Personality Distinguish Based on Text Messages Using Machine Learning

Zainab Hama Namdar Suhad Malallah Kadham Khalil Ibrahim Ghathwan

University of Technology/ Computer Science Department

Baghdad – Iraq

E_mail: cs.19.73@grad.uotechnology.edu.iq

Abstract

Day after day, the number of users of text messages on social media is growing, knowing that, the impersonation on these sites is also growing, and often the identity of the sender is indistinguishable. In order to get rid of blackmail and threats and protect the users of these sites by recognizing text messages dialogues, this paper is therefore an attempt to identify the individual through personal message dialogues. The text selection method is used and its classification performance is verified using six machine learning methods (Random Forest, Decision Tree(J48), KNN, Logistic Regression, Naive Bayes, and SGD). The results showed that the Decision Tree and Random Forest outperformed other classification methods with a precision of about 99%.

Keywords: Machine Learning, Classification, TF-IDF and Text Message.

تمييز الشخصية بالاعتماد على الرسائل النصية باستخدام التعلم الآلي

زينب حمه نامدار سهاد مال الله خلف خليل ابراهيم غثوان

الجامعة التكنولوجية/قسم علوم الحاسوب

بغداد- العراق

الخلاصـــة

يومًا بعد يوم، يتزايد عدد مستخدمي الرسائل النصية على وسائل التواصل الاجتماعي، مع العلم أن انتحال الهوية على هذه المواقع يتزايد أيضًا، وغالبًا ما يتعذر تمييز هوية المرسل. من أجل التخلص من الابتزاز والتهديدات وحماية مستخدمي هذه المواقع من خلال التعرف على حوارات الرسائل النصية، فإن هذه الورقة هي محاولة للتعرف على الفرد من خلال حوارات الرسائل الشخصية. استخدمت طريقة اختيار النص والتحقق من أداء التصنيف باستخدام ستة طرق للتعلم الآلي هي الغابة العشوائية وشجرة القرار(J48) وKNN والانحدار اللوجستي وساذج بايز و(SGD). اظهرت النتائج طريقتي شجرة القرار والغابة العشوائية التي تم إجراؤهما تميزتا على طرق التصنيف الأخرى بدقة تصل إلى حوالي 99٪.

الكلمات المفتاحية: التعلم الآلي والتصنيف وTF-IDF والرسائل النصية

Introduction

Text mining can be defined as one of the significant tools used to obtain valuable information, including the classification and clustering of text by (Chen *et al*., 2020). Text classification might be utilized in a lot of applications in many areas, such as spam detection in opinion reviews, digital library systems, Email message classification, sentiment analysis assessment, film analysis reviews, text summary, marketing sentiment analysis, and mining of Arabic opinion. There are several researches in terms of text classification. With regard to natural languages, like English, Latin, Chinese and Turkish text by (Bahassine, *et al*., 2020). There are many techniques or algorithms that can be used to process data, but this study focuses on one which is known as TF-IDF, which is a numerical statistic showing the keywords' value for specific documents, or it can be inferred that it contains those keywords that can be used to identify or categorize specific documents as illustrated by (Bafna, *et al*., 2016). The machine learning algorithms are used in various areas such as classification problems, and regression problems as (Ghodke, *et al*., 2020) mentioned. The aim of this work is to provide protection for the user from the impersonation of his identity by others and the exploitation of his personal text messages. One of the areas that have been used in this study is discovering messages in the personal writing behavior through the daily use of text messages on social media, knowing that the discovery of the individual is through the behavior of his word from his personal textual dialogue. The text classification method was used for implementation purposes. While the model might be categorized into tree major steps:

- Preprocessing Step: Tokenization, Normalization, Removing Symbols, Stop Word, and Stemming.
- Features Extraction Step: The associated features were chosen from the original text in the presented step, also they have presented the text utilized in TF-IDF.
- machine learning: Several methods have been used to teach the model how to distinguish the Text Dialog.

Materials and Methods

Related Works

The following few works are related to this technique: have used SVM, KNN, and NB for sentiment analysis related to Arabic data-set of tweets as well as Facebook comments. In addition, they utilized TF-IDT for extracting features. The results indicated that the precision that has been achieved via the use of the NB is approximately 66.20, the precision that has been achieved via SVM is approximately 75% and the precision that has been achieved via the use of the KNN is approximately 70.97%. by (Duwairi, *et al*., 2014), SVM, NB, and Stochastic Gradient Descent (SGD) with TF-IDF for a classification system to categorize Bangla text document. The results of applying SGD was 93% with (Kabir, *et al*., 2015), have used KNN, NB, and SVM for recognizing the personality of a user with the use of the Big5 personality model from the tweets that have been posted in Indonesian and English languages. The results of applying KNN = 58% NB = 60% SVM = 59% by Pratama et al., (2015), have used NB, SVM, Logistic Regression and Decision Tree. For the application on SMS spam categorization the results of applying Naïve Bayes was 97% by (Arivoliet, *et al*., 2017). They utilized a couple of techniques for ML, and those were SVM and NB for Semantic Sentiment Analysis related to the text in

Arabic. Also, they utilized Arabic WordNet (AWN) and BOW concepts as an external knowledge for extracting the features, while the experimental results indicating that utilizing concept features enhances ATSA performance in comparison to basic BOW representation. The precision of naïve Bayes was about 85.99% by Sana (Alowaidi, *et al*., 2017), utilized SVM, NB and Neural Net, TF-IDF, LIWC, Emulex, and Concept Net for the personality prediction from online texts and the results showed that SVM with all the feature vectors reached the best accuracy overall MBTI dimensions are 88% by (Bharadwaj, *et al*., 2018), utilized SVM, NB and Logistic Regression with n-gram features weighted as well as TFIDF values for detecting offensive language and hate speeches on Twitter, the result of applying SGD with n-gram is 95.6% by (Gaydhani, *et al*., 2018), examined the individuals' public opinion in social media particularly Twitter and Facebook specified for having huge data that is difficult to analyze. By using k means, Naïve Bayes, and KNN algorithms, the accuracy was tested. NB's accuracy is between 80.526% and 82.500%, whereas the combination K-means and NB has accuracy between 80.323% and 81.523% by (Li, *et al*., 2018), have used KNN, random forest and Logistic Regression for BBC news text classification model such algorithms generated maximum result with TF-IDF is logistic regression with a result of 97% with (Shah, *et al*., 2020), utilized algorithms of decision tree such as Decision Stump, Hoeffding Tree, J48, REP Tree as well as Random Forest to recognized online scam or computer fraud. After comparing each algorithm's results, it is indicated that J48 reached the minimum error rate and maximum accuracy between other classifiers by (Palad, *et al*., 2020), utilized SVM, Gini, KNN, and Bagging and Boosting for developing an automatic model for the open-ended Physics questions utilizing the algorithms of text classification. Also, the results of using AdaBoost.M1 method had the best performance by (Çınar, *et al*., 2020), have used Random Forest, Naive Bayes, Lazy Random Forest to detect whether the comment, SMS or text message is SPAM or Normal message with two techniques called to hold out and K-fold Cross-validation. The proposed MLRF has the best performing capacity on text comment classification is 82.5% with (Ghodke, *et al*., 2020).

A New Proposed Model

In this model, we used a dataset (Daily Dialog) The proposed work depends on the high-quality multi-turn data collection, Regular dialogues. Using The language which is human-written and less distracting. Manually, they also mark the established dataset with communication by (Li, *et al*., 2017) The steps of the new model have been shown in Figure (1).

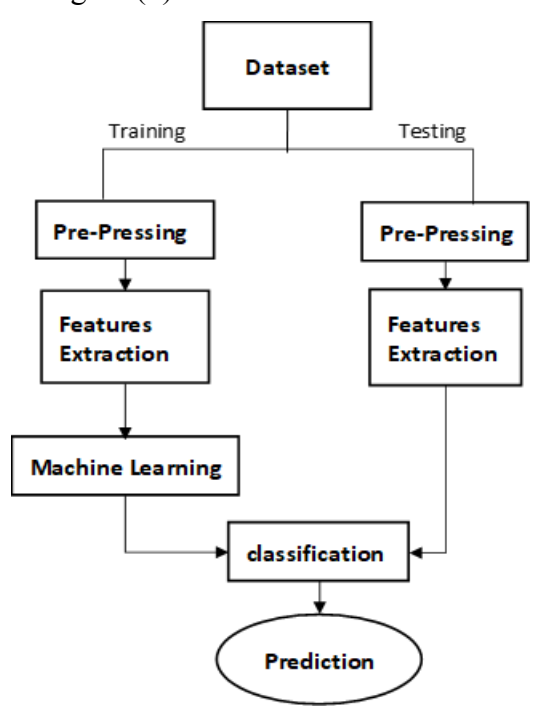

Figure (1) The Main Steps of the Proposed Model.

Step 1 Pre-Pressing

This step is important to prepare the dataset for Feature extraction. The Pre-pressing steps are shown in Figure (2), and it includes the following steps:

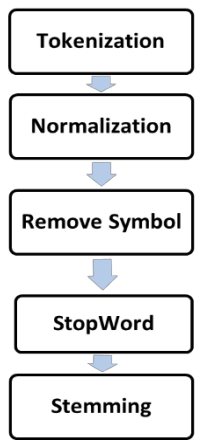

Figure (2) Pre-Pressing Steps

Tokenization: This is considered as the first step in NLP, in which the text messages are divided into tokens with the use of space to discrete each one of the words from the other, such tokens might be words, numbers, and symbols.

Normalization: This is a process of unification regarding many forms of the same letter in text messages through Converting from upper case to lower case.

Remove Symbol: In this step the remove all symbols is meaning remove (all non-letters) such as !, ?, @, *, #, $, %, &, ( ), [ ], { }, , =, !=, +, -, _, ,, ;, :, :-, ", ', \ , /, and numbers.

Stop Word: Text classification involves several terms which do not contain essential meaning to be used in classification algorithms, such as {"her", "hers", "will", "just", "don", "should", "now",. .}. Removing them from texts and records is the most common way of dealing with these terms remove.

Porter Stemmer: It is an operation that reduces the phrases to the frameworks of their uninflected basis. The stem is not close to the root sometimes. Anyway, it's helpful because while the stems are not in the real root, usually the related words map to one stem. Stemming is a significant stage in the process of text mining.

Step 2: Features Extraction (TF-IDF)

TF-IDF can be defined as one of the techniques majorly utilized for determining the word's significance in corpus text. There is a proportional increase in the word's meaning with the number of times that such words appeared in a conversation. On the other hand, the meaning will be decreased with the frequency in the corpus that such words appeared. TF (Term Frequency) shows how many times a given word occurs in the dialog. The TF is standardized in order to avoid TF prejudice against long dialogues, and the computation in the following way:

$$TF = \frac{W_i}{\sum_k w_i} \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

In which wi: representing the word frequency regarding the ith word, while k representing the total number of dialogues in the corpus. Assuming that D is the total number of dialogs in the corpus, while dw is the total number of dialogs that contain the term w, then IDF (inverse document frequency) can be determined in the following way.

$$IDF = \log \frac{D}{d_{W_i}+1} \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Then we Calculate the TF-IDF:

$$TF\text{-}IDF = \frac{W_i}{\sum_k w_i} \times \log \frac{D}{d_{W_i}+1} \ldots\ldots\ldots (3)$$

Step 3: Machine learning

using six machine learning classification methods to teach the model how to distinguish the text dialog.

Naive Bayesian: The Naive Bayesian classifier is a supervised learning algorithm Classification model of NB which is used for calculating the class's posterior probability. It applies Bayes Theorem for the purpose of predicting

the probability that a certain feature set is belonging to a specific label. For researches with over 6 authors: then, horizontally add the author names, move to 3rd row if required for over 8 authors.

$$P(\text{label}|\text{features}) = \frac{p(\text{label})*p(\text{features}\,|\text{label})}{P(\text{features})} \dots\dots\dots (4)$$

P (label) represents prior probability.
P (features| label) represents a prior probability in which a certain feature set was evaluated as a label. P (features) was the prior probability that a certain feature set occurred.

Logistic Regression (LR): LOR might be multinomial, ordinal, or binomial. The latter, sometimes referred to as binary LR, handling situations in which the results acquired for the dependent variable has just 2 likely forms, "1" and "0" (that might be representing "no" vs. "yes" or "false" vs. "true"). Logistic function was utilized for evaluating the relationship between the categorically dependent variable and at least one independent variables.
The equation for LR:

$$\log y = \frac{\exp(\alpha*x1+\beta*x2+y)}{1+\exp(\alpha*x1+\beta*x2+y)} \quad \text{.....(5)}$$

Where y is the bias, α and β is the weight, and x1, x2 is the features.

K-Nearest Neighbor (KNN): KNN can be defined as an approach to classify the data on the basis of the nearest training set in feature space. Also, it has been utilized for deducing the text's general prediction. Furthermore, the weighted sum in KNN classification is written in the following way:

$$\text{Score}(di, d) = \sum_{dj=knn(d)} \text{sum}\,(d, dj)\ \delta(dj, cj) \text{.....(6)}$$

In which KNN (d) is indicating the set of KNN of text d. In the case when dj belongs to ci, δ(dj,cj) equals 1, or otherwise 0. For test text d, it should belong to the class that has the highest resulting weighted sum. In order to compute the sum of (d,dj), Euclidean distance is used for representing the usual manner in which humans thinking of distance in the real- world.

$$D_{euclidean}(x,y) = \sqrt{\sum_{i=1}^{m}(xi - yi)^2} \dots (7)$$

Where KNN (d) defines the set of K nearest neighbors with text d. If dj belongs to ci, then δ(dj, cj) equals 1, or else 0. It should belong to the class which has the highest resulting weighted total for test text d. We use the Euclidean distance to measure the number (d, dj), which is the normal way humans think of distance in the real world.

Decision Tree(J48): An efficient way of producing is a traditional decision tree technique Models in a form of tree structure, the technique of Decision Tree breaks into smaller and smaller subsets, down a dataset into small sub-sets. On the other hand, an associated decision tree was formed gradually. The outcome of a tree with decision nodes as well as leaf nodes is part of such a method. Decision trees are able to treat both categorical and numerical data. J48 algorithm is using the entropy function to test the terms' classification from the test set.

$$\text{Entropy (Term)} = -\sum_{j=1}^{n} \frac{termj}{term} \log 2 \frac{termj}{term} \dots (8)$$

Random Forests: A large number of decision trees have been developed in this algorithm. As together they work. Decision trees act as pillars of That algorithm. The group of random forests is known as the group of Decision trees with nodes recognized at the pre-processing stage. The best nodes are identified at the pre-processing stage following multiple trees' construction. The function was chosen from a random subset of features. Another definition of creating a decision tree that it is built using the algorithm of the decision tree. Therefore, these trees are random forests

that were utilized for distinguishing new objects from the input vector. With regard to, each one of the constructed decision trees that will be utilized. Assuming that class the tree votes, then the random forest is selecting the classification with the majority of votes regarding all trees in classification. Based on 2 factors, there were many error probabilities in the random forest:

- Chances are that 2 trees in the forest might have an association with each other, contribute to increasing the error rate.
- Each one of the trees has its own power. Thus, a good classifier is a tree that has a low error rate, and vice versa. The mathematical formula related to random forest classifier is as follows:

n_(Ij)= w_l C_j–w_left(j) C_left(j) – w_right(j) C_right(j) ..... (9)

n_(i˙)sub(j) = the importance of node j
w sub(j) = weighted number of samples reaching node j
C sub(j) = the impurity value of node j
left(j) = child node from left split on node j
right(j)= child node from the right split on node j

Stochastic Gradient Descent (SGD): It is a well-known approach used to solve problems with ML. It is a technique of optimization where a sample is randomly chosen instead of whole data samples in each iteration. The weight update method for gradient descent and stochastic gradient descent at jth iteration is represented by equations (10) and (11):

$$w_j := w_j — \alpha\, \partial J\, \partial w_j \ldots\ldots\ldots (10)$$

$$w_j := w_j — \alpha\, \partial Ji\, \partial w_j \ldots\ldots\ldots (11)$$

Here, α indicates the learning rate, J represents the cost of all examples of training and Ji is the cost of the example of ith training. To measure the sum of the gradient of the cost function of all the samples is computationally expensive; therefore, it takes a lot of time to complete each iteration. In each iteration, SGD takes one sample randomly to solve this problem and calculates the gradient. While more iteration is required to converge, with shorter training time, it can meet the global minima.

Performance Measures

A lot of evaluation metrics were used for comparing the performances related to supervised classification ML approaches. Frequently, the evaluation metrics were utilized in supervised ML and allow testing the algorithm's efficiency. Also, a confusion matrix was utilized for comparing the message detection performance as seen below.

Ttable (1) A confusion Matrix was utilized for evaluating the Message Detection Performance.

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

TP, FP, TN, and FN are concepts indicated in the following way:
- True positive (TP): represents the positive instances classified correctly.
- False Negative (FN): represents the positive instances classified incorrectly.
- False Positive (FP): represents the negative instances classified incorrectly.
- True negative (TN): represents the negative instances classified correctly.

Precision Measure

Precision Measure is the proportion which is related to correctly classified expected text for a certain class.

Precision =(|TP|)/(|FP|+ |TP|)…….(12)

Recall Measure

Recall Measure is the proportion related to all text for a given class which are correctly classified

Recall =(|TP|)/(|FN|+ |TP|)……(13)

F1-measure

F1-measure might be utilized for estimating document results. The precision and recall were blended by classifiers.

F1-measure = 2 × ((Recall×Precision|)/(Recall+Precision)). ……..(14)

Results and Discussion

A lot of experiments were carried out for finding the final predictions. Table (2) is showing the results regarding six algorithms tested for classification.

Table (2) The Decision Tree and Random Forest Results with Precision, Recall, and F1-Score.

Algorithms	Precision	Recall	F1-Score
Decision Tree	99%	99%	99%
Random Forest	99%	99%	99%
SGD	87%	87%	87%
Logistic Regression	83%	83%	83%
Naive Bayes	81%	81%	81%
KNN	80%	80%	80%

Conclusion

In this paper, six machine learning methods (random forest, logistic regression, KNN, decision tree, naive Bayes, SGD) and the TF-IDF method were used to extract features based on the Regular Dialog dataset and results obtained from six machine learning methods were compared. The best thing about the system that it is a decision tree classifier and random forest to identify the entity by means of personal messages is best done than other classifiers. KNN methods gave the worst results because they do not take all potential possibilities.

Future Work

For the future, these six machine learning algorithms will be added to the unsupervised process. Also, the use of a deep learning system Instead of a machine learning method may be a good suggestion used for better precision. At the same side, building an application that identifies the sender of text messages Operates on all social media platforms.

References

Alowaidi, S; Saleh, M; and Abulnaja, O. (2017). Semantic Sentiment Analysis of Arabic Texts. International Journal of Advanced Computer Science and Applications, 8(2), 256-262.

Arivoli, P. V; Chakravarthy, T; and Kumaravelan, G. (2017). Empirical Evaluation of Machine Learning Algorithms for Automatic Document Classification. International Journal of Advanced Research in Computer Science, 8(8).

Bafna, P.; Pramod, D.; and Vaidya, A. (2016). Document Clustering: TF-IDF Approach. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 61-66). IEEE.

Bahassine, S.; Madani, A.; Al-Sarem, M.; and Kissi, M. (2020). Feature Selection Using an Improved Chi-square for Arabic Text Classification. Journal of King Saud University-computer and Information Sciences, 32(2), 225-231.

Bharadwaj, S.; Sridhar, S.; Choudhary, R.; and Srinath, R. (2018, September). Persona Traits Identification Based on Myers-briggs Type Indicator (MBTI)-A Text Classification Approach. in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1076-1082). IEEE.

Chen, X.; Qi, J.; Zhu, X.; Wang, X.; and Wang, Z. (2020). Unlabelled Text

Mining Methods Based on Two Extension Models of Concept Lattices. International Journal of Machine Learning and Cybernetics, 11(2), 475-490.

Çınar, A.; Ince, E.; Gezer, M.; and Yılmaz, Ö. (2020). Machine Learning Algorithm for Grading Open-ended Physics Questions in Turkish. Education and Information Technologies, 1-24.

Duwairi, R. M.; and Qarqaz, I. (2014, August). Arabic Sentiment Analysis Using Supervised Classification. in 2014 International Conference on Future Internet of Things and Cloud (pp. 579-583). IEEE.

Gaydhani, A.; Doma, V.; Kendre, S.; and Bhagwat, L. (2018). Detecting Hate Speech and Offensive Language on Twitter Using Machine Learning: An N-gram and TF-IDF Based Approach. arXiv preprint arXiv:1809.08651.

Ghodke, T.; and Khadse, V. (2020). Effective Text Comment Classification Using Novel ML Algorithm-modified Lazy Random Forest (No. 3987). EasyChair.

Kabir, F.; Siddique, S; Kotwal, M. R. A; and Huda, M. N. (2015, March). Bangla Text Document Categorization Using Stochastic Gradient Descent (SGD) Classifier. in 2015 International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 1-4). IEEE.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. (2017). Dailydialog: A Manually Labelled Multi-turn Dialogue Dataset. ArXiv Preprint arXiv:1710.03957.

Li, Y.; Wang, X.; and Xu, P. (2018). Chinese Text Classification Model Based on Deep Learning. Future Internet, 10(11), 113.

Palad, E. B. B; Burden, M. J. F; Torre, C. R. D.; and Uy, R. B. C. (2020). Performance Evaluation of Decision Tree Classification Algorithms Using Fraud Datasets. Bulletin of Electrical Engineering and Informatics, 9(6), 2518-2525.

Pratama, B. Y.; and Sarno, R. (2015, November). Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM. In 2015 International Conference on Data and Software Engineering (ICoDSE) (pp. 170-174). IEEE.

Shah, K.; Patel, H.; Sanghvi, D.; and Shah., M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for The Text Classification. Augmented Human Research, 5(1), 1-16.