on 500 negative and 500 positive tweets in the Egyptian dialect. They approached SA by using two machine learning classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), and Semantic Orientation (SO). These models combined with unigram and bigram features. While ML approach achieved 78.8% accuracy, SO achieved 75.9% accuracy (2012).

Other researchers in [8] adopted semi-supervised ML that enhanced with semantic and name entities tagging features, applied on a dataset of 40000 Arabic words. Classifiers are SVM, NB, Logistic Regression (LR), and Convolutional Neural Networks (CNNs). The ensemble learning model remarked the highest precision, recall and f-measure of 85.52%, 39.49%, 54.03% respectively (2012).

Moreover, contributors of [2] worked on four corpora of 513 negative, 613 positive, and 848 negative tweets. They conducted supervised ML using SVM, NB, and Decision Tree (D-Tree) models. Results achieved accuracies of: 68.05% for Obama dataset, 84.43% for Messi dataset, 68.06% for iPhone dataset, and 65.60% for Shia dataset (2013).

In [10], authors worked on seven corpora in multiple dialects by using supervised ML models: SVM and NB, which enhanced with features as emoticon, segment's number, and text length. The highest accuracy achieved 85.03% using SVM (2013).

Furthermore, researchers in [15] applied SVM classifier on a large corpus of 340000 tweets in Kuwaiti dialect. Further features enhanced the model including emoticon, part of speech tagging (POST), and n-gram variables. Classifier achieved 76% precision and 61% recall (2014).

In [3], researchers conducted a ML approach to classify 2026 tweets in health domain. Their approach enhanced with a combination of unigram and bigram features as well as term frequency–inverse document frequency (TF*IDF). Classifiers include NB, SVM, LR, Stochastic Gradient Descent (SGD), and CNNs. Experiments show that linear SVM using SGD outperforms all other classifiers by 85% to 91% accuracy (2017).

In addition, researchers in [5] conducted supervised ML to classify 24,028 reviews in hotel domain. Their approach enhanced with morphological, syntactical, and semantical features, namely: T1: Aspect Category Identification, T2: Opinion Target Expression Extraction, and T3: Sentiment Polarity Identification. The dataset consists of provided for training (19,226). SVM achieved the highest accuracy of 53% for T1, 59% for T2, and 19% for T3 (2018).

III. Dataset Construction

The dataset used for this research, collected from Google (https://maps.google.com) and Zomato (https://www.zomato.com/lebanon) on reviews' services. In this section, we will present how we collected the dataset.

A. *Data Collection*

The dataset collected using Google map and Zomato. First, we used an important feature within Google maps to notify costumer service providers inside Lebanon as a whole, including the five main Governorates and districts of Lebanon: Beirut, Mount Lebanon, Nabatiyeh, North Lebanon, and South Lebanon. The collected reviews are about 3500 from Google. Secondly, Zomato gives only access to the top five reviews for each service provider, so we have collected 416 reviews manually over a period from October 23, 2018 to November 22, 2018. Overall, the corpus has reached to 3916 reviews. The following table provides a sample from the corpus that represent each review with a client's rating:

rating	review	English Equivalent
2	هذا الفندق ينقصه بعض الاشياء داخل الغرف مثلا عدم وضوح القنوات التلفزيونية وجود اعطال في الحمامات ولا تستطيع وضع الملابس داخل الحمام الا على المغسلة وليس به وجبة افطار صباحية اما موقعه فهو جيد وقريب على اماكن التسوق وقريب جدا من المطار	This hotel lacks some things inside the rooms, for example, the lack of clearness on television channels, there is a problem in the bathrooms, and you cannot put the clothes inside the bathroom except on the laundry, also no breakfast, but its location is good and close to the shopping areas and very close to the airport
4	لطيف ولكن الغرف الفندقية تحتاج صيانة كادر الخدمة يجب ان يكون بالمستوى المطلوب	Nice but the hotel rooms need maintenance and the staff of services must be professional
5	مكان جميل جدا وحسن الخلق والضيافه	Very nice place, good attitude and hospitality
3	بحاجة الى اعادة تأهيل للمفروشات	Needs rehabilitation of furniture
5	فندق ممتاز ومعاملة راقية جدا	Excellent hotel and very upscale treatment
5	رائع وجميل	Wonderful and beautiful

Fig. 1. Corpus Sample

Corpus generated in Excel CSV (utf-8 Comma delimited) format. It contains 2313 5-stars and 734 4-stars, 418 3-stars, 148 2-stars, and 303 1-star reviews.

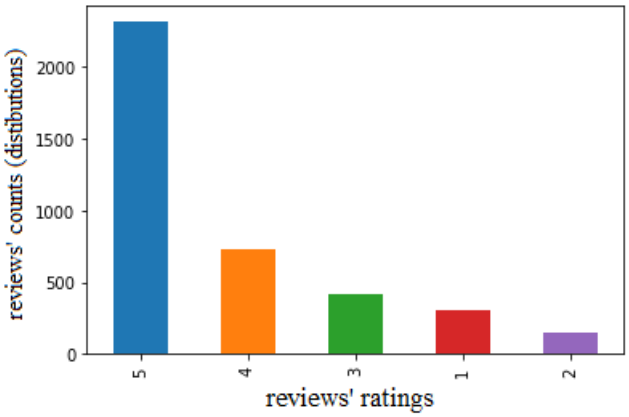

Fig. 2. Corpus Ratings' Distributions

IV. The Characteristics of Arabic Language Reviews

In the following, we presented few observed linguistic characteristics that occur frequently in Arabic reviews found in the collected corpus.

1) Exaggerations: Internet users tend to exaggerate while expressing their opinions in social media. Consider the following examples:

Arabic Review	Corrected Text	English Equivalent
روووووووووووووووووووووو وعه	روعه	Amazing
مكان حلو كتيييييييير	مكان حلو كثير	Very sweet place
جمييييل	جميل	Beautiful
ووووواووو	واو	Waw
راااائع	رائع	Wonderful
ييممم رائعه	يم رائعه	Yamy it is wonderful

Fig. 3. Arabic Exaggerations

2) Code Switching and Mixing: Code mixing, or switching is the alternating use by bilingual language users of two or more languages within an utterance [14].

Arabic Review	English Equivalent
Its recommended ولا اطيب	It is recommended the most delicious
veryyyyyy delecious plat du jour. كل يوم بقدموا 3 طبق رئيسيه .	Very delicious daily plate. Everyday they present 3 main plats
Generally not recommended at all! المطعم غير جيد بالمرة	Generally not recommended at all! The restaurant is not good at all
كتير شيك Updated the traditional Lebanese breakfast	Very chic updated the traditional Lebanese breakfast

Fig. 4. Arabic Code Mixing

3) Sarcasm: Social media is full of sarcasm. The algorithm does not differentiate between sarcastic and sincere sentences.

كول و شكور (😂)(😂 Eat and give your thanks)
طلوب اليوم وتعشا بكرا (Request today and dinners tomorrow)

Fig. 5. Arabic Sarcasm Examples

In the first example, the costumer may criticize the service because it has a negative value but it is still may regarded as positive. On the second example, the costumer is mocking the response delaying of a restaurant.

4) Question Sentences: question sentences have niether positive nor negative polarity.

هل الاسعار مناسبة أو مرتفعة (Is the prices affordable or high)
في غرفة؟ (Is there a room)

Fig. 6. Examples of Question Arabic Sentences

5) Opposite Representation: This type of sentences is based on linguistic challenges, which have negative lexical polarity but indicating positiviness.

ولا اطيب واهم فريق عمل (nothing is more delicious than that and important work team)
رائع ولا أطيب كتير طيب (wonderful and nothing is more delicious than that very delicious)

Fig. 7. Opposite Representation Examples

"ولا" is a negative word, but here it is an intensifier for the adjective "طيب"/delicious.

6) Sense Disambiguation: There are plenty of words (often called "homophones") are pronounced as same as many other words but differ in the way of spelling and meaning.

نايس وكريم (Nice and generous)
صاحب المطعم السيد كريم حمزة [...] يقدم لزبائنه اشهى الأطباق. (The owner of restaurant Kareem Hamzieh serves the most delicious plates for his costumes)

Fig. 8. Words Disambiguation

There is a distinction of meaning between the first and the second reviews in the use of كريم/Kareem. The first one indicates generosity the user experienced, while the second notes restaurant's owner first name.

7) Language Variation: From corpus observation, reviews generally written in Modern Standard Arabic (MSA) not in Arabic Lebanese dialect. The following table is a sample of language use:

Arabic Review	English Equivalent
مكان جميل جدا وحسن الخلق والضيافه	Very beautiful place and sincere hospitality and behaving
المكان جيد و لكن الأسعار إلى حد ما مبالغ بها	Good place but the prices are exaggerated somehow
مناظر خلابه واكل مميز ولذيذ واسعار مناسبة	Amazing views and special food and spicy and the affordable prices
افضل فندق و يوجد فيه مطعم وغرف و مصلى	Best hotel and it has restaurant and rooms and a place to pray
جيد جدا واشكر القائمين على ادارة المشفى	Very good and thanx to those who responsible for managing the hospital

Fig. 9. Corpus Language Variety

V. METHODOLOGY

In this section, we will present the methodology of approaching sentiment classification for machine learning.

A. Sentiment Assignment

a) Sentiment assigned by Reviews' values: Positive sentiment of "+1" assigned to reviews with values of 5, 4 and 3, while reviews with values of 2 and 1 assigned to a negative sentiment of "0". As it is shown in the figure below, the "0" sentiment bar contains 451 reviews of negative polarity, while the bar with "1" sentiment contains 3465 reviews of positive polarity.

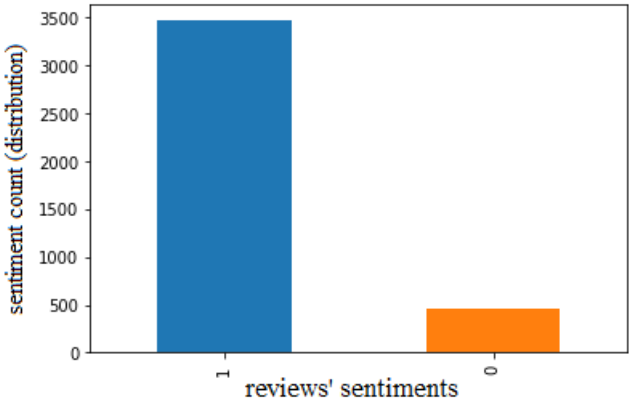

Fig. 10. Corpus Sentiment Distribution

B. *Data split*

Data splitting is necessary for machine learning and testing phases, so we randomly split the corpus in our hand to 80% for training set and 20% for testing set. The training set is 3132 examples with 353 negative and 2779 positive reviews. On the other hand, the testing set is 784 examples with 98 negative and 686 positive reviews.

C. *Feature Extraction*

In this research, we approached SA task with term frequency–inverse document frequency feature (TF*IDF), which is the product of two frequencies tf and idf (Manning. 2009). For a word to have high TF*IDF in a document, it must appear a lot of times in a document and must be absent in the other documents. It must be a signature word of the document [12].

$$tf * idf,d = tft,d \times idft \quad (2)$$

tf is basically the output of the bag-of-words (BoW) model. For a specific document, it determines how important a word is by looking at how frequently it appears in the document. If a word appears a lot of times, then the word must be important. Moreover, idf for a word to be considered a signature word of a document, it shouldn't appear that often in the other documents. Thus, a signature word's document frequency must be low, meaning its inverse document frequency must be high.

The total amount of extracted TF*IDF features are 7095 words. The following table represents an example of 6 top features with smallest and largest coefficients.

Words with smallest Coefficients	Words with largest Coefficients
غير (Non)	جميل (beautiful)
سيئة (bad)	رائع (wonderful)
مش (not)	لذيذ (delicious)
سيئ (bad)	طيب (spicy)
صفر (zero)	رائعة (wonderful)
لا (no)	مميز (special)

Fig. 11. TF*IDF feature with largest and smallest coefficients

D. *Logistic Regression*

As we are going to classify reviews in a positive or negative class, LR adopted because of high efficacy in binary classification tasks. LR uses a threshold boundary to isolate the positive reviews from the negative ones. LR uses a Logistic function to estimate probabilities between positive or negative label y and data features w given by input x. Thus, LR, as denoted in equation (3), uses sigmoid function to get the likelihood directly by minimizing infinitive +∞ and -∞ into a scale between 0 to 1 [11].

$$p(y = \pm 1|x, w) = \frac{1}{1 + e^{-w^T h(x)}} \quad (3)$$

where the probability of the review to be equal to positive class (+1) or negative class (-1) depends on the input (x) given its learnt coefficient (w). The equation computed by 1 divided by 1 plus exponential (e) powered to the dot product of transpose learnt coefficient (w) of each feature (h) in the given input (x).

The following presents the squeezing of sigmoid function for sentiment probability in which "0" represents negative, "+1" represents the positive, while "0.5" represents the uncertainty neither of both sentiments.

$$y = \begin{cases} 0, & x < 0 \\ 0.5, & x = 0 \\ +1, & x \geq 0 \end{cases} \quad (4)$$

Hence, LR is feed with reviews as input (x), sentiment as the target (y), TF*IDF as feature of input (h(x)), and features' coefficients as (w). In addition, precision and recall, Receiver operating characteristics (ROC) curve used for evaluating the performance of LR on test data.

1) Precision and Recall: Because every mistake is costly and effective for the whole process of classification, a one may simply prefer to reduce the percentage of false positives to be less than, say, 3.5% of all positive predictions. This is where precision comes in. A complementary metric is recall, which measures the ratio between the number of true positives and that of (ground-truth) positive reviews.

2) Receiver Operating Characteristics (ROC) curve: It is a metric for binary classification, which considers all possible thresholds. Various thresholds result in different true positive/false positive rates. As you decrease the threshold, you get more true positives, but also more false positives.

VI. RESULTS

Figure (12) below shows the macro, micro and weighted averages of precision and recall obtained when applying LR. The macro-precision for the classifier is equal to 0.84. The classifier gave slightly better results when predicting positive reviews, and that is well returned to the large percentage of positive reviews. Also, the table illustrates that the values of recall for positive predication is high, while it is too low when predicting negative reviews. Thus, this results in low macro-average 0.54%.

	precision	recall
0	0.80	0.08
1	0.88	1.00
micro avg	0.88	0.88
macro avg	0.84	0.54

Fig. 12. Precision and Recall Values

Figure (13) shows the ROC curve the classifier achieved by drawing True Positive (TP) on the y-axis and False Positive (FP) on the x-axis. The curve is too close to the middle nearby the red-cut line, which is not good as it supposed to be closer to the upper left. The area under the curve scores 0.54 that indicates a fail-grade classifier.

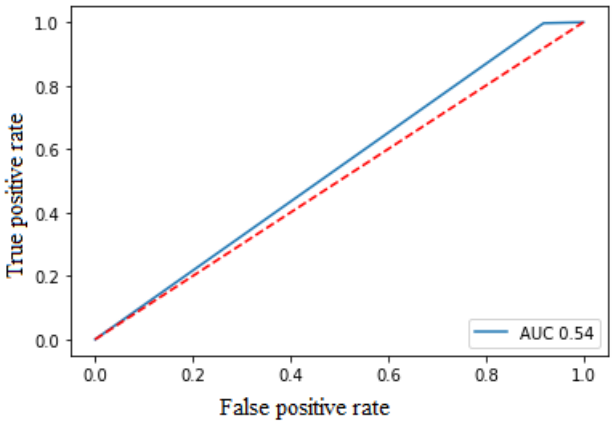

Fig. 13. Receiver operating characteristics (ROC) curve

VII. Conclusion and Future Works

This research has tackled a highly important task of sentiment analysis for Arabic language in the Lebanese context on 3916 reviews' services from Google and Zomato. Experiments show three main findings: 1) The classifier is confident when used to predict positive reviews, 2) while it is biased on predicting reviews with negative sentiment, and finally 3) the low percentage of negative reviews in the corpus contributes to the diffidence of LR.

Definitely, we would expand further into the research for including much more reviews' services in the Lebanese context for sharpening the performance of the classifier. And, we would expand our research to include more reviews in other regions as Syria, Jordan and Palestine. In addition, diverse ML classifiers and deep learning ones would be conducted to reach to the highest performance.

References

[1] Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). Sentiment Analysis Using Common-Sense and Context Information. Computational Intelligence and Neuroscience., vol. 2015, pp. 1–9.

[2] Ahmed, S., Pasquier, M. and Qadah, G. (2013). Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text. Proceedings of the International Conference on Innovations in Information Technology (IIT) Abu Dhabi, United Arab Emirates, pp. 72-77.

[3] Alayba, A., Palade, V., England, M. and Iqbal, R. (2017). Arabic Language Sentiment Analysis on Health Services. Proc. 1st International Workshop on Arabic Script Analysis and Recognition (ASAR '17), pp. 114-118.

[4] Albraheem, L. and Al-Khalifa H. (2012). Exploring the problems of Sentiment Analysis in Informal Arabic. Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, Bali, Indonesia, pp. 415-418.

[5] AL-Smadi, M., Qawasmeh, O., Talafha, B., Al-Ayyoub, M., Jararweh, Y., & Benkhalifa, E. (2016). An enhanced framework for aspect-based sentiment analysis of hotels reviews: Arabic reviews case study. International conference for internet technology and secured transactions (ICITST-2016). IEEE, pp. 98–103.

[6] Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), pp. 345–354.

[7] Collomb, A, Costea, C., Joyeux, D., Hasan, O., & Brunie, L. (2013). A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. University of Lyon, INSA-Lyon, Villeurbanne, France, F-69621.

[8] Elarnaoty, M., AbdelRahman, S., and Fahmy, A. (2012). A Machine Learning Approach for Opinion Holder Extraction in Arabic Language. Proceedings of the International Journal of Artificial Intelligence & Applications, abs/1206.1011.

[9] El-Beltagy, S. and Ali, A. (2013). Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study. Proceedings of the 9th International Conference on Innovations in Information Technology (IIT), Abu Dhabi, United Arab Emirates, pp. 215-220.

[10] El-Beltagy, S., Khalil, T., Halaby, A., and Hammad, M. (2013). Combining Lexical Features and a Supervised Learning Approach for Arabic Sentiment Analysis. A. Gelbukh (Ed.): CICLing,2016, Part I, LNCS 7816, Springer-Verlag, Berlin, pp. 89-97.

[11] Logistic Regression. (n.d.). Retrieved December 12, 2018, from http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html

[12] Manning, C. D., Raghavan, P., & Schütze, H. (2009). An Introduction to Information Retrieval (Online ed.). Retrieved October 27, 2018, from https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

[13] Moralwar, S., & Deshmukh, S.. (2015). Different Approaches of Sentiment Analysis. International Journal of Computer Sciences and Engineering, Volume-3, Issue-3, pp. 160-165, E-ISSN: 2347-2693.

[14] Muysken, P. (2000). Bilingual speech: A typology of code-mixing. Cambridge University Press.

[15] Salamah, J. and Elkhlifi, A. (2014). Microblogging Opinion Mining Approach for Kuwaiti Dialect. Proceedings of the International conference on Computing Technology and Information Management, Dubai, UAE, pp. 388-396.

[16] Salem, F. (2017). The Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World (Vol. 7). Dubai: MBR School of Government.

[17] Shaalan, K., Siddiqui, S., & Monem, A. (2016). Sentiment Analysis in Arabic. NLDB 2016, LNCS 9612, pp. 409–414, doi: 10.1007/978-3-319-41754-7_41.

[18] Shoukry, A. and Rafea, A. (2012). Preprocessing Egyptian Dialect Tweets for Sentiment Mining. In: 4th Workshop on Computational Approaches to Arabic Script-Based Languages, pp. 47–56.

[19] Shukla, S. (2017). Regression and Classification | Supervised Machine Learning. Retrieved August 17, 2018, from https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning.

[20] Vinodhini, G., & Chandrasekaran, R. (2017). A sampling based sentiment mining approach for e-commerce applications. Information Processing & Management, 53(1), pp. 223–236.

[21] Yoo, K. H., & Gretzel, U. (2008). What motivates consumers to write online travel reviews? Information Technology & Tourism, 10(4), pp. 283–295.