Asian Journal of Information Technology 18 (1): 20-27, 2019
ISSN: 1682-3915
© Medwell Journals, 2019

Arabic Semantic Classifier of Arabic Social Media "Twitter" Users

[1]Jamal Alnaeb, [2]Issam Salman and [1]Mohamad Bassam Kurdy
[1]Department of Web Sciences, Syrian Virtual University, Damascus, Syria
[2]Czech Technical University Prague, Parague, Czech Republic

Abstract: Text classification from text has gained a lot of interest in the last years, yet, some languages such as Arabic (with its different spoken dialects) has not been given such attention. In this study, we present our work in the text classification of Arabic texts with a focus on spoken Arabic dialects on Twitter messages. Therefore, we have constructed a corpus of Arabic spoken Tweets and implemented two approaches to automatically classify Tweet messages. One of the approaches uses different machine learning while the other approach uses semantic web to do the classification. Further, a comparison has been made between the two methods to determine the best configuration of the Tweets recognition system.

Key words: Text analysis, data mining, Tweet classification for Arabic text, Twitter, spoken arabic dialects, comparison

INTRODUCTION

The field of "Text analysis" and "Emotion analysis" from text is one of the promising fields in "Text mining" for its services and great benefits to decision makers in political, social, financial and marketing domains. This field has doubled its importance with the growing of social networks and the feedbacks, since, the social networks users currently have exceeded 2 billion. Moreover, the typical users spend about 4-6 h a day browsing these networks, sharing their information and express their opinions, evaluating products they buy, the books they read and appraising the services they receive not to mention sharing ideas, needs and requirements.

All mentioned above have become the data highly valuable in terms of their (importance and value of information as well as their novelty and realism as they result from the interaction of the users themselves). These data vary according to, the topics presented (medical, cultural, tourism, artistic, policy, commercial and etc.).

We can say that the social media platforms are now considered as the top source of data contributed by users from around the world. The valuable data and its importance urge the majority of researchers concentrate on the "Text analysis" and "Emotion analysis" to initiate tools and techniques help business organizations to build their reputation based on a deep understanding of their customer's directions and tendencies.

The lack of appropriate tools for word-processing that support Arabic language, in particular, the spoken dialects is due to the difficulty of Arabic grammar and the unfamiliarity/ignorance of the large number of researchers with such a grammar. These reasons have led to rare and limited studies on specific subjects which is often specific scope between the comparisons and the suggestions without showing a real impact or results for building real and effective applications in the processing of Arabic language.

On the another hand, word processing tools that support English language are various and making a lot of job (determine the beginning and the end of each sentence in the text determine the verb, subject and object in the sentence). In addition to these tools, there are dictionaries as applications which are often available in English. This is very helpful for research activity in this field.

In this study, two approaches (Machine learning algorithm semantic web) to classify the Tweet message written in Arabic (Arabic spoken dialect) into seven different categories (Politics, economy, health, sport, culture and arts, science and technology, woman) will be presented.

With a focus on studying user's Tweets on Twitter where we used more than 3475 Tweets in Arabic language-various Arabic spoken dialect. These Tweets collected and classified manually in this study in order to develop a semantic web-based classification model which achieve 81.4% accuracy.

Challenges and importance of research: The main challenge in text analysis from Arabic text is when the researchers have considered that the research studied for the processing of the Arabic language is rare and limited because the difficulty of processing it and its

Corresponding Author: Jamal Alnaeb, Department of Web Sciences, Syrian Virtual University, Damascus, Syria

ramifications compared to the other languages. Besides, there are other challenges related to processing Arabic language, the Arabic language splits into two parts: the standard language that is used in books, newspapers and official meetings, the second part is that used among people inside the social media which vary according to, the region in addition to the difficulty of Arabic grammar in terms of morphology and the lack of familiarity of most the researchers with these rules.

Literature review: A lot of researches had been accomplished in the field of studying the analysis and classification of, Tweets within social media sites "Twitter". Herein some of these researches.

Alabdullatif *et al*. (2016) which aims to classify Arab users on Twitter based on their behavior and interests into five categories. In addition it displays the user's specific advertisements according to, his/her interest. Basit had applied the Naive Bayes NB algorithm which achieved a high accuracy of 90.32%.

Alabbas *et al*. (2017) was about classifying Twitter Tweets written in Arabic (informal colloquial) into two categories: positive and negative. They used many techniques for processing words and each word represented using the TF_IDF mathematical equation. Further, they trained the model by used several machine-learning algorithms. The accuracy achieved 0.933%.

Daood *et al*. (2017) their study was designed to classify the Tweets written in Arabic as the "Syrian dialect" on Twitter into six emotional categories. It has applied three classifiers: Naive Bayes NB-CRF ("Conditional Random Fields") sequential minimal optimization. Moreover, the study has conducted a series of treatments for the text strings which included (removing the formations, removing the duplicate characters, removing the links and references, standardization some characters) in addition to removing stop words and then using several dictionaries and has been used as an "ISRI" parser to return the word to its root. The two researchers have presented the data radically and has used three weighting equations: (TF_IDF_Weight_TWF-Modified TF). the achieved accuracy was 66.9% F-measure.

By Barbosa and Feng (2010) the study goal was to classify English language Tweets on Twitter into three emotional categories (positive, negative and neutral). The researchers here applied Supervised Machines (SMO) to train three training blogs the result was achieved by sequence where the first one was 77%, the second was 82% and the third was 89%.

Dataset: One of the most important challenges in this research is the lack of suitable Arabic datasets and lexicons which serve the subject of the research.

We have collected manually more than 3.475 Tweets from the popular Twitter social media pages such as Al Jazeera, Sky News and Shaba News. The collected Tweets were categorized manually into seven groups (Politics, economy, health, sport, culture, arts, science, technology and woman) to help us building the training data that train/reinforce the model in addition, we have collected 285 Tweets to test the model. Therefore, we have a dataset of 3760 Tweets, 3475 Tweets for training and 285 Tweets for testing.

MATERIALS AND METHODS

In this research, we classified the Tweet messages into seven categories as mentioned earlier (Politics, economy, health, sport, culture and arts, science and technology, woman) from Tweets written in Arabic dialect. The syntactical structure of a word will be considered by using N grams (on the full-form words). After collecting the Tweets for each category as dataset, we processed them by:

- Eliminating @Users, #hashtags and URLs
- Normalization (such as Hamza "أ" standardization)

Then we refined the dataset by filters which exclude the punctuation marks and stop words. All aimed at decreasing the dimension length of the word in the dataset.

The Tweets were represented as feature vector taking the Tweet's words as features in this vector. Also, we improved the feature vector to be represented by the seven categories to reduce the vector size. Before all of that it was so important to collect several dictionaries. These dictionaries had been collected during our work. The words of these dictionaries had been taken from textual training data in addition to common Arabic words that social media users use.

Dictionaries

Dictionary for repeated words: This dictionary includes all the words of the training data as well as the repetition of each word within the training data.

Dictionary weights words: We need to take into consideration each word weight. Therefore, for each word in the training data "Tweets", we calculated the weight of that word within each category. Then, we built a set of dictionaries that indicate the weight of the words within the category. In this research, we compared between two methods to measure this weight.

TF_IDF Term Frequency Inverse Document Frequency (Mohsen *et al*., 2016): “Q_1” number of the term [x] in Tweets which refer to emotion [e]. “Q_2” number of the terms within Tweets which refer to emotion [e]:

$$f_{ij} = Q_1 / Q_2$$

“N” number of all Tweets with in training data.“dfi” number of all Tweets which contain term [x] with in training data:

$$\text{Freq. of term}_x \text{ in emotion}_e = \left(\frac{f_{ij}}{\max_{fij}}\right) \times \frac{\log(\frac{N}{df_i})}{\log(2)}$$

Weighed-TWF (Do and Choi, 2015): “D” number of all Tweets with in training data. “ne” number of the Tweets which refer to emotion [e] and contain term [x]. “Q” Total number of terms in the d:

$$\text{Normalized}_{\text{Tweet frequency}} = ne/Q$$

$$\{(ne < D) \text{ then weight } = 1/ne \text{ else weight } = 0$$

$$\text{Freq of term}_x \text{ in emotion}_e = NTF{*}weight$$

Dictionary of synonyms N-gram: This dictionary contains a collection of words and it is build to come together to indicate one meaning such as (“صباح ا لخير” “عليكم السلام”) and this dictionary was built on (1 and 2 gram), for example, the word “السلام” alone and the word “عليكم” alone in 1 gram. In 2-gram is taken one word “عليكم السلام”.

Dictionary of quality words gain-ratio: By gain-ratio we mean the most qualitative words in other words, the set of words inside the Tweet which has a strong signal to all Tweets. This dictionary serves to identify the most influential words on the entire blog where a dictionary was built containing a set of qualitative words with each weight.

Classification: We used two approaches to classify the Tweet. One of which is: we trained many types of classifiers:

- SVM (SMO) “Support Vector Machine”(Platt, 2000)
- NB “Naive Bayes” (Mozina *et al*., 2004)
- Bayes Net (Allen and Darwiche, 2013)
- CRF “Conditional Random Fields” (Toepfer *et al*., 2010)
- Random tree (Mishra and Ratha, 2016)
- J48 (Moskovitch *et al*., 2007), the result of the classification algorithm models was evaluated according to, the algorithm 1

Algorithm 1; Classification algorithms machine learning:

“Precision, recall and F-measure”
Precision = #correct guesses/#total
guesses
Recall = #correct guesses/#total
F-measure = 2PR/(P+R)
where #correct guesses is the number of
statements marked correctly as expressing
an emotion X by the classifier, #total
guesses is the total number of statements
that are marked by the classifier as
expressing the emotion X (including
correct and wrong guesses) and #total is
the number of
Statements expressing the emotion X in
the dataset

Classification using “semantic web”: The semantic web is defined as an extension of the current structure of the web. The resources are defined using a descriptive language which allows the semantic content of the resource to be expressed in such a way that the machine can interpreted. A number of languages and mechanisms have been proposed to express the resource’s semantic content. Each of these mechanisms is concerned with a number of practical aspects and omits others. All these languages and mechanisms have been consolidated into a more comprehensive, dynamic and holistic structure allowing the expression of the semantic content of resources to be an effective expression understood by man both natural and the machine as well as the broad potential offered by this structure to query resources. The general structure of any application in the semantic web consists of three basic components the ontology, the group of object and the software class. We will discuss these components in some detail as in Fig. 1.

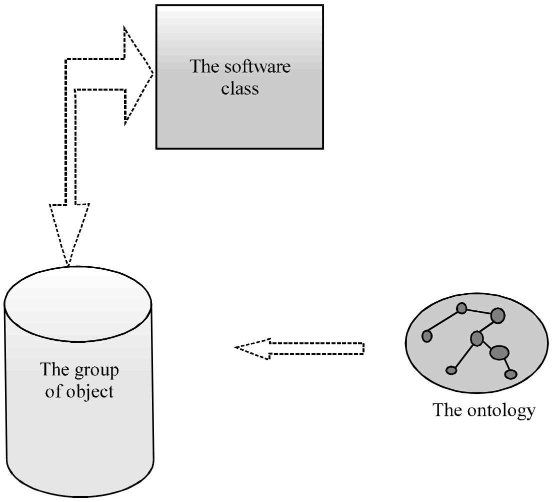

Fig. 1: The general structure of application in the semantic web

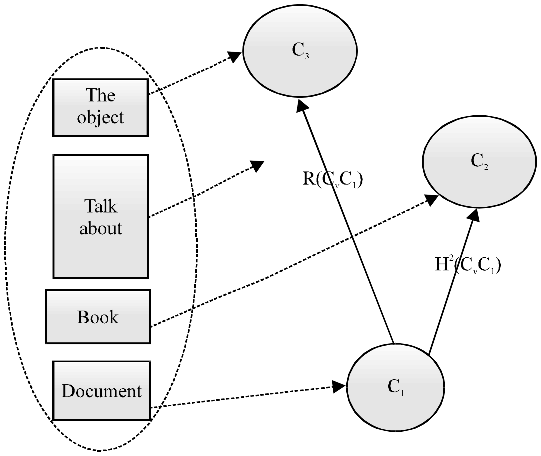

Fig. 2: An example of using ontology

The ontology: Defining as a description of concepts, purposes and entities in a specific area as well as relations between them. In other words, the ontology is a model of the common relations and objects that are generally applicable across a wide range of domain ontologies. It usually employs a core glossary that contains the terms and associated object descriptions as they are used in various relevant domain ontologies. Figure 2 represents sample ontology. The using of ontologies in the field of knowledge representation and management as well as word processing applications to information retrieval systems which have allowed the representation of the concepts of the field studied as well as the relationships between them and specialized editors do it through a set of standard languages. OWL (Ontology Web Language) is one of the most popular and is a standard W3C-certified language based on XML and URI.

The retrieval of the knowledge of the ontology requires a query language as is the case when retrieving data stored in databases. SPARQL (Simple Protocol and RDF Query Language) is a standard language for query within the field.

The group of object: Is defined as a database that expresses the set of purposes created from ontology.

The software class: Is a set of dependencies and algorithms that use the previous set of purposes to provide a number of functions for which the application was built such as searching, comparing, translating documents and etc. The software class can be an inference engine that depends on the set of objects as facts and the set of rules that are defined in the ontology as rules of inference. In our research, we represent our training data in ontology file and the extension is RDF file as Fig. 3. Figure 4 shows the procedure of using "semantic web classifiers".

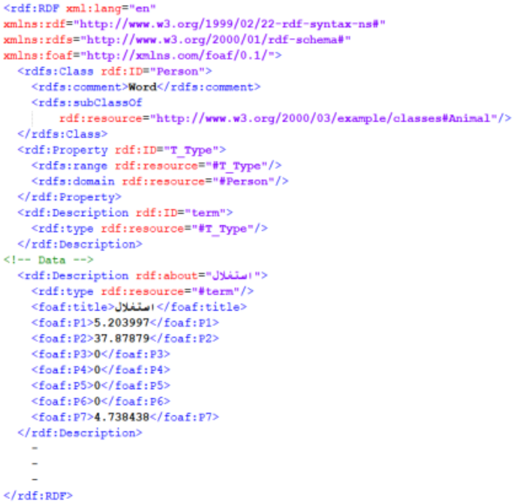
```
<rdf:RDF xml:lang="en"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <rdfs:Class rdf:ID="Person">
    <rdfs:comment>Word</rdfs:comment>
    <rdfs:subClassOf
        rdf:resource="http://www.w3.org/2000/03/example/classes#Animal"/>
  </rdfs:Class>
  <rdf:Property rdf:ID="T_Type">
    <rdfs:range rdf:resource="#T_Type"/>
    <rdfs:domain rdf:resource="#Person"/>
  </rdf:Property>
  <rdf:Description rdf:ID="term">
    <rdf:type rdf:resource="#T_Type"/>
  </rdf:Description>
<!-- Data -->
  <rdf:Description rdf:about="استقلال">
    <rdf:type rdf:resource="#term"/>
    <foaf:title>استقلال</foaf:title>
    <foaf:P1>5.203997</foaf:P1>
    <foaf:P2>37.87879</foaf:P2>
    <foaf:P3>0</foaf:P3>
    <foaf:P4>0</foaf:P4>
    <foaf:P5>0</foaf:P5>
    <foaf:P6>0</foaf:P6>
    <foaf:P7>4.738438</foaf:P7>
  </rdf:Description>
    -
    -
    -
</rdf:RDF>
```

Fig. 3: Descriptive our training data in ontology

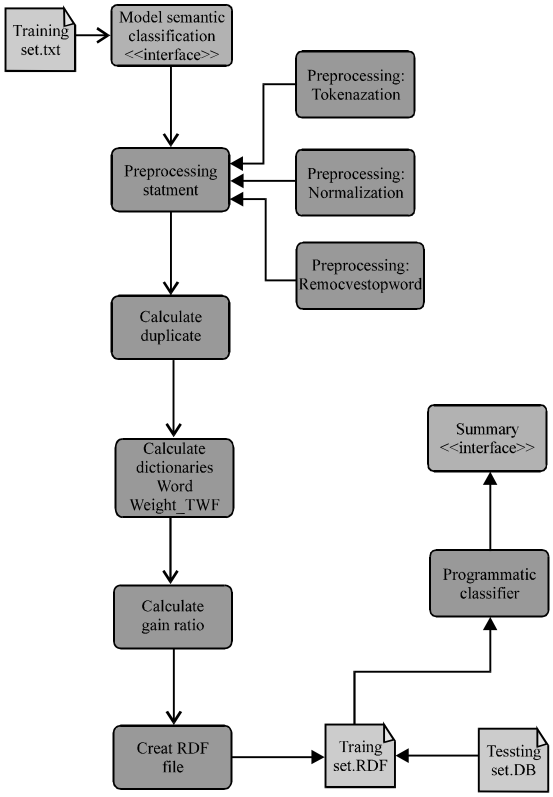
Training set.txt
Model semantic classification <<interface>>
Preprocessing: Tokenazation
Preprocessing statment
Preprocessing: Normalization
Preprocessing: Remocvestopword
Calculate duplicate
Calculate dictionaries Word Weight_TWF
Summary <<interface>>
Programmatic classifier
Calculate gain ratio
Creat RDF file
Traing set.RDF
Tessting set.DB

Fig. 4: The procedure of using "semantic web classifiers"

Vector space modelling: The Tweets are represented as a feature vector by considering each word in the Tweets

Table 1: An example of feature vector

Variables	Policy	Economy	Health	Art and culture	Science and technology	Woman	Sport
آبطال	0.0	0.0	4.210477	0.0	0.0	0	10.78592
العالم	0.0	4.210477	4.210477	0.0	0.0	0	10.78592
بالشطرنج	0.0	0.0	0.0	0.0	0.0	0	10.78592
يتواجون	0.0	0.0	0.0	0.0	0.0	0	0.0
كآس	0.0	0.0	0.0	0.0	0.0	0	10.78592
الملك	0.0	4.210477	0.0	2.10477	0.0	0	0.0
سلمان	4.210477	0.0	0.0	0.0	6.771168	0	0.0
Vector	4.210477	10.78592	10.39695	2.10477	6.771168	0	49.95544

Table 2: The finale vector before using the effect of gain ratio

Variables	Policy	Economy	Health	Sport	Culture and art	Science and technology	Woman
Vector	25.70255	16.22305	4.360773	233.0564	13.23304	11.01545	13.61452

Table 3: The final vector after using the effect of gain ratio

Variables	Policy	Economy	Health	Sport	Culture and art	Science and technology	Woman
Vector	395.2652	250.2708	67.03432	3580.052	203.4919	169.4183	209.284

as an attribute in the feature vector. Its value may take multiple forms: the simplest form represents each word by a number that reveals the term frequency in the source text "Tweet". The process starts by taking all the unigrams and bigrams from Tweets then keeping those which exceed a certain frequency to be represented as features in the final vector.

The second form is more advanced as when assigning weights to the words within the training data, we take into consideration both of syntactic features (N-grams frequency, total number of the words and characters) and stylistic features (punctuation marks).

Actually, the problem with all of these methods is the high dimensionality of the feature vector. The solution was to reduce the feature vector containing all training data words to a feature vector that consists of seven features only "a feature for each category". The value of each feature will be an accumulative number that stands for the total weighted sum of sentence's words according to, the category that this value refers to. For example, the following feature vector extremely indicates "Sport". "آبطال العالم بالشطرنج يتواجدون في كأس الملك سلمان" which mean world chess champions are in the King Salman Cup (Table 1-3).

Factors affecting our feature vector

The effect of the most influential words "gain ratio" on the final vector: After extraction and analyzing the final attributes vector of the Tweet, the analytical algorithm rechecks the Tweet to insure the existence of any influential words within. If effective word [S] appears, the algorithm will increment the Value [V] of the cumulative Counter [C] of each category [I] by the result of the multiplication of [V] value of this counter with the impact strength of this word [SM] on this category [I].

RESULTS AND DISCUSSION

Experiments: We have conducted multiple tests in order to compare between a numbers of Arabic Tweets classifying methods.

"Tweet words" as feature vector: Initially, we represented the dataset as feature vector taking the Tweet's words as attributes of these vectors and we applied the classifiers (SVM "SMO", Naive Bayes, conditional random fields) using this feature vector. Figure 5 presents the results. In order to explore the different feature vector implementation, we have implemented different models. Using S1-S4 sets. And we improved the feature vector to be represented by the seven categories.

- Weighted words model (S1): using TF_IDF
- Weighted words model (S2): using Weight-TWF
- Model (S3): using gain ratio and TWF
- Model (S4): using 1 and 2 gram and ratio and TWF

S1 Model: In this experiment, we study the effect of the weighting models used TF_IDF (Fig. 6). We can notice when representing the dataset using words as feature vector attributes the classifier results were less by 60-70% than when we improved the feature vector to be represented by the seven categories.

S2 Model: In this experiment, we study the effect of the weighting models used Weight_TWF (Fig. 7).

Comparison between S1 and S2 Model: In this experiment, we comparison between S1 based on TF_IDF and S2 based on Weight_TWF . We notice that using S2 "Weight_TWF" enhanced the results by 2% more than when using "TF_IDF (Table 4 and 5).

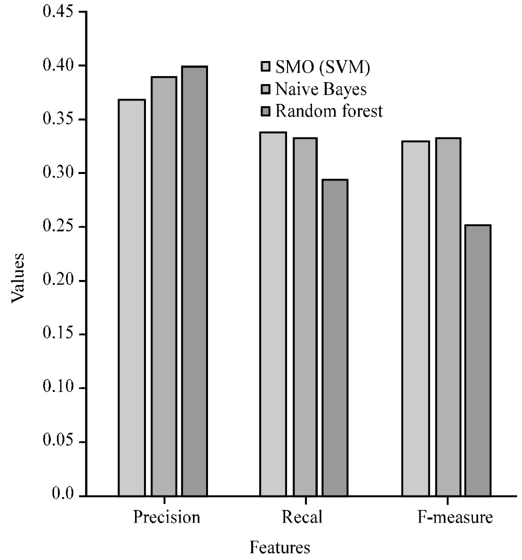

Fig. 5: Comparison between different classifiers using Tweet word as feature vector

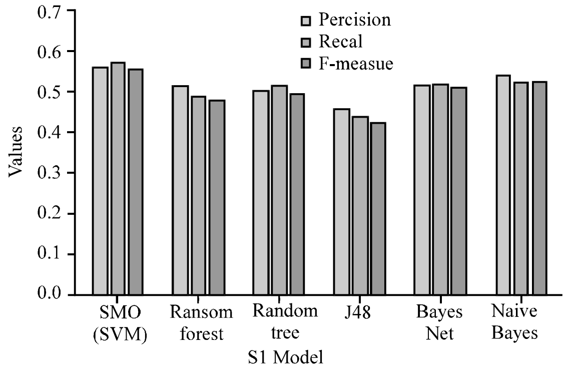

Fig. 6: The comparison between multi-classifiers based on TF_IDF

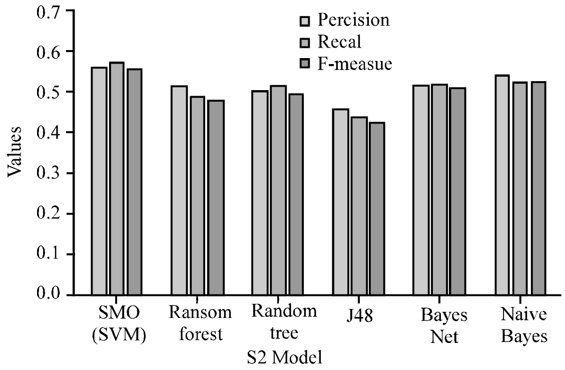

Fig. 7: The comparsion between multi-classifiers on Weight_TWF

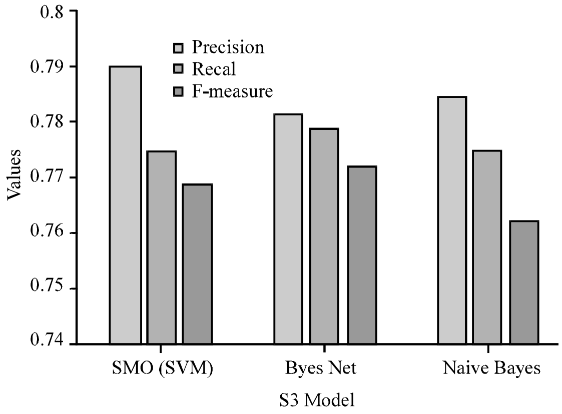

Fig. 8: The result of using multi classifiers based on Weight_TWF and gain ratio

Table 4: The comparative between classifiers using Weight_TWF and using TF_IOF

Model	Algorithm	Precision	Recall	F-measure
S1	SMO (SVM)	0.566	0.575	0.563
S1	Random forest	0.544	0.519	0.511
S1	Random tree	0.443	0.463	0.445
S1	J48	0.457	0.446	0.449
S1	Bayes Net	0.456	0.457	0.491
S1	Naive Bayes	0.528	0.551	0.544
	Average	0.524	0.502	0.491
S2	SMO (SVM)	0.568	0.575	0.563
S2	Random forest	0.519	0.495	0.483
S2	Random tree	0.51	0.519	0.497
S2	J48	0.463	0.442	0.432
S2	Bayes Net	0.518	0.523	0.514
S2	Naive Bayes	0.547	0.53	0.527
	Average	0.520	0.514	0.502

Table 5: The comparative between classifiers using TF_IDF and Weight_TWF and gain ratio

Models	Algorithm	Precision	Recall	F-measure
S1	SMO (SVM)	0.566	0.575	0.563
S1	Bayes Net	0.446	0.456	0.491
S1	Naive Bayes	0.528	0.551	0.544
	Average	0.513	0.527	0.532
S2	SMO (SVM)	0.568	0.575	0.563
S2	Bayes Net	0.518	0.523	0.514
S2	Naive Bayes	0.547	0.530	0.527
	Average	0.533	0.527	0.512
S3	SMO (SVM)	0.790	0.775	0.769
S3	Bayes Net	0.782	0.779	0.772
S3	Naive Bayes	0.785	0.775	0.762
	Average	0.785	0.776	0.767

S3 Model "gain ratio" vs. "Weight_TWF": In this experiment, we studied the effect of the weighting models used Weight_TWF because weighting models gave better accuracy than weighting models "TF_IDF" and we applied algorithm gain ratio, using the three Classifiers (NB, Naive Network, SVM (SMO) (Fig. 8).

Table 6: The comparative between classifiers using TF_IDF and Weight_TWF and Gain ratio and N gram

Models	Algorithm	Precision	Recall	F-measure
S1	SMO(SVM)	0.566	0.575	0.563
S1	Bayes Net	0.446	0.456	0.491
S1	Naive Bayes	0.528	0.551	0.544
	Average	0.513	0.527	0.532
S2	SMO(SVM)	0.568	0.575	0.563
S2	Bayes Net	0.518	0.523	0.514
S2	Naive Bayes	0.547	0.530	0.527
	Average	0.533	0.527	0.512
S3	SMO(SVM)	0.790	0.775	0.769
S3	Bayes Net	0.782	0.779	0.772
S3	Naive Bayes	0.785	0.775	0.762
	Average	0.785	0.776	0.767
S4	SMO(SVM)	0.784	0.768	0.761
S4	Bayes Net	0.80	0.775	0.767
S4	Naive Bayes	0.780	0.768	0.761
	Average	0.788	0.770	0.761

Table 7: The result of using semantic web

		Category name						
Variables	Test set	Policy	Economy	Health	Sport	Culture and art	Science and technolog	Woemen
Total instance	285	39	41	41	41	41	41	41
Cwrecay classified instances	232	36	23	31	36	33	31	36
Incorrectly classified instances	53	3	18	4	5	8	10	5
Correctly classified instances	81.4%	92.3%	56.10%	90.24%	87.80%	80.48%	75.60%	91.1%
Incorrectly clessified Instances	18.6%	7.7%	43.90%	9.76%	12.20%	19.52	24.40%	8.9%

S3 Comparison between S1 and S2 Model: The using of "Gain Ratio" enhanced the results of classifiers by 25% more than when using only "Weight_TWF".

S4 Model "Weight TWF" vs. "Gain Ratio" vs. "N gram": In this experiment, we study the effect of the weighting models used Weight_TWF and we applied algorithm gain ratio and principle of N gram based on the three classifiers (NB, Naive Network, SVM (SMO) (Fig. 9).

S4 Comparison between S1-S3 Model: We concluded that the principle of using "n gram" did not make any effect on the result (Table 6).

The classifiers using semantic web: We conclude from the above that the most accurate results were when using the TWF Model, taking into account the gain ratio value. Based on this summary, we tried to experiment with a classification methodology that differs from the previous methodology we built an ontology by describing all the words of the training data along with weights of each word in each category. We then applied SPARQL queries within the. Net environment using the dotNetRDF library to identify the news categories of the test data. The accuracy of the identification was more accurate at 3% than at best when we relied on a methodology to classify and use classification algorithms (Table 7).

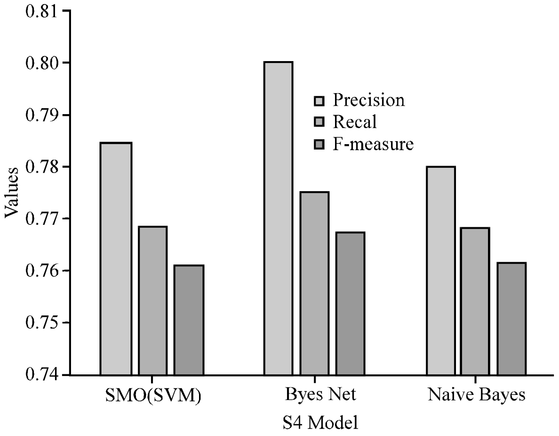

Fig. 9: The result of using multi classifiers based on Weight_TWF and gain ratio and N-gram

CONCLUSION

In this research, we proposed two different approaches models to recognize the basic seven categories-policy, economy, health, sport, culture and art, science and technology, woman of Arabic Tweets (Arabic dialectal Tweets). In order to test our models, we have built a balanced dataset of Arabic dialectal Tweets and manually annotated it. We compared the results of several machine-learning algorithms such as SVM, Naive Bayes, CRF, Naive Net, J48. We also compared the results of our proposed models.

We have adopted the three equations to test the accuracy (Precision-recall-F-measure) and have reached the best possible when using the SMO algorithm. We have achieved (Precision: 78.4% -Recall 76.8% F-measure: 76.1%) and we have achieved accuracy 81.4% when the semantic web concept was applied.

REFERENCES

Alabbas, W., H.M. Al-Khateeb, A. Mansour, G. Epiphaniou and I. Frommholz, 2017. Classification of colloquial Arabic Tweets in real-time to detect high-risk floods. Proceedings of the 2017 International Conference on Social Media, Wearable And Web Analytics (Social Media), June 9-20, 2017, IEEE, London, UK., ISBN:978-1-5090-5058-1, pp: 1-8.

Alabdullatif, A., B. Shahzad and E. Alwagait, 2016. Classification of arabic Twitter users: A study based on user behaviour and interests. Mob. Inf. Syst., 2016: 1-11.

Allen, D. and A. Darwiche, 2013. Optimal Time-Space Trade off in Probabilistic Inference. In: Advances in Bayesian Networks, Gamez, J.A., S. Moral and A. Salmeron (Eds.). Springer, New York, USA., ISBN:9783540398790, pp: 39-55.

Barbosa, L. and J. Feng, 2010. Robust sentiment detection on Twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, August 2010, Stroudsburg, PA., pp: 36-44.

Daood, A., I. Salman and N. Ghneim, 2017. Comparison study of automatic classifiers performance in emotion recognition of Arabic social media users. J. Theoret. Applied Inform. Technol., 95: 5173-5183.

Do, H.J. and H.J. Choi, 2015. Korean twitter emotion classification using automatically built emotion lexicons and fine-grained features. Proceedings of the 29th International Pacific Asia Conference on Language, Information and Computation: Posters (PACLIC 29), October 30-November 1, 2015, Deparment of Computer Science and Engineering Shanghai Jiao Tong University, Shanghai, China, pp: 142-150.

Mishra, A.K. and B.K. Ratha, 2016. Study of random tree and random forest data mining algorithms for microarray data analysis. Intl. J. Adv. Electr. Comput. Eng., 3: 1-7.

Mohsen, A.M., H.A. Hassan and A.M. Idrees, 2016. Documents emotions classification model based on TF_IDF weighting measure. Intl. Scholarly Sci. Res. Innovation, 10: 252-258.

Moskovitch, R., N. Nissim, D. Stopel, C. Feher and R. Englert *et al.*, 2007. Improving the detection of unknown computer worms activity using active learning. Proceedings of the Annual International Conference on Artificial Intelligence, September 10-13, 2007, Springer, Berlin, Heidelberg, ISBN:978-3-540-74564-8, pp: 489-493.

Mozina, M., J. Demsar, M. Kattan and B. Zupan, 2004. Nomograms for visualization of naive Bayesian classifier. Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 20-24, 2004, Pisa, Italy, pp: 337-348.

Platt, J., 2000. Fast training of Support Vector Machine using Sequential minimal optimization. Microsoft Way Redmond, Redmond, Washington, USA.

Toepfer, M., P. Kluegl, A. Hotho and F. Puppe, 2010. Conditional random fields for local adaptive reference extraction. Department of Computer Science, Wurzburg, Germany.