Journal of Computer Science

Original Research Paper

Sentimental Analysis on Health-Related Information with Improving Model Performance using Machine Learning

Wael M.S. Yafooz and Abdullah Alsaeedi

Department of Computer Science, College of Computer Science and Engineering, Taibah University, Saudi Arabia

Article history
Received: 05-01-2021
Revised: 18-02-2021
Accepted: 20-02-2021

Corresponding Author:
Wael M.S. Yafooz
Department of Computer Science, College of Computer Science and Engineering, Taibah University, Saudi Arabia
Email: waelmohammed@hotmail.com

Abstract: Social media platforms are extensively used in exchanging and sharing information and user experience, thereby resulting in massive outspread and viewing of personal experiences in many fields of life. Thus, informative health-related videos on YouTube are highly perceptible. Many users tend to procure medical treatments and health-related information from social media particularly from YouTube when searching for chronic illness treatments. Sometimes, these sources contain misinformation that cause fatal effects on the users' health. Many sentimental analyses and classifications have been conducted on social media platforms to study user post and comments on many life science fields. However, no study has been conducted on the analysis of Arabic user comments, which provide details on herbal treatments for people with diabetes. Therefore, this study proposes a model to detect and discover emotions/opinions of YouTube users on herbal treatment videos is proposed through an analysis of user comments by using machine learning classifiers. In addition, a new Arabic Dataset on Herbal Treatments for Diabetes (ADHTD), which is based on user comments from several YouTube videos, is introduced. This study examines the impact of four representation methods on ADHTD to show the performance of machine learning classifiers. These methods remove repeating characters in Arabic dialect and character extension known as 'TATAWEEL' or 'MAD', stemming of Arabic words, Arabic stop words removal and N-grams with Arabic words. Experiments has been conducted based aforementioned methods to handle imbalanced proposed dataset and identify the best machine learning classifiers over Arabic dialect textual data. The model has achieved a higher accuracy that reached 95% when using Synthetic Minority Oversampling TEchnique (SMTOE) techniques to balanced dataset than imbalanced dataset.

Keywords: Sentiment Analysis, N-gram, Support Vector Machine, Logistic Regression, SMOTE

Introduction

Social media platforms occupy a large part of our daily life activities. They allow users to exchange information in seconds. YouTube is the most popular video platform, generating billions of views through the uploaded video content. More than two billion logged-in users visit YouTube every month, allowing people to share their perspectives on daily activities, thoughts, experiences, advertisements and educational resources (Bhuiyan *et al*., 2017; Burns *et al*., 2020). YouTube facilitates the users in liking, commenting and sharing ideas. The advancing technology has made YouTube easily accessible to users. Thus, it has become popular among kids, adults and the elderly crowd as a mode of education and entertainment. However, users face difficulties in differentiating the desirable and undesirable content due to the massive unstructured data on YouTube (Awal *et al*., 2018; Chen *et al*., 2017).

As a solution for this, researchers utilize natural language processing methods, such as text mining and sentiment analysis through machine learning or deep learning approaches in ranking and analyzing the best suitable content through user comments and the number of views and likes (Awal *et al*., 2018; Choi and Segev, 2020; Dabas *et al*., 2019; Tahir *et al*., 2019; Vedula *et al*., 2017). These comments can indicate user perspectives, emotions and opinions against the content that can be positive, neutral, or negative using sentiment analysis. Sentiment analysis is the process of extracting and discovering user

opinions. It can be useful for service improvement and to obtain user feedback on products and services. Besides, it can also provide users with the best approach when making choices between different video content on YouTube in agreement with their requirements as the content desirable for kids, etc. (Tahir *et al*., 2019; Alghowinem, 2018).

Sentiment analysis is carried out using machine learning approaches. Especially the supervised classification techniques. The most popular Machine Learning Classifiers (MLCs) that are used in the process are Naiva Bays (NB), Decision tree, Random Forest (RF), k-Nearest Neighbours (KNN), Support Vector Machine (SVM) and Logistic Regression (LR). These MLCs are used in classifying data into two parts; positive or negative. This classification is carried out based on the previous training process on the dataset called the training dataset. The accuracy of the model performance depends on the training dataset and the pre-processing steps. The pre-processing steps that is to use in dataset preparation can easily include in English language textual datasets. A lot of scholarly attention is given to the user comments on social media in other dialects than in English.

A lot of studies were conducted on sentiment analysis upon many domains such as advertisement classifications (Vedula *et al*., 2017; Gauba *et al*., 2017; Chauhan and Meena, 2019), business intelligence (feedback on products) (Aufar *et al*., 2020; Tafesse, 2020), education (Veletsianos *et al*., 2018; Lee *et al*., 2017; Anggraini and Tursina, 2019; Thelwall and Mas-Bleda, 2018) and the quality of videos dedicated to kids (Tahir *et al*., 2019; Alghowinem, 2018; Araújo *et al*., 2017). However, only a few studies were conducted on health-related misinformation. Many YouTube videos on disease treatments and health information are usually created through individual experiences. Thus, the content in such videos can mislead the followers while raising health issues among them. Most of the scholarly studies are focused on user opinions (sentiment analysis) and on detecting the rumours/misinformation on health-related issues through social media platforms, such as Twitter and Facebook (Daabes and Kharbat, 2019; Oksanen *et al*., 2015; Sicilia *et al*., 2017; Alayba *et al*., 2017; Alsaeedi and Khan, 2019; Alsaeedi, 2019).

To the best of my knowledge, there had been no studies focusing on building a dataset on herbal treatment for diabetes in the Arabic dialect utilizing the user comments on YouTube videos. Therefore, this study addresses two issues namely; sentiment analysis on Herbal Treatments for Diabetes in YouTube content and improving the performance of MLCs on Arabic datasets using sentiment analysis case study. Therefore, this study proposes a model to detect the user opinion on herbal treatment on YouTube videos related to diabetes based on user comments by using the MLCs. In addition, a new dataset called *Arabic Dataset on Herbal Treatments for Diabetes* (ADHTD) is introduced. ADHTD dataset includes 21,320 user comments from 22 YouTube videos on herbal treatments for diabetes. After removing the noise, data cleaning and pre-processing, the remaining 4,111 comments were divided into two classes (positive and negative) through the human annotation process. Additionally, this study examine the impact of four representation methods on ADHTD to show the model performance. These methods remove repeating characters in Arabic dialect and characters extension known as 'TATAWEEL' or 'MAD', stemming of Arabic words, Arabic stop words removal and N-grams with Arabic words. Furthermore, handle imbalance dataset and identify the best MLCs in terms of accuracy. The experimental findings show that the stop word removal method cannot effectively use on model performance comparing with the removal of repeating characters, character extension and stemming of Arabic words that will decrease the model performance. Besides, LR and SVM outperform KNN, DT and NB. The worst accuracy in all experiments recorded with KNN. The model has achieved a higher accuracy level that reached 95% when using *Synthetic Minority Oversampling TEchnique* (SMTOE).

The rest of this paper organized as follows: Section 2 presents the related studies. The methods, model architecture and ADHTD dataset is explains in section 3. Section 4 includes the discussion and results of the proposed model while the conclusion of this paper in section 5.

Related Studies

Many literary studies had been conducted on sentiment analysis on social media networks. These studies focus on the study of user opinions, known as sentiment analysis. These studies can also be categorized as feedback on products (Ray and Chakrabarti, 2017; Aufar *et al*., 2020; Tafesse, 2020; Veletsianos *et al*., 2018; Fang and Zhan, 2015; Nguyen *et al*., 2015; Alghowinem, 2018), education (Lee *et al*., 2017; Anggraini and Tursina, 2019; Thelwall and Mas-Bleda, 2018; Chauhan and Meena, 2019; Relucio and Palaoag, 2018), health misinformation (Daabes and Kharbat, 2019; Oksanen *et al*., 2015; Alayba *et al*., 2017; Chua and Banerjee, 2017; Song *et al*., 2019, construction of Arabic datasets (Al Mukhaiti *et al*., 2017; Elnagar *et al*., 2018; Al-Horaibi and Khan, 2016; Al-Rubaiee *et al*., 2016) and information quality for kids (Tahir *et al*., 2019; Alghowinem, 2018; Araújo *et al*., 2017). In literary reviews, a less attention is given to health-related issues, especially on Arabic language. In this study, we focus more on health-related issues and Arabic sentiment analysis.

Authors in (Qi *et al*., 2017) focuses on the analysis of rumours about cancer on the internet by collecting data from two websites. The experiment had been conducted with the assistance of doctors, nurses and students specialise in medicine to classify and verify data into two

classes namely, dread and wish. The ANOVA statistical method was used due to the diminutiveness of the dataset. In the same way, (Song *et al*., 2019) conducted experiments by collecting health information from china's website. The dataset of 872 was verified by 218 participations. It was found that dread health rumours are more credible than wish rumours. While (Sicilia *et al*., 2017), proposed a model to detect health related rumours on twitter about Zika virus with a dataset of 800, which was classified into rumours and non-rumours. It focuses on feature extraction based on three levels of influencing the users and network by using random forest classifier. The model performance of approximately 89% was achieved in detecting rumours.

Alayba *et al*. (2017) introduced a new Arabic dataset on health-issues by extracting the related information from user comments on Twitter, applying several machine learning algorithms and deep learning methods using Convolutional Neural Networks (CNN) on the imbalanced dataset. The experiment shows that the model performance achieved approximately a success rate of 85 to 90%. However, the issue of the imbalanced dataset was not considered. Besides, the deep learning concept was applied to a small dataset. Daabes and Kharbat (2019) demonstrated the importance of the study of Arabic health information through cancer-based YouTube videos by assembling the video specifications such as the number of likes, comments, views and the amount of dislikes.

Al-Tamimi *et al.* (2017) carried out sentiment analysis on Arabic a language from YouTube user comments. Approximately 5,986 comments were collected and annotated into three classes. The dataset was collected from the top-rated videos on YouTube and its unbalanced dataset. Several machine learning algorithms with model performance were applied which reached a success rate of 88.8%. Najadat and Abushaqra (2018) proposed a Multimodal sentiment analysis on twenty-one Arabic videos collected from YouTube. Five machine learning models were applied to the voice and facial features of the person where the model performance achieved 76%.

An Arabic dataset based on hotel reviews was introduced by (Elnagar *et al*., 2018). This imbalanced dataset contained 373772 reviews which can be materialized in further research functions. Six types of machine learning classifiers were applied on this balanced and imbalanced dataset, along with three experiments namely; Polarity, rating and lexicon classification.

Al-Horaibi and Khan (2016) constructed a general dataset from Twitter with selected Naïve Bayes and decision trees as machine learning classifiers While measuring the sentiment analysis on the Arabic language. It is mentioned that Naïve Bayes had the best accuracy rate of 64.85% while the decision tree has an accuracy rate of 53.75%.

Al-Rubaiee *et al.* (2016). introduced a dataset with 1121 tweets to study student emotions on e-learning. They applied the two machine learning classifiers Naïve Bayes and support vector machine with four experiments on two to three classes. The best accuracy rate had been noticed when Naïve Bayes classifier on two classes rating 84.62%. Even though the dataset was small (Heikal *et al*., 2018) used the Deeping learning CNN and Long Short-Term Memory (LSTM) while utilizing the methods of ensemble to obtain the best accuracy rate which was at 64.46%. The experience was on ASTD: Arabic Sentiment Tweets Dataset (Nabil *et al*., 2015).

Methods and Model Architecture

This section presents the methods and model architecture of this research. The model architecture consists of seven phases namely; data collection, data cleaning, data pre-processing, annotation, feature extraction, build model and model evaluation as shown in Fig. 1.

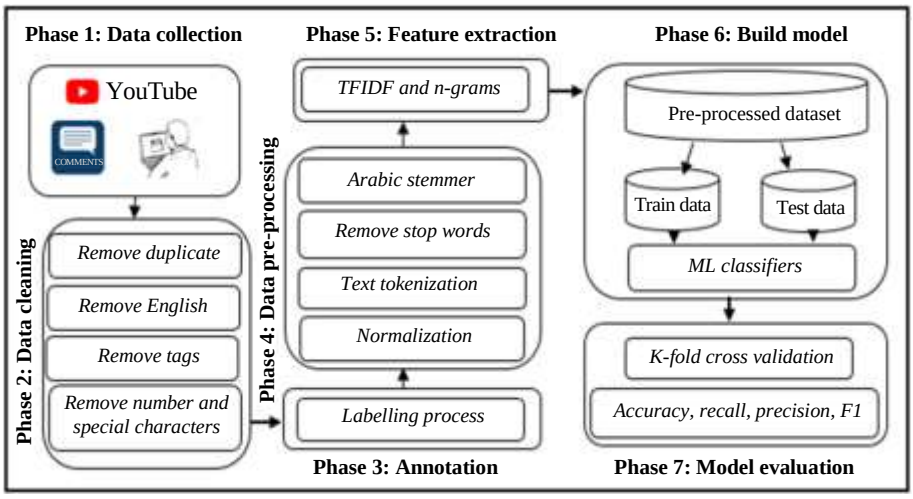

Fig. 1: Model architecture

Data Collection

This study aims to analyze the user opinions against the herbal treatments for diabetes based on YouTube content. Such datasets need to be prepared and collected. This initial step of data collection is known as Phase 1. In phase 1, the criterias of the data collection are the videos published between 2015-2020 with more than 100K views, 5000 likes and 2000 comments. These are published by people, professional doctors, herbal treatment experts and the users of such medical substances.

Data was collected from 22 YouTube videos. Besides, the focus was given to the comments by the speakers of Arabic dialect. Some keywords were used in order to retrieved videos such as Herbal treatments for diabetes "علاج السكرى بلاعشاب", diabetes cure " علاج مرض السكرى",, cure from diabetes without medications " علاج السكرى بدون ادوية. Python version 3.8 and YouTube API was utilized to extract the user comments on the videos. Figure 2 presents the sample of extracted comments from a YouTube Video.

There are several steps to carry out before combining the extracted comments into a single file. A total of 22 files were allocate for each comments extracted YouTube video. After the said steps, the dataset was named as The *Arabic Dataset on Herbal Treatments for Diabetes* (ADHD). This was followed by the data cleaning process.

Data Cleaning

This is the Phase 2. Upon data collection and storage, the following data cleaning steps was carried out to remove the noise:

- Duplicate comments
- English sentences [*a-z*, *A-Z*]
- Whitespaces
- Words with length size = one
- HTML tags <,>, br, </,/>
- Numbers and special characters [0-9], [@,#,&,!,.,;(,),<,>,-,..,%,*]

Then, the annotation process will start to identify the positive or negative user comments.

Annotation Process

In Phase 3, the annotating process is a human-based, manual process conducted with assistance from three native Arabic speakers as annotators. If at least two annotators agreed that comments are positive, they consider being positive, failing which it will consider negative. If there is any ambiguity from the annotators, the comments are not to be removed or considered.

The user comments with negative and positive denotations were inspected and verified, reducing the data set from 21362 to 4111 comments. Hence 1013 comments were positive and 3098 were negative, ADHD is considered an imbalanced dataset because the number of positive comments is greater than the number of negative comments. Table 1 describes the balanced dataset and the original imbalanced dataset with attributes of classes that are positive, negative, including the maximum and minimum length of characters in comments.

Table 1: ADHTD Description

Items	Description
Positive class	3098
Negative class	1013
Total number comments	4111
Maximum length of character	696
Minimum length of character	1

	Comment	Class
0	جزاك الله خير الجزاء يا شيخ	1
1	يالشيخ جزاك الله خيرا في حالة استخدام وصفه الحسل وصارت حالته هبوط مثلا العسل هل استخدم السكر او الحسل	1
2	يالشيخ جزاك الله خير فرد علي عادي استخدم حليب ولبن مزروع السمسم اثناء استخدام الخلطه حليب في حالت هبوط السكر مثلا العسل مع العلم السكر نوع اول بتي عيرها ١١	1
3	جزاك الله فيك يابيدي الفاضل ❤️🌹 راح تجربها ارجوك سؤال هل الوصفه مع الاستمرار في اخذ دواء السكري او الاقتصاع عنه اثناء اخذها وشكرا	1
4	شخذا الله يحفظك هل يمكن اخذ وصفه الحسل بعد الكل لان اني لما شاهدها على الريق بتعص بالصب شديد	1
...	...	...
3995	معلومات غير دقيقه من دكتوره لاتعرف معنى الفاعليه الصح يحدد سماح نصائها	0
3996	هذه المراه جاهله تضع معلومات غير صائبه غير صالحه في شهر 12 اكتشفت انه عندي السكري وكان 500 والحمد لله الآن في الثامنات والتسعينات و بعد الاكل في ساعتين 120 واكل الملح البحري والتحمه والدهون وامتنعت عن جميع التشوربات والسكر والدهون الشايه امتنع عنها نهائيا لا تصدقوا هذه جاهله	0
3997	دكتوره ماتعرف الملح بالسكر	0
3998	الا بالموت اهسن كل شي ممنوع تعيش بـ قرفه و اوراق ؟؟؟	0
3999	غير صحيح القرنفل ضار لمرضى السكري قاعمه واحده تحتوي حوالي 17 غرام من الكربوهيدرات هذه كميه هائله	0

4000 rows × 2 columns

Fig. 2: Extracted comments

Based on the preprocessed ADHTD dataset, the pre-processing step is carried out in phase 4.

Data Pre-Processing

In Phase 4, the data pre-processing step is performed. It consists of four steps. They are normalization, stop word removal, tokenization and Arabic stemming. Python version 3.8 was utilized in carrying out the aforesaid steps. The words were returned to the normal form through normalization. The Arabic stop words were removed using the NLTK package while tokenization was used in separating the words. Also, the following condition had been considered which are frequently appear in Arabic dialect comments. Such preprocessing steps help in improving the MLCs, particularly with Arabic comments:

- Remove the repeated character such as " رااااائع" to " رائع"
- Remove the character extension in Arabic language such as "جميـــــــل" to "جميل"

For instance: Repeated character

Before (Normal)	After
ان الفيديو رااااائع و مفيد جداااا	ان الفيديو رائع و مفيد جدا
long character (extend)	
Before (Normal)	After
المعلومـــــــــات مفيـــــــــــــدة لكل مريض ســـــــــكر	المعلومات مفيدة لكل مريض سكر

Feature Extraction and N-Gram

In Phase 5, after preprocessing steps performed in phase 4, the extracted words/terms from comments represent a form of n-grams. The n-gram was applied to five forms of unigram, bigram, trigram, 4-grams and 5-grams while the best results were achieved with bigram and trigram which means no performance can be noticed after that. Then, the n-grams converted numeric values using "CountVectorizer" and "sklearn.feature_extraction.text" by python version 3.8 program language. The term frequency-inverse document was applied through "TfidfTransformer" in the "sklearn" package. The TFIDF algorithm as in (1):

$$W(w,C) = TF(w)C\ Log\frac{N}{CF(T)} \tag{1}$$

Where:

$TF(w)\ C$ = Denotes number of word (w) in comment (C)
CFt = Denotes number of comments containing word (w)
N = Denotes is the total number of comments in dataset

The output of this phase is used as input for MLCs in phase 6.

Build Model

In phase 6, the most popular MLCs were selected (Pranckevičius and Marcinkevičius, 2017; Maxwell *et al*., 2018; Saharudin *et al*., 2020). The implementation was executed using the "sklearn" package by Python 3.8. These MLCs are; NB, DT, RF, KNN, SVM and LR. The results were analyzed, discussed and presented in the next section. The dataset had been divide into three phases for training and testing. They are, Phase 1: 70% tranning and 30% testing data, phase 2: 60% training and 40% testing and phase 3: 80% training and 20% testing. The results did not show a major difference in the model performance in the three phases. The results present in this research are based on 70% training and 30% for testing.

Model Evaluation

In phase 7, the MLCs are evaluated. There are two types of the evaluation conducted on the dataset and performance of MLCs. First, cross-validation is conducted to evaluate the quality of the dataset to void the over-fitting problem. The cross-validation splits data into 3, 5 and 10 folds. The result shows the accuracy is to test data (MLCs performance), as shown in Table 1 with into 5 folds. Secondly, the performance MLCs in term of accuracy has been evaluated using Precision (2), Recall (3) and F1-score (4) and accuracy (5):

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FP} \tag{3}$$

$$F1 - score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where, TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

Experiments, Results and Discussion

This section presents the experiment settings and the results. There are two types of experiments has been carried out are; Experiments with word representation methods and Experiments with Handling Imbalanced Dataset.

Experiments with Word Representation Methods

The experiment has been carried out to examine the four word representation methods on extracted features to determine its impact on the performance of the proposed model. These methods remove repeating

characters in the Arabic dialect and characters extension known as 'TATAWEEL' or 'MAD', stemming of Arabic words, Arabic stop words removal and N-grams with Arabic words. They also determine the best MLCs and handle the imbalance dataset. The MLCs that has been included NB, DT, RF, KNN, SVM and LR. In all experiments, the cross validation process was carried out with five folds to avoid the over-fitting problem as shown in Table 2. In addition, unigram, bigram, trigram and 4-gram were applied in all experiments. There are two types of experiments has been carried out.

In the first Experiment (E1), these aforementioned methods were not considered. The results of this Experiment (E1) are shown in Table 3. The model performance in terms of accuracy using VSM and LR reached 90 and 89%, respectively and the worst accuracy was 69% using KNN.

In the second Experiment (E2), remove repeating characters in Arabic dialect and characters extension which is known as 'TATAWEEL' or 'MAD', stemming were considered. The in stemming words revert to original form, the repetition of character is a common form of writing comments in Arabic dialect particularly in social media and the character extension is usually used by Arabic commenters. In this experiment, the performance of MLCs has improved compared with that in (E1), as shown in Fig. 2-5 with all forms of N-gram, Unigram, Bigram, Trigram and 4-gram, respectively.

Table 2: Results of 5-Folds cross-validation

MLCs	CV 1 (%)	CV 2 (%)	CV 3 (%)	CV 4 (%)	CV 5 (%)	Mean (%)
KNN	69.5	70.4	71.1	69.4	66.0	69.3
SVM	87.1	90.8	89.5	87.4	87.4	88.4
NB	70.4	72.1	72.8	72.1	73.5	72.2
LR	88.8	89.8	89.8	88.1	88.4	89.0
DT	84.2	79.9	81.3	80.2	83.4	81.7

Table 3: MCLs with not word representation methods

MLCs	Class	Precision	Recall	F-score	Accuracy (%)
KNN	0	0.64	0.42	0.51	69.00
	1	0.71	0.85	0.77	
SVM	0	0.69	0.86	0.77	90.00
	1	0.96	0.9	0.93	
NB	0	0.86	0.47	0.61	72.00
	1	0.68	0.94	0.79	
LR	0	0.64	0.88	0.74	**89.00**
	1	0.97	0.89	0.93	
DT	0	0.25	0.8	0.39	80.00
	1	0.98	0.8	0.88	

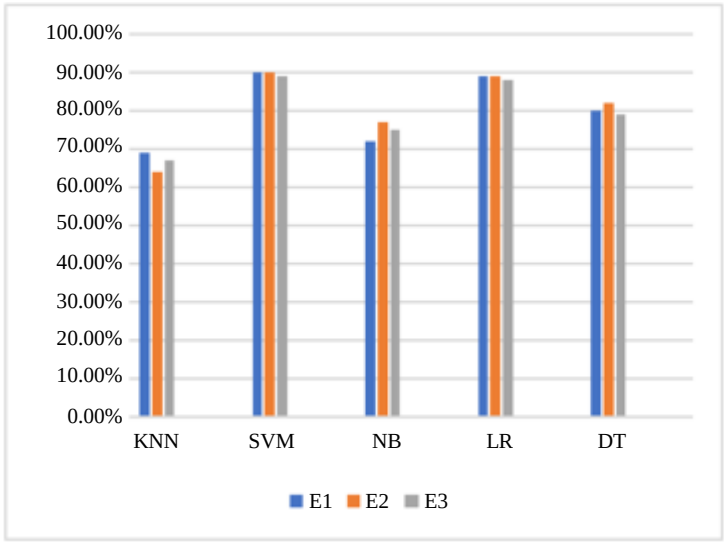

Fig. 2: Accuracy of model with Unigram

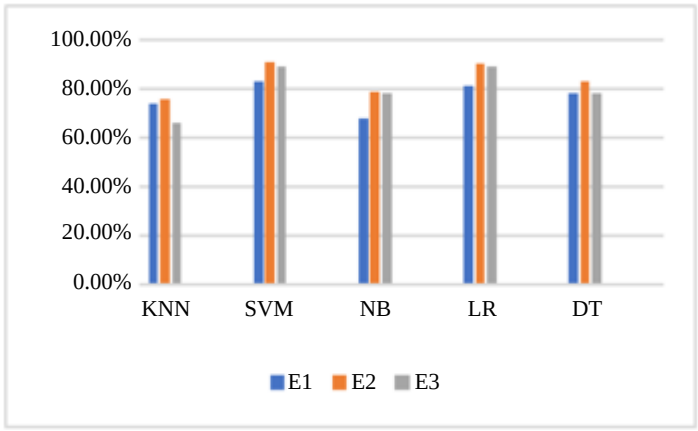

Fig. 3: Accuracy of model with Bigram

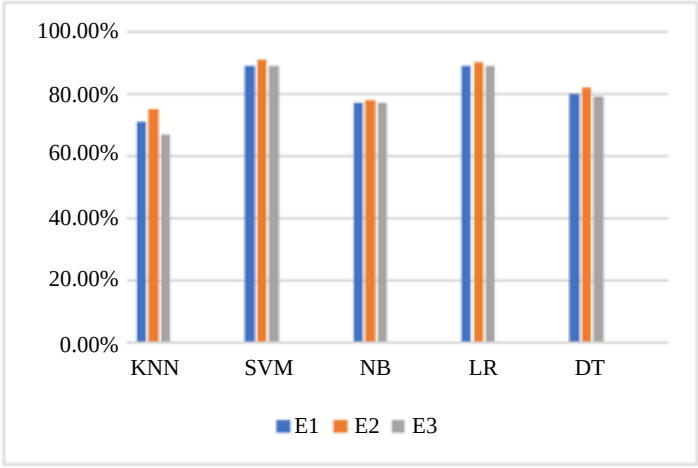

Fig. 4: Accuracy of Model with Trigram

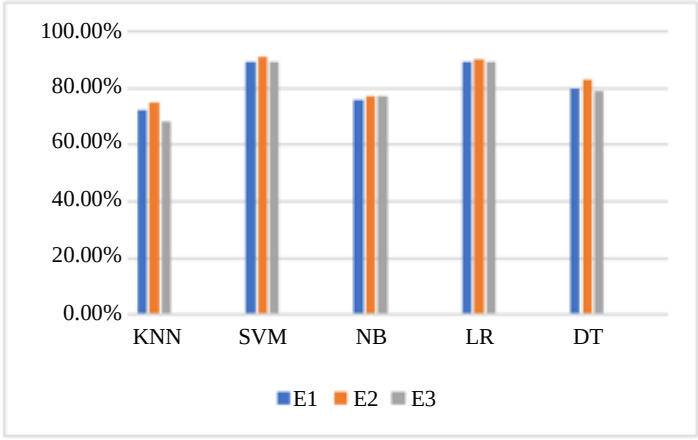

Fig. 5: Accuracy of model with 4-Gram

In addition in E2, VSM and LR had the best accuracy. The experiment results show that the model performance in terms of accuracy in trigram and 4-gram is the same. It reached 91 and 90% for VSM and LR, respectively. These results are better than those of Unigram and Bigram, as shown in Fig. 2-5. However, the accuracy of KNN decreased to 64, 65, 67 and 67 with Unigram, Bigram, Trigram and 4-gram, respectively, as shown in Fig. 2-5.

Table 4: Balanced Dataset (Using SMOTE)- Unigram 5-Fold Cross-validation

MLC	CV 1 (%)	CV 2 (%)	CV 3 (%)	CV 4 (%)	CV 5 (%)	Mean (%)
KNN	70.16	70.16	71.74	70.59	69.67	70.46
SVM	93.09	93.55	95.62	95.39	94.58	94.44
NB	86.41	86.29	88.70	87.54	88.47	87.48
LR	89.40	91.24	91.35	91.35	89.97	90.66
DT	85.83	85.48	87.54	84.54	86.39	86.10

Table 5: Accuracy of MCLs with imbalanced dataset

	Using oversampling		Using under sampling		Using SMOTE	
MLCs	Unigram (%)	Trigram (%)	Unigram (%)	Trigram (%)	Unigram (%)	Trigram (%)
KNN	74.34	77.84	72.53	70.39	71.11	75.15
SVM	94.73	94.84	87.01	89.31	94.94	95.27
NB	84.88	85.42	81.25	84.54	87.90	89.03
LR	91.39	92.04	87.17	90.46	92.58	92.85
DT	86.18	85.31	81.91	82.40	85.85	86.23

In the third Experiment (E3), where the stop word removal in the NLTK of Arabic dialect is used, no significant improvement in the model performance given the slight decrease in all four classifiers, except for KNN, which shows slight improvement in Unigram from 64 to 67% and Bigram from 65% to 66 but no improvement and decrease in accuracy in Trigram and 4-gram.

Overall, the performance of MLCs in terms of accuracy in E2 and E3 outperform E1. This improvement was achieved due to use of the method for removing repeating characters in Arabic dialect and characters extension known as 'TATAWEEL' or 'MAD', stemming in E2. Such methods can improve the performance of the MLCs, especially when handling Arabic dialect textual data extracted from social media. While using stop word removal for Arabic textual data it does not show any improvement. Furthermore, using N-grams method with methods of removing repeating characters in the Arabic dialect and the characters extension, stemming of Arabic words, the accuracy of the model has shown improvement with bigram and trigram while the accuracy improvement has not been noticed using 5-grams.

Experiments with Handling Imbalanced Dataset Techniques

In this research, the dataset that is considered is imbalanced. Therefore, imbalanced dataset techniques were applied to the imbalanced dataset. These techniques were oversampling, under sampling and SMTOE (Jeatrakul *et al*., 2010). These techniques were used to make the two classes in each class close to each other in the number of comments.

In all three techniques, the cross-validation process has been conducted and results of 5-Folds cross-validation based on the SMOTE technique shown in Table 4. In these experiments, the n-gram of unigram and trigram was utilized. The results of MLCS using oversampling, under sampling and SMOTE techniques shown in Table 5. The best accuracy level had been achieved by SMOTE (trigram) using the two classifiers SVM and LR with a rate of 94.94 and 92.58% respectively in unigram and a rate of 95.27 to 92.85% respectively in the trigram.

In using SMOTE technique, the performance of the model has been significant improved with MLCs except with KNN were noticed the decrease of accuracy. In under sampling the number of comments is reduced from majority class (high number of comments/rows in positive class) to the number of comments to minority class (low number of comments/rows in negative class. In opposite, oversampling which randomly duplicates comments to reach a close number of comments in the majority class. While SMOTE used a different way of increasing the minority class to majority class. It calculate the average distance between the data points in space to increase the number of comments (data point in space) to reach the size of the majority class. Therefore, this method is achieved a good accuracy compared with under sampling or oversampling.

The main contribution of this study as follows:

1. Introduce a new dataset on diabetes herbal treatment in the Arabic dialect that can be useful to many researchers in data mining or related research areas
2. Examine the performance of MCLs on the proposed dataset. The experiments' results demonstrate that MLCs, namely SVM and LR had the best accuracy rate and outperformed DT, KNN and NB in terms of accuracy
3. Evaluate the four methods on the Arabic text to improve the performance of MLCs. These methods remove repeating characters in Arabic dialect and character extensions known as 'TATAWEEL' or 'MAD', stemming of Arabic words, Arabic stop

words removal and N-grams with Arabic words. The experimental results show that there is an improvement in model performance when using remove repeating characters in Arabic dialect and characters extension known as 'TATAWEEL' or 'MAD', stemming of Arabic words. No record improvement using Arabic stop words removal were found. This result can be seen with unigram, bigram, trigram, 4-grams and 5-grams. Generally, E2, bigram and trigram achieve a better accuracy rate compared to unigram and 5-grams. While results with 4-grams are the same

4. Evaluate SMOTE with N-grams on the proposed imbalanced Arabic dataset. Experiments results show that SMOTE outperformed under and upper sampling methods with all MLCs except the KNN, the model performance reducer in accuracy
5. This model used to detect health-related issues is based on user comments on YouTube videos, particularly diabetes videos

Conclusion

A lot of misinformation is spread over the digital world, such information can impact on people if related to health. Diabetes herbal treatment via YouTube has been detected through this study by analysis user opinions based on user comments. Besides, this study examines the impact of four representation methods on Arabic datasets to evaluate the model performance. Therefore, based on the finding of this study using the mentioned methods can improve the models or research that will be conducted on data mining based on textual data. This study examines the impact of four representation methods on Arabic datasets to evaluate the model performance. The experiments results show that the model performance can be improved by remove repeating characters in Arabic dialect and characters extension known as 'TATAWEEL' or 'MAD', stemming of Arabic words, while using Arabic stopword removal, the model performance is slight increase with KNN only compared with another machine learning classifiers. Furthermore, the results demonstrates that the VSM and LR outperform all the machine learning classifiers in all experiments while the worst is using KNN. The best accuracy level was recorded when applying SMOTE on the imbalanced dataset which achieved a rate of 94 to 92% using SVM and LR, respectively compare with under sampling or oversampling techniques.

Author's Contributions

Wael M.S. Yafooz: Proposed the idea, carried out the experiments and Writing the manuscript.

Abdullah Alsaeedi: Corrections and help in the writing.

Ethics

The authors confirm that this article has not been published in any other journal. The corresponding author confirms that all the authors have read and approved the manuscript and there are no ethical issues involved.

References

Al Mukhaiti, A. J. S., Siddiqui, S., & Shaalan, K. (2017, September). Dataset built for Arabic sentiment analysis. In International Conference on Advanced Intelligent Systems and Informatics (pp. 406-416). Springer, Cham. https://doi.org/10.1007/978-3-319-64861-3_38

Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2017, April). Arabic language sentiment analysis on health services. In 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR) (pp. 114-118). IEEE. https://doi.org/10.1109/ASAR.2017.8067771

Alghowinem, S. (2018, September). A safer youtube kids: An extra layer of content filtering using automated multimodal analysis. In Proceedings of SAI Intelligent Systems Conference (pp. 294-308). Springer, Cham. https://doi.org/10.1007/978-3-030-01054-6_21

Al-Horaibi, L., & Khan, M. B. (2016, July). Sentiment analysis of Arabic tweets using text mining techniques. In First International Workshop on Pattern Recognition (Vol. 10011, p. 100111F). International Society for Optics and Photonics. https://doi.org/10.1117/12.2242187

AL-Rubaiee, H. S., Qiu, R., Alomar, K., & Li, D. (2016). Sentiment analysis of Arabic tweets in e-learning. Journal of Computer Science. https://doi.org/10.3844/jcssp.2016.553.563

Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 10(2), 361-374. https://doi.org/10.14569/IJACSA.2019.0100248

Alsaeedi, A. (2019). EFTSA: Evaluation Framework for Twitter Sentiment Analysis. JSW, 14(1), 24-35. https://doi.org/ 10.17706/jsw.14.1.24-35

Al-Tamimi, A. K., Shatnawi, A., & Bani-Issa, E. (2017, October). Arabic sentiment analysis of youtube comments. In 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) (pp. 1-6). IEEE. https://doi.org/10.1109/AEECT.2017.8257766

Anggraini, N., & Tursina, M. J. (2019, November). Sentiment Analysis of School Zoning System On Youtube Social Media Using The K-Nearest Neighbor With Levenshtein Distance Algorithm. In 2019 7th International Conference on Cyber and IT Service Management (CITSM) (Vol. 7, pp. 1-4). IEEE. https://doi.org/10.1109/CITSM47753.2019.8965407

Araújo, C. S., Magno, G., Meira, W., Almeida, V., Hartung, P., & Doneda, D. (2017, September). Characterizing videos, audience and advertising in Youtube channels for kids. In International Conference on Social Informatics (pp. 341-359). Springer, Cham. https://doi.org/10.1007/978-3-319-67217-5_21

Aufar, M. andreswari, R., & Pramesti, D. (2020, August). Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study. In 2020 International Conference on Data Science and Its Applications (ICoDSA) (pp. 1-7). IEEE. https://doi.org/10.1109/ICoDSA50139.2020.9213078

Awal, M. A., Rahman, M. S., & Rabbi, J. (2018, October). Detecting Abusive Comments in Discussion Threads Using Naïve Bayes. In 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET) (pp. 163-167). IEEE. https://doi.org/10.1109/ICISET.2018.8745565

Bhuiyan, H., Ara, J., Bardhan, R., & Islam, M. R. (2017, September). Retrieving YouTube video by sentiment analysis on user comment. In 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA) (pp. 474-478). IEEE. https://doi.org/10.1109/ICSIPA.2017.8120658

Burns, L. E., Abbassi, E., Qian, X., Mecham, A., Simeteys, P., & Mays, K. A. (2020). YouTube use among dental students for learning clinical procedures: A multi-institutional study. Journal of dental education. https://doi.org/10.1002/jdd.12240

Chauhan, G. S., & Meena, Y. K. (2019). YouTube Video Ranking by Aspect-Based Sentiment Analysis on User Feedback. In Soft Computing and Signal Processing (pp. 63-71). Springer, Singapore. https://doi.org/10.1007/978-981-13-3600-3_6

Chen, Y. L., Chang, C. L., & Yeh, C. S. (2017). Emotion classification of YouTube videos. Decision Support Systems, 101, 40-50. https://doi.org/10.1016/j.dss.2017.05.014

Chua, A. Y., & Banerjee, S. (2017). To share or not to share: The role of epistemic belief in online health rumors. International journal of medical informatics, 108, 36-41.

Choi, S., & Segev, A. (2020). Finding informative comments for video viewing. SN Computer Science, 1(1), 47. https://doi.org/10.1007/s42979-019-0048-2

Daabes, A. S. A., & Kharbat, F. F. (2019). A content analysis of Arabic YouTube videos for cancer treatment. International Journal of Health Governance. https://doi.org/10.1108/IJHG-05-2019-0035

Dabas, C., Kaur, P., Gulati, N., & Tilak, M. (2019, November). Analysis of Comments on Youtube Videos using Hadoop. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 353-358). IEEE. https://doi.org/10.1109/ICIIP47207.2019.8985907

Elnagar, A., Khalifa, Y. S., & Einea, A. (2018). Hotel Arabic-reviews dataset construction for sentiment analysis applications. In Intelligent Natural Language Processing: Trends and Applications (pp. 35-52). Springer, Cham. https://doi.org/10.1007/978-3-319-67056-0_3

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 5. https://doi.org/10.1186/s40537-015-0015-2

Gauba, H., Kumar, P., Roy, P. P., Singh, P., Dogra, D. P., & Raman, B. (2017). Prediction of advertisement preference by fusing EEG response and sentiment analysis. Neural Networks, 92, 77-88. https://doi.org/10.1016/j.neunet.2017.01.013

Heikal, M., Torki, M., & El-Makky, N. (2018). Sentiment analysis of Arabic Tweets using deep learning. Procedia Computer Science, 142, 114-122. https://doi.org/10.1016/j.procs.2018.10.466

Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010, November). Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In International Conference on Neural Information Processing (pp. 152-159). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17534-3_19

Lee, C. S., Osop, H., Goh, D. H. L., & Kelni, G. (2017). Making sense of comments on YouTube educational videos: a self-directed learning perspective. Online Information Review. https://doi.org/10.1108/OIR-09-2016-0274

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing, 39(9), 2784-2817. https://doi.org/10.1080/01431161.2018.1433343

Nabil, M., Aly, M., & Atiya, A. (2015, September). ASTD: Arabic sentiment tweets dataset. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 2515-2519). https://doi.org/10.18653/v1/D15-1299

Najadat, H., & Abushaqra, F. (2018). Multimodal sentiment analysis of Arabic videos. Journal of Image and Graphics, 6(1), 39-43. https://doi.org/10.18178/joig.6.1.39-43

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42(24), 9603-9611. https://doi.org/10.1016/j.eswa.2015.07.052

Oksanen, A., Garcia, D., Sirola, A., Näsi, M., Kaakinen, M., Keipi, T., & Räsänen, P. (2015). Pro-anorexia and anti-pro-anorexia videos on YouTube: Sentiment analysis of user responses. Journal of Medical Internet Research, 17(11), e256. https://doi.org/10.2196/jmir.5007

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing, 5(2), 221. https://doi.org/10.22364/bjmc.2017.5.2.05

Qi, J., Banerjee, S., & Chua, A. (2017). Analyzing medical personnel's perceptions of online health rumors. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1). http://www.iaeng.org/publication/IMECS2017/IMECS2017_pp457-460.pdf

Ray, P., & Chakrabarti, A. (2017, February). Twitter sentiment analysis for product review using lexicon method. In 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) (pp. 211-216). IEEE. https://doi.org/10.1109/ICDMAI.2017.8073512

Relucio, F. S., & Palaoag, T. D. (2018, January). Sentiment analysis on educational posts from social media. In Proceedings of the 9th International Conference on E-Education, E-Business, E-Management and E-Learning (pp. 99-102). https://doi.org/10.1145/3183586.3183604

Saharudin, S. N., Wei, K. T. & Na, K. S. (2020). Machine Learning Techniques for Software Bug Prediction: A Systematic Review. Journal of Computer Science, 16(11), 1558-1569. https://doi.org/10.3844/jcssp.2020.1558.1569

Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2017, November). Health-related rumour detection on Twitter. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1599-1606). IEEE. https://doi.org/10.1109/BIBM.2017.8217899

Song, X., Zhao, Y., Song, S., & Zhu, Q. (2019). The role of information cues on users' perceived credibility of online health rumors. Proceedings of the Association for Information Science and Technology, 56(1), 760-761. https://doi.org/10.1002/pra2.165

Tafesse, W. (2020). YouTube marketing: how marketers' video optimization practices influence video views. Internet Research. https://www.emerald.com/insight/content/doi/10.1108/INTR-10-2019-0406/full/html

Tahir, R., Ahmed, F., Saeed, H., Ali, S., Zaffar, F., & Wilson, C. (2019, August). Bringing the kid back into YouTube kids: detecting inappropriate content on video streaming platforms. In 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 464-469). IEEE. https://doi.org/10.1145/3341161.3342913

Thelwall, M., & Mas-Bleda, A. (2018). YouTube science channel video presenters and comments: Female friendly or vestiges of sexism?. Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-09-2017-0204

Vedula, N., Sun, W., Lee, H., Gupta, H., Ogihara, M., Johnson, J., ... & Parthasarathy, S. (2017, November). Multimodal content analysis for effective advertisements on Youtube. In 2017 IEEE International Conference on Data Mining (ICDM) (pp. 1123-1128). IEEE. https://doi.org/10.1109/ICDM.2017.149

Veletsianos, G., Kimmons, R., Larsen, R., Dousay, T. A., & Lowenthal, P. R. (2018). Public comment sentiment on educational videos: Understanding the effects of presenter gender, video format, threading and moderation on YouTube TED talk comments. PloS One, 13(6), e0197331. https://doi.org/10.1371/journal.pone.0197331