Arabic Sentiment Analysis on Chewing Khat Leaves using Machine Learning and Ensemble Methods

Wael M.S. Yafooz
Computer Science Department
Taibah University
Madinah Munawarah, Saudi Arabia
wyafooz@taibahu.edu.sa

Essa Abdullah Hezzam
Information Systems Department
Taibah University
Madinah Munawarah, Saudi Arabia
essa_alkadasi@yahoo.com

Waseem Alromema
Information Systems Department
Taibah University
Madinah Munawarah, Saudi Arabia
wromema@taibahu.edu.sa

Abstract-Sentiment analysis plays an important role in obtaining speakers' opinions or feelings towards events, products, topics, or services, helping businesses to improve their products. Moreover, governments and organizations investigate and solve current social issues by analyzing perspectives and feelings. This study evaluated the habit of chewing Khat (qat) leaves among the Yemeni society. Chewing Khat plant leaves, is a common habit in Yemen and East Africa. This paper proposes a model to detect information about the Khat chewing habit, how people explore it, and the preference for Khat leaves among Arabic people. A dataset consisting of user comments on 18 youtube videos was prepared through several natural language processing techniques. Several experiments were conducted using six machine learning classifiers and four ensemble methods. Support Vector Machine and Linear Regression had almost 80% accuracy, whereas xgboot was the most accurate ensemble method reaching 77%.

Keywords-sentiment analysis; machine learning; classification; ensemble methods

I. INTRODUCTION

Nowadays, the study of user opinions has attracted substantial attention in social perspectives, focusing on services, products, and habits through many data mining applications, recommender systems, and business intelligence applications. The analysis and interpretation of user opinions are altogether known as sentiment analysis, which is an area of natural language processing, also kwon as the voice of the customer in business intelligence [1, 2]. Business owners need to be aware of feedback to improve future performance. Such a time-consuming and difficult task is used in the analysis of huge unstructured data gathered through social media or internet comments. Several studies have been conducted, classifying sentiments as positive, negative, or neutral [3, 4, 6]. More complex sentiment analysis [7-10], often referred to as fine-grained, classify datasets into five classes, namely very positive, positive, neutral, negative, and very negative. Moreover, aspect-based sentiment analysis [11-14] classifies datasets by extracting entities from text.

Users' comments are often an outcome of their opinions, and they can be considered as the main factor in evaluating services or products. Some studies focused on education [15, 16], while other researches focused on detecting health misinformation on social media users [18, 19]. Khat is a type of plant that pleases and stimulates, and chewing Khat leaves is a commonly seen habit in Yemen and East Africa [21-23]. Although it is customary in these countries, several experiments showed its direct impact on human organs. This paper presents a model to study consumers' opinions regarding the habit of chewing Khat leaves in Yemeni and East African society. At first, the dataset was collected by extracting user comments from 18 youtube videos. The annotation process classified the data into positive and negative fractions. Several NLP processes were executed to prepare the data for Machine Learning Classifiers (MLCs) and Ensemble Methods (EMs).

II. RELATED WORKS

Several studies utilized sentiment analysis in different ways. Multilingual student comments, obtained through student feedback, were used to evaluate online courses' effectiveness and teachers' performance in [3, 15-17]. In [3], the dataset was collected using approximately 4000 student comments through surveys conducted on 25 university courses to evaluate the performance of a professor who had been teaching for 10 years, while the sentiment analysis was directed including positive, negative, and eight more emotions. Similarly, authors in [15] proposed a system to evaluate a lecturer's performance by collecting data through student surveys via a rating system in a form of numerical data. The MLC Naïve Bays was employed to predicate the positive and negative students' sentiments toward the lectures. A recurrent neural network of long and short term memory in deep learning was utilized in [16]. The dataset was collected from 3000 positive, negative, and neutral student comments on 30 courses. The performance improved when using the softmax activation function, reaching 89%, 99%, and 90% during training, testing, and validation, respectively. Deep learning was applied on a course evaluation dataset with 3000 student comments using three predefined classes in [17], while the results showed that relu and softmax performed better.

Sentiment analysis is used to identify the main factors affecting the success of businesses, particularly start-ups. In [1], user comments were extracted from Twitter using topic modeling and applying supervised vector machine learning to divide comments into three main classes. The textual analysis was applied based on the entities trained in the previous phase

Corresponding author: Wael M.S. Yafooz

using Nvivo software. In [4], an analysis of a massive amount of user comments (approximately 1.6 million) from the Yelp Challenge Dataset was conducted. The dataset was divided to 20% for testing and 80% for training, using four machine learning classifiers. The best accuracy rate reached 92.6% and 92.3% under Stochastic Gradient Descent and Linear Support Vector Classification respectively. Similarly, the same dataset was utilized in [2] to analyze restaurant reviews through a hybrid classifier ensemble method using Naïve Bayes, Support Vector Machines, and Genetic Algorithms.

Some health sector studies have also been conducted [18, 19]. Authors in [18] focused on tweets on breast cancer, collecting user comments from approximately 845 cancer patient accounts with 48,000 posts. The logistic regression classifier and a Convolutional Neural Network was utilized in the process, and the model's performance accuracy was 97.6%. Besides, it was found that positive experiences had more shares, providing more awareness to the general public. Descriptive statistics of text mining and topic modeling were utilized in [19]. Unstructured data from 3 million news articles on Reuters assisted in identifying the 10 major health issues published in news articles from 2007 to 2017. On the contrary, the analysis of user reviews on mobile health applications was prioritized in [8], collecting data from 104 mobile health applications with approximately 88,125 user reviews. The data were categorized based on each comment's functionality (such as usability, content, customer support, and ethics), the polarity concept was divided into three classes, and five machine classifiers were applied. The best accuracy was recorded at 89.42% through Stochastic Gradient Descent.

III. METHODS

This section describes the main model phases, as shown in Figure 1. There are four phases: data acquisition, pre-processing, machine learning classifiers, and model evaluation.

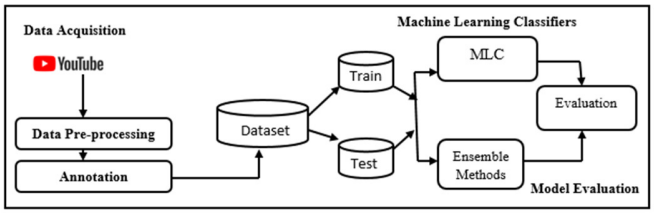

Fig. 1. Model architecture.

A. Phase 1: Data Acquisition

The dataset was collected using Python 3.8 programming language and YouTube API (googleapiclient package), for information extracting from 18 videos related to chewing Khat. The criteria for selecting videos were: published date between 2015-2020, more than 50K views, more than 10K likes, and focus on Arabic speakers. Moreover, some keywords were used to locate the videos, such as Khat, Khat is dangerous, and disadvantages of Khat. The main attributes for the extracted video information were: commenter_id, commenter_name, comment, video_id, number of views, number of likes, and date. Table I shows the dataset description and the minimum and maximum length of user comments.

TABLE I. DATASET DESCRIPTION

Items	Description	Max length	Min length	Average
Negative	1436	427	1	17
Positive	1296			
Total	2732			

The initial step of data preprocessing was carried out, removing English or duplicate comments. The next step, data annotation, was a manual process conducted with the assistance of three annotators that were Ph.D. holders, Arabic native speakers, and computer science specialists. Data annotation classified the comments into negative and positive. Some unrelated, unclear, or ambiguous comments were removed. If two annotators classified comments as either positive or negative then comments were considered respectively, otherwise, the comments were removed.

B. Phase 2: Pre-processing

The natural language pre-processing steps were: data cleaning, tokenization, normalization of Arabic words, lemmatization, deletion of special characters, and removal of repeating characters. Then, the annotation was performed by three annotators into positive and negative. These pre-processing steps increased accuracy by removing "TSHKEEL", "TATWEEL", and "HAMZAH" using Python 3.6 and a package called “tashaphyne”.

C. Phase 3: Machine Learning Classifiers

Two types of MLCs were used: classic MLCs and Ensemble Methods (EMs). The MLCs were Linear Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), K-nearest Neighbor (KNN), Stochastic Gradient Descent (SGD), and Decision Tree (DT). The EMs were Random Forest (RF), Adaboost (ADA), Gradient Booting (BG), and xgboost (XG).

D. Phase 4: Model Evaluation

The model's performance was verified using Precision (1), Recall (2), F-Score (3), Accuracy (4), and 5-fold cross-validation on the dataset.

$$\text{Precision} = \frac{\text{Retrieved and Relevant Documents}}{\text{All Retrieved Documents}} \quad (1)$$

$$\text{Recall} = \frac{\text{Retrieved and Relevant Documents}}{\text{All Relevent Documents}} \quad (2)$$

$$\text{F} - \text{Score} = 2\ \text{X}\ \frac{\text{Precision X Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{Number of correct predications}}{\text{Total number of predications}} \quad (4)$$

IV. RESULTS AND DISCUSSION

This section presents the experiments and the results. Two experiments were conducted on the dataset: classic MLCs and EMs. In both experiments, the dataset was divided into 70% for training and 30% for testing, while 5-fold cross-validation was applied.

A. Classic Machine Learning Classifiers

As mentioned above, six classic MLCs were used. The n-gram with Unigram, Bigram, and Trigram was used with all six classifiers to examine their performance. Table II shows the results using Unigram in the six classifiers. It can be observed

that the SVM classifier had the highest performance accuracy (80.12%), while the lowest accuracy was noted on KNN (65%).

TABLE II. CLASSIFIERS' PERFORMANCE USING UNIGRAM

MLCs	Class	Precision	Recall	F-score	Accuracy
LR	Negative	84%	78%	81%	79.39%
	Positive	75%	81%	78%	
NB	Negative	63%	78%	70%	71.95%
	Positive	81%	68%	74%	
SVM	Negative	84%	79%	81%	80.12%
	Positive	76%	82%	79%	
KNN	Negative	94%	60%	73%	65.00%
	Positive	34%	84%	49%	
SGD	Negative	81%	77%	79%	77.80%
	Positive	74%	79%	76%	
DT	Negative	74%	71%	73%	71.34%
	Positive	68%	72%	70%	

Table III shows the MLCs performance using bigram. SVM had the highest accuracy (79.76%), whereas the lowest performance was noted on KNN (65.24%).

TABLE III. CLASSIFIERS' PERFORMANCE USING BIGRAM

MLCs	Class	Precision	Recall	F-score	Accuracy
LR	Negative	83%	78%	80%	78.90%
	Positive	75%	80%	77%	
NB	Negative	63%	78%	70%	72.07%
	Positive	82%	68%	74%	
SVM	Negative	83%	79%	81%	79.76%
	Positive	76%	81%	78%	
KNN	Negative	91%	61%	73%	65.24%
	Positive	37%	81%	51%	
SGD	Negative	79%	78%	79%	77.80%
	Positive	76%	78%	77%	
DT	Negative	77%	69%	73%	70.49%
	Positive	64%	72%	68%	

Table IV shows the MLCs performance using trigram. The highest accuracy was 79.51% using SVM, whereas the lowest was noted again for KNN (65.12%). Figure 2 depicts the overall MLCs results for Unigram, Bigram, and Trigram. Although SVM had the highest accuracy, it was followed closely by both LR and SGD at almost 80%. NB's and DT's accuracies were near 70%, whereas KNN was less accurate.

TABLE IV. CLASSIFIERS PERFORMANCE USING TRIGRAM

MLCs	Class	Precision	Recall	F-score	Accuracy
LR	Negative	83%	78%	80%	78.90%
	Positive	75%	80%	77%	
NB	Negative	63%	78%	69%	71.59%
	Positive	81%	67%	73%	
SVM	Negative	83%	78%	81%	**79.51%**
	Positive	76%	81%	78%	
KNN	Negative	91%	61%	73%	65.12%
	Positive	37%	80%	51%	
SGD	Negative	76%	78%	77%	76.71%
	Positive	77%	75%	76%	
DT	Negative	78%	67%	72%	69.15%
	Positive	59%	72%	65%	

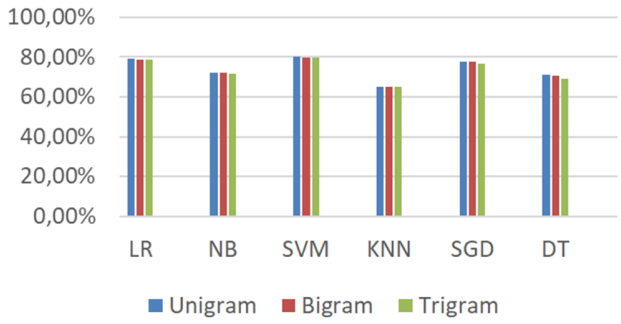

Fig. 2. Accuracy comparison for 6 MLCs on three n-grams.

B. Ensemble Methods

The four mentioned above common methods were used. Table IV shows the accuracy of these EMs. The highest accuracy was recorded for XG, while the lowest was noted for GB. Figure 3 demonstrates the accuracy of the DT classifier using Unigram, Bigram, and Trigram compared to RF. As it can be noted, RF outperformed DT.

TABLE V. ENSEMBLE CLASSIFIERS' PERFORMANCE

EMs	Class	Precision	Recall	F-score	Accuracy
RF	Negative	83%	73%	78%	75%
	Positive	68%	79%	73%	
ADA	Negative	80%	75%	77%	75%
	Positive	71%	77%	74%	
GB	Negative	81%	73%	77%	74%
	Positive	68%	77%	72%	
XG	Negative	82%	76%	79%	77%
	Positive	72%	79%	75%	

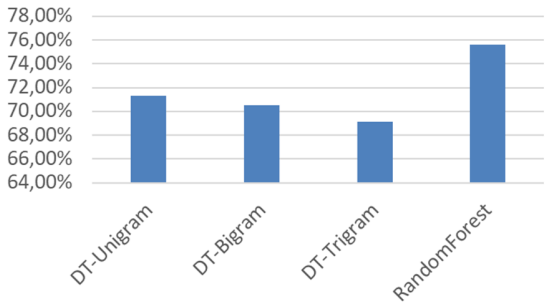

Fig. 3. Accuracy comparison between Decision Tree and Random Forest.

V. CONCLUSION

This paper presented a study on users' opinions on chewing Khat in Yemen and East Africa, using a dataset collected from YouTube comments. Several natural language processing steps were carried on the dataset to get the best performance using classifiers. Classic MLCs and Ems were applied. The best performance in terms of accuracy was recorded when using SVM, followed by Linear Regression. The best accuracy using EMs was recorded for XG.

REFERENCES

[1] J. R. Saura, P. Palos-Sanchez, and A. Grilo, "Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining,"

Sustainability, vol. 11, no. 3, Jan. 2019, Art. no. 917, https://doi.org/10.3390/su11030917.

[2] M. Govindarajan, "Sentiment analysis of restaurant reviews using hybrid classification method," in *Proceedings of 2nd IRF International Conference,* Chennai, India, Feb. 2014, pp. 127–133.

[3] S. Rani and P. Kumar, "A Sentiment Analysis System to Improve Teaching and Learning," *Computer*, vol. 50, no. 5, pp. 36–43, May 2017, https://doi.org/10.1109/MC.2017.133.

[4] A. Salinca, "Business Reviews Classification Using Sentiment Analysis," in *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Timisoara, Romania, Sep. 2015, pp. 247–250, https://doi.org/10.1109/SYNASC.2015.46.

[5] U. P. Gurav and S. Kotrappa, "Sentiment Aware Stock Price Forecasting using an SA-RNN-LBL Learning Model," *Engineering, Technology & Applied Science Research*, vol. 10, no. 5, pp. 6356–6361, Oct. 2020, https://doi.org/10.48084/etasr.3805.

[6] J. Carrillo-de-Albornoz, J. R. Vidal, and L. Plaza, "Feature engineering for sentiment analysis in e-health forums," *PLOS ONE*, vol. 13, no. 11, 2018, Art. no. e0207996, https://doi.org/10.1371/journal.pone.0207996.

[7] M. Madhukar and S. Verma, "Hybrid Semantic Analysis of Tweets: A Case Study of Tweets on Girl-Child in India," *Engineering, Technology & Applied Science Research*, vol. 7, no. 5, pp. 2014–2016, Oct. 2017, https://doi.org/10.48084/etasr.1246.

[8] O. Oyebode, F. Alqahtani, and R. Orji, "Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews," *IEEE Access*, vol. 8, pp. 111141–111158, 2020, https://doi.org/10.1109/ACCESS.2020.3002176.

[9] S. Angelidis and M. Lapata, "Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 17–31, Aug. 2018, https://doi.org/10.1162/tacl_a_00002.

[10] Z. Wang, C. S. Chong, L. Lan, Y. Yang, S. B. Ho, and J. C. Tong, "Fine-grained sentiment analysis of social media with emotion sensing," in *2016 Future Technologies Conference (FTC)*, Dec. 2016, pp. 1361–1364, https://doi.org/10.1109/FTC.2016.7821783.

[11] J. Luo, S. Huang, and R. Wang, "A fine-grained sentiment analysis of online guest reviews of economy hotels in China," *Journal of Hospitality Marketing & Management*, vol. 30, no. 1, pp. 71–95, Jan. 2021, https://doi.org/10.1080/19368623.2020.1772163.

[12] C. Yang, H. Zhang, B. Jiang, and K. Li, "Aspect-based sentiment analysis with alternating coattention networks," *Information Processing & Management*, vol. 56, no. 3, pp. 463–478, May 2019, https://doi.org/10.1016/j.ipm.2018.12.004.

[13] M. Song, H. Park, and K. Shin, "Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean," *Information Processing & Management*, vol. 56, no. 3, pp. 637–653, May 2019, https://doi.org/10.1016/j.ipm.2018.12.005.

[14] W. Xue and T. Li, "Aspect Based Sentiment Analysis with Gated Convolutional Networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, Jul. 2018, vol. 1, pp. 2514–2523, https://doi.org/10.18653/v1/P18-1234.

[15] F. F. Balahadia, M. C. G. Fernando, and I. C. Juanatas, "Teacher's performance evaluation tool using opinion mining with sentiment analysis," in *2016 IEEE Region 10 Symposium (TENSYMP)*, May 2016, pp. 95–98, https://doi.org/10.1109/TENCONSpring.2016.7519384.

[16] I. A. Kandhro *et al.*, "Sentiment Analysis of Students' Comment by using Long-Short Term Model," *Indian Journal of Science and Technology*, vol. 12, no. 8, pp. 1–16, Feb. 2019, https://doi.org/10.17485/ijst/2019/v12i8/141741.

[17] I. A. Kandhro, S. Z. Jumani, F. Ali, Z. U. Shaikh, M. A. Arain, and A. A. Shaikh, "Performance Analysis of Hyperparameters on a Sentiment Analysis Model," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 6016–6020, Aug. 2020, https://doi.org/10.48084/etasr.3549.

[18] E. M. Clark *et al.*, "A Sentiment Analysis of Breast Cancer Treatment Experiences and Healthcare Perceptions Across Twitter," *arXiv e-prints*, vol. 1805, p. arXiv:1805.09959, May 2018.

[19] M. Zolnoori *et al.*, "Mining news media for understanding public health concerns," *Journal of Clinical and Translational Science*, pp. 1–10, Oct. 2019, https://doi.org/10.1017/cts.2019.434.

[20] F. Saeed, W. M.S. Yafooz, M. Al-Sarem, and E. A. Hezzam, "Detecting Health-Related Rumors on Twitter using Machine Learning Methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, 2020, https://doi.org/10.14569/IJACSA.2020.0110842.

[21] A. Al-Alimi, E. Halboub, A. K. Al-Sharabi, T. Taiyeb-Ali, N. Jaafar, and N. N. Al-Hebshi, "Independent determinants of periodontitis in Yemeni adults: A case-control study," *International Journal of Dental Hygiene*, vol. 16, no. 4, pp. 503–511, 2018, https://doi.org/10.1111/idh.12352.

[22] M. Hijazi, H. Jentsch, J. Al-Sanabani, M. Tawfik, and T. W. Remmerbach, "Clinical and cytological study of the oral mucosa of smoking and non-smoking qat chewers in Yemen," *Clinical Oral Investigations*, vol. 20, no. 4, pp. 771–779, May 2016, https://doi.org/10.1007/s00784-015-1569-2.

[23] M. A. Al-Duais and Y. S. Al-Awthan, "Association between qat chewing and dyslipidaemia among young males," *Journal of Taibah University Medical Sciences*, vol. 14, no. 6, pp. 538–546, Dec. 2019, https://doi.org/10.1016/j.jtumed.2019.09.008.

[24] B. Kalakonda, S. A. Al-Maweri, H.-M. Al-Shamiri, A. Ijaz, S. Gamal, and E. Dhaifullah, "Is Khat (Catha edulis) chewing a risk factor for periodontal diseases? A systematic review," *Journal of Clinical and Experimental Dentistry*, vol. 9, no. 10, pp. e1264–e1270, Oct. 2017, https://doi.org/10.4317/jced.54163.