SMAD: Text Classification of Arabic Social Media Dataset for News Sources

Amira M. Gaber[1], Mohamed Nour El-din[2], Hanan Moussa[3]
Information System Department, Faculty of Computer and Artificial Intelligent, Cairo University, Giza, Egypt[1, 2, 3]
Higher Institute of Computer Science and Information Systems, Culture and Science City Academy, 6 October-Giza-Egypt[1]

***Abstract*—Due to the advances in technology, social media has become the most popular means for the propagation of news. Many news items are published on social media like Facebook, Twitter, Instagram, etc. but are not categorized into various different domains, such as politics, education, finance, art, sports, and health. Thus, text classification is needed to classify the news into different domains to reduce the huge amount of news available over social media, reduce time and effort for recognizing the category or domain, and present data to improve the searching process. Most existing datasets don't follow pre-processing and filtering processes and aren't organized based on classification standards to be ready for use. Thus, the Arabic Natural Processing Language (ANLP) phases will be used to pre-process, normalize, and categorize the news into the right domain. This paper proposes an Arabic Social Media Dataset (SMAD) for text classification purposes over the social media using ANLP steps. The SMAD dataset consists of 15,240 Arabic news items categorized over the Facebook social network. The experimental results illustrate that the SMAD corpus gives accuracy of about 98% in five domains (Art, Education, Health, Politics, and Sport). The SMAD dataset has been trained tested and is ready for use.**

Keywords—Text classification; Arabic text classification; Arabic Natural Language Processing (ANLP)

I. INTRODUCTION

Recently, the news media has transformed from hardcopy like newspapers, radios, and magazines to digital forms integrated with the internet to organize social media platforms like Facebook, Twitter, blogs, channels, and other digital media formats. Online social media has become a great way to connect people with each other around the whole world. Users of social media share news, communicate with other people, and create more posts and tweets related to the news than they consume. Consequently, a huge amount of incredible news is created and propagated through social media, which has a serious impact on society and individuals. Various social media needed to categorize their news into different domains, like politics, education, finance, art, sports, and health. So text classification is used to reduce the huge amount of news available over the social media. It is useful for reducing time and effort for recognizing the category or domain, and the data will be pretreated to improve the searching process and performance of classification.

The online news published on social media propagates over the network in different languages and formats, such as texts, images, videos, and unstructured formats. It is difficult to detect and classify the news and check its veracity, especially in the Arabic language, where it needs human expertise. However, Arabic Natural Language Processing (ANLP) are computational techniques that can be used for identifying the reality of text news based on facts and handling the Arabic language.

A. Arabic Language

This work concentrates on Arabic language news. The Arabic language is one of the greatest languages in the world. As this language possesses special spelling, grammatical rules, and punctuation marks. However, the text classification for this language is a challenging task because its structural essentials are complex. In the Arabic language, the text consists of some features which can be classified into external or internal features. The external features do not relate to the content of the text document which includes the author name, publication date, publishing house, etc. In contrast, the internal features relate to the text content and its linguistics features including lexical words and grammatical characteristics [1]. As a result, the following is the characteristics of the Arabic language structure and should be treated and deal with them to classify the news into the correct domain:

- The direction of Arabic language reading and writing from right to left.
- This language possesses 28 letters and there aren't any upper-case letters
- Arabic pronouns can be a singular, dual, or plural; masculine or feminine pronoun.
- Has three grammatical cases: nominative, accusative, and genitive.
- Words are classified into three main parts of speech, nouns, verbs, and particles.
- Nouns include adjectives and adverbs.
- Arabic verbs can have suffixes that change the overall meaning or the tense of the word.
- All verbs and some nouns have morphological rules.
- Arabic sentence can be a noun phrase or verbose phrase in which the verb can be passive.
- The subject pronouns may be removed from the sentence [1] [4].

B. Arabic Natural Processing Language (ANLP)

Most researchers tend to do and implement a set of tools that can be used in Arabic Natural Language Processing (ANLP) to help in the preparations for the processing structure for it. Because the Arabic language is the widest language due to the number of users using it and according to research introduced in 2015, it is the mother tongue of over 300 million people [2].

There are various developing tools and applications like tokenizers, Part of Speech (POS), Bag of Words (BOW), sentence segmentation, syntactic parsers, Matchers, etc. and various approaches are used for classification such as:

- Lexicon-based approach: The concerned data will be classified into a class or more based on linguistic rules (lexicon-based).
- Machine learning (ML) approach: The classification processes can be supervised, or unsupervised, or semi-supervised learnings. The seven stages of ML classification shown in Fig. 1.
- Hybrid approach: The classification integrates between the rule-based approaches besides the ML-based approach to achieve the optimum performance [3].

This paper presents SMAD dataset, a new Arabic social media dataset built across Facebook social media for news sources using the hybrid approach ANLP standard classification to cover five different domains (Sports, Arts, Health, Education and Political) domains. In addition, several algorithms in the ML approach will be used to help the classification with the principle of ANLP to reach high accuracy classification. The KNN algorithm will be used as a classifier. After the data was assembled and organized, the pre-processing methods and filtering are applied to make the data ready using ANLP.

This paper is organized as follows. Section II introduces the related work for the existing Arabic corpus. Section III presents the text classification methodology. Section IV presents the applied methodology that shows the formation of SMAD corpus, Section V displays the experiment results. Section VI will discuss the proposed methodology with others. Finally, Section VII concludes this work.

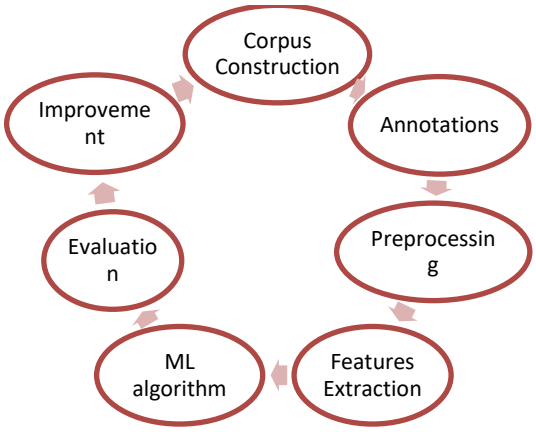


Fig. 1. Arabic Natural Language Processing (ANLP) Stages.

II. Related Work

There are several studies categorize the text and build datasets tested against the quality measurement metrics. Riyad et al. [4], proposed an Arabic text classifier based on the Document Frequency threshold (DF) besides the Support Vector Machines (SVM) algorithm with a precision of 0.95. The datasets were collected from various Arabic newspaper online websites like Al-Jazeera, Al-Hayat, Al-Ahram, Al-Nahar, and Al-Dostor.

Syiam et al. [5], proposed a new Arabic text classification depending on a Hybrid approach of document frequency. Egyptian newspapers like El-Ahram, El-Akhbar, and El-Gomhoria were used to collect datasets. The used classifiers are the Key Nearest Neighbor (KNN) and Rocchio the classifier performance accuracy was 0.98.

Harrag et al. [6], performed a classification in which data collected from the Arabian scientific encyclopedia. They used the decision tree algorithm with an accuracy of 0.93.

Chantar and Corne [7], applied the Particle Swarm Optimization (PSO) Algorithm with Support Vector Machin (SVM) classifier. Datasets were collected from specialized web sites as Al-Jazeera, Al-Hayat, and Al-Ahram. The result of the classification is more accurate and efficient.

In Saraç and Ayşe Özel [8], they concentrate on web documents they use a Firefly Algorithm and J48 classifier to test data from WebKB and Conference datasets. The applied algorithm returned an accuracy (between 0.56 and 0.93).

Rohaidah et al. [9], introduced a new Sentiment Analysis approach using the k-NN classifier. The dataset was collected from customer review datasets with maximum precision results equal to 0.892.

Guessoum et al. [10], performed a text classification on OSAC3 corpus (Open-Source Arabic Corpora) which was gathered from several different websites (BBC Arabic, CNN Arabic, etc.). It contains 22,429 textual records. Every text document is a part of one of ten separate categories (Economics, History, Religion, Health, Education and Family, Sports, Astronomy, Law, Stories, and Cooking Recipes).

There are number of Arabic datasets like, DAA [11] is a dataset in which nine categories have been processed and standardized with 400 documents for each category, Akhbar-Alkhaleej [12] is a popular Arabic Dataset with 5690 Arabic news documents gathered regularly from the online newspaper "Akhbar-Alkhaleej". It consists of five categories: Alwatan [13] is an Arabic Dataset with 20,291 Arabic news documents collected regularly from its online newspaper, Al-Jazeera-News [14] Arabic Dataset (Alj-News) is an Arabic dataset with 1500 documents. It consists of five categories (Sport, Economy, Science, Politics, and Art), NADA, [11] is an Arabic dataset consists of two corpora OSAC and DAA it used a s (Dewey Decimal Classification scheme (DDC) and Synthetic Minority Over-Sampling Technique (SMOKE) to reprocess and filtering to enhance the results to reach high accuracy.

All previous work concentrated on how to classify the articles using different classifiers which collected its precision and accuracy measurements. All Arabic datasets collected data

from web sources and didn't take into consideration the sources of the news collected from the social media sources. This paper will construct a new dataset collected from different news websites (BBC Arabic, Al-Watan, El-youm7, etc.) to classify the news over the Facebook social media because it is a widest mean for spreading news using the text classification methodology for the Arabic Natural Language Processing (ANLP).

III. Text Classification Methodology

Text classification is a process of retrieving strong meaningful bulk of text [15] then segmenting them into meaningful sentence, topic, words or character for the text analysis [16]. There are many reasons for using text classification. One of the main reasons is the breaking down the text into smaller give more meaning and contrast than the whole document. Another one is the smaller text useful in accessing and analyzing the text.

A. Arabic Text Classification Steps

Fig. 2, shows the text classification process steps which are: 1) stemming, which returns back to the root of the word; 2) stop word removal: it removes unnecessary words, 3) Indexing is the process of creating an internal representation of the documents. It consists of three phases: a) Construction, which builds a vector consisting of all the words that appear in the document; b) Term Selection: choose the main words according to some criteria; c) Term Weighting: count how many occurrences the main word appears; 4) classifier construction, which learns the classifier for each category characteristics by training it on a set of documents to classify correctly; 5) Evaluate Classifier Which Apply a test set to check whether the classification process completed correctly or not [5].

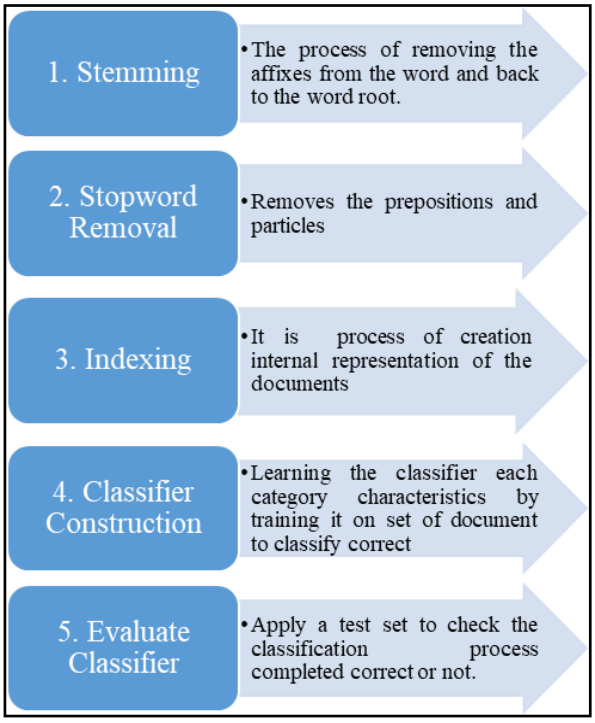


Fig. 2. The Arabic Text Categorization Steps.

IV. SMAD Methodology

This work focused on classifying Arabic news into different domains for each news source that published news on Facebook. It can be considered as a new step in classification and detection of content targeted at the Arabic language over Facebook and social media. For this purpose, Arabic Natural Language Processing (ANLP) and machine learning are used, to achieve this purpose. Fig. 3 shows the phases of SMAD formation methodology.

Fig. 3. SMAD Methodology Phases.

A. Dataset Collection Phase

Data collection step it is a crucial step to gather the information of the research and it is accuracy depend it for all rest of steps [17]. In this step, the SMAD dataset will be collected. It is new corpora dataset consist of 15,240 textual topic of news collected from (BBC Arabic, Al-Watan, El-youm7, etc.) websites scrapped for five domains: Sports, Art, Education, Health and Political with size 2MB. For our study, the randomly train data about 2000 textual topics in different domain for training phase and about approximately 1000 textual topics news items extracted from Facebook social media for testing phase.

B. Dataset Construction Phase

This phase consists of two steps data preprocessing and indexing step.

1) Data Pre-processing step using ANLP: Data preprocessing considered as different transformations applied to the data (data gathered from various sources different in style which are not feasible for the analysis) before introducing it to the classification methodology. Data preprocessing is important to raise the efficiency of machine learning algorithms to put the data into a suitable form for the next processing steps [12] to facilitate the analysis of data. the SMAD methodology will follow steps of Arabic text classification explained in details in Section III:

- The Stemming and stop words removal steps, execute the stemming process which includes the stop word process by using the "Root-Based Stemmer" technique to remove the affixes from the word and back to the word root, matches the root word against a set of suggested 67 patterns that represent most of word forms to reduce the number of words used in the indexing step to get more accurate results and finally executes the stop word removal step which removes the prepositions and particles of the word. Fig. 4 will explain the steps applied in this step.
  - Remove all numbers exists in the text.
  - Splits text into words.

- Preprocess the content of the news like the header of the news.
  - Removes the stop words from the text like (على، عن، إلى ، من) / (on, on, to, from).
  - Removes the punctuation letters or spaces Such as comma and semicolon, question marks and exclamation marks.
  - Removes the diacritics like (ً, ٌ, ٍ, ّ, ُ, ِ, َ, ْ) which are signs above or below letters. used the grammatical case.
  - Removes the non-Arabic letters and special. symbols in text.
  - Removes all words which have a small. length (أنا (me) / oh (يا)).
- Deletes the repeated words and specifies the distinct terms from each news item and records the number of repetitions for each word.

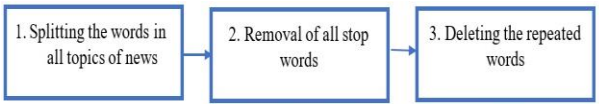


Fig. 4. Data Pre-processing Phase.

2) The Indexing step: This step executes the indexing process which consists from two phases.

a) Term selection, tokenization technique will be used to split text into words, symbols, phrases as a token then construct a super vector which contains all most important terms.

b) Term weighting, there are four techniques to weight the terms: *i) Boolean weighting* it give the result 1 if the frequency word in the text greater than 0, otherwise give zero frequency, *ii) term Frequency weighting* it records the frequency of all words in the document, *iii) Term Frequency-inverse weighting (TF-IDF)* it is used to measure how many times an important word exists in a document, *iv) Normalized-TF-IDF weighting :* similar to TF-IDF but take into its consideration the different length of the text [5].

This phase implemented by using the TF-IDF technique because this technique constructs a vector contains the most important words and the less important ones as well, by using the scoring schema and take into consideration the rarity of words.

For a term T present in document D [18]. Eq. 1 specifies the TF-IDF formula.

$$TF\_IDF = TF \times IDF \quad (1)$$

The Term Frequency (TF) is used to measure that how many times a term exists in a document. It is calculated by Eq. (2) [19].

$$TF = \left[\frac{\text{Number of times term appearance in text}}{\text{Number of all words in the text}}\right] \quad (2)$$

The IDF (Inverse Document Frequency) is an approach to measure the importance of a particular word that can be measured by taking the logarithm for the output of dividing the number of all topics by the number of topics containing the text. Calculated by Eq. (3).

$$IDF = LOG\left[\frac{\text{Number of all Terms}}{\text{Number of terms contain the text}}\right] \quad (3)$$

The following algorithm constructs a vector of the important words in each domain and then weights them.

The higher the TF-IDF weight value of the term, the stronger relationship to the text they appear in [18].

***Algorithm 1**: Indexing Algorithm*

***Input**: News Titles Scrapped from multiples source $s_1, s_2 \ldots s_i$*
***Output**: News id, News Title, News Domain*
*1. **for** all articles A_0 to A_i*
2. tokens= tokenize (A_i)
*3. **for** all $t_i \in$ tokens **do***
*4. **if** t_i **not** a stop word **or** pun **or** small length*
5. tokenlist[] = t_i
*6. **end if***
7. score= TF_IDF (t_i)
8. tokenvector.put=(t_i, score)
9. write t_i in the domain file
*10. **end for***
*11. **return** (News_id, News_title, domain)*

The following is the pie chart illustrates the number of training data used in each domain and testing data shown in Fig. 5.

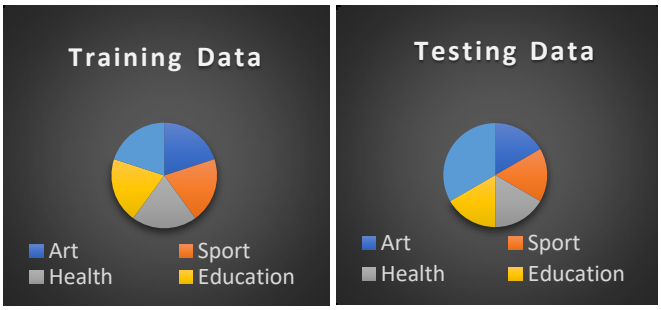


Fig. 5. Training and Testing Data Pie Chart.

C. Dataset Evaluation Phase

This phase consists of two steps Build the classifier and Evaluate Classifier using the performance matrices.

1) Build classifier: In this step, the KNN algorithm used as a classifier , it is an efficient technique for text classification, it is used to test given text to be classified. This algorithm searches for the k nearest neighbors among the pre-classified training text based on similarity measure, and ranks these similarity scores, the prediction of the correct category of the test text decided by the weight of the candidate categories, if more than one neighbor belong to the same category then the sum of their scores is used as the weight of that category, the category with the highest score is assigned to the test text provided that it exceeds a predefined threshold, more than one category can be assigned to the test text [20]. For a similarity measure calculated by Eq. (4).

$$\text{Sim}(A,B) = \frac{\sum_{i=1}^{r} w_{ij} \times w_{ik}}{\sqrt{\sum_{i=1}^{r} w_{ij}^2} \sqrt{\sum_{i=1}^{r} w_{ik}^2}} \quad (4)$$

Where A and B are the two verctors constructed to get the similiarty between them

,i,k are the text representing A and B.

, r is the number of terms in the feature space .

2) Evaluate classifier: All the news of the corpus is collected and handled using new corpus using TF-IDF technique, the number of trained texts used about 2000 news items for each domain and now it needs to input test data for the classifier to categorize news to the appropriate classes. The performance metrics are the best indicators to evaluate the classifier. There are four basic quality metrics of any classifier Precision (P), Recall (R), F-measure, and accuracy.

- Precision: it is the answer of the question; what are the of positive identifications were actually correct?
- Recall: it is the answer of the question, what are the actual positives of the data was identified correctly?
- F-score: is the harmonic mean of Precision and Recall, as precision and recall alone cannot provide the best evaluation of the model.
- Accuracy / classification error: is the performance measure, it is a ratio of correctly predicted observation to the total observations.

To calculate each of these measures, it's should define the following:

- TP (true positive) – the set of news that are in the correct category and are predicted truly.
- TN (true negative) – the set of news that are not in the correct category and are predicted to be in a different category.
- FP (false positive) – the set of news that are in a different category and are predicted to be in the correct category.
- FN (false negative) – the set of news that are in a different category and were predicted false category.

These four performance quality metrics are measured by the following equations [21].

- Precision (p) = $\frac{T_p}{T_p+F_p}$
- Recall (R) = $\frac{T_p}{T_p+F_n}$
- F-measure = $\frac{2PR}{(R+P)}$
- accuracy = $\frac{T_p+T_n}{T_p+T_p+f_p+f_n}$

V. Experimental Results

In this section, the evaluation and the performance of the proposed methodology is shown using six real datasets. Section A will present experimental setup details, including dataset construction by using the Arabic text categorization steps, and measure its performance by calculating its accuracy and quality metrics measurements. Section B will present the main results of the accuracy and quality metrics (Recall, precision, and F measure) of the SMAD dataset and then compares the performance improvement of the personalized model, with other similarity metrics for baseline datasets in different domains.

A. Experimental Setup

Datasets: To evaluate our model, experiments were conducted on 15,240 news items collected from the (BBC Arabic, Al-Watan, El-youm7, etc.) website and Facebook in five domains, like sports, political health, art, politics, and education. The dataset was collected at a size of 2 MG with different files. Each file corresponds to one domain generated as a CSV file. The "SMAD" dataset trained on the 2000 news items and tested on a data benchmark of 1000 news item for each domain using the KNN classifier to classify the news into the correct domain at a specific time. Fig. 6 shows the performance of classifying SMAD through the recall, precision, and F1 measure quality metrics.

The proposed model will be compared with other baseline models on these six datasets compared with the previous Arabic datasets OSAC, DAA, Akhbar-Alkhalee, Aljezzera, NADA, and Alwatan corpus used in categorizing domains in the Arabic language w.r.t. the recall, precision, and F1 measure quality metrics.

Fig. 7 shows the comparison of accuracy measurement for all datasets, SMAD corpus gives accuracy of about 98% in five domains while accuracy of 98% while the accuracy of NADA is 93.8792%, accuracy of OSAC is 98.1758%, accuracy of DAA is 80.9087%, accuracy of Alj-News is 93.1%, accuracy of Alwatan is 96.1% and the accuracy of AkhbarAlkhaleej is 88.7%.

B. Main Results

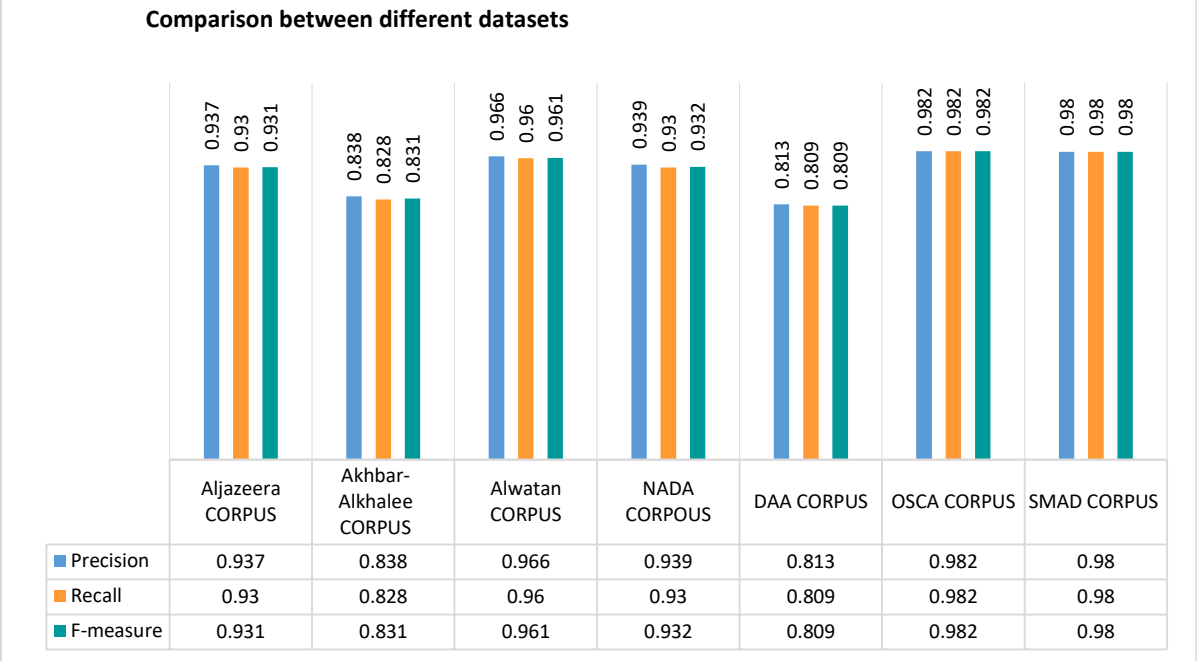
The performance quality measurements of the proposed model will summarize the key observations and show a detailed comparison for the mentioned Arabic datasets with respect to recall, precision, and F-measures. These performance metrics are specialized in different domains and record the weighted average for each domain in each news source shown in Table I.

C. Methodology Implementation Architecture

Fig. 8 illustrates the Data Flow Diagram (DFD) of the proposed methodology, which shows how the SMAD dataset implemented in the following steps:

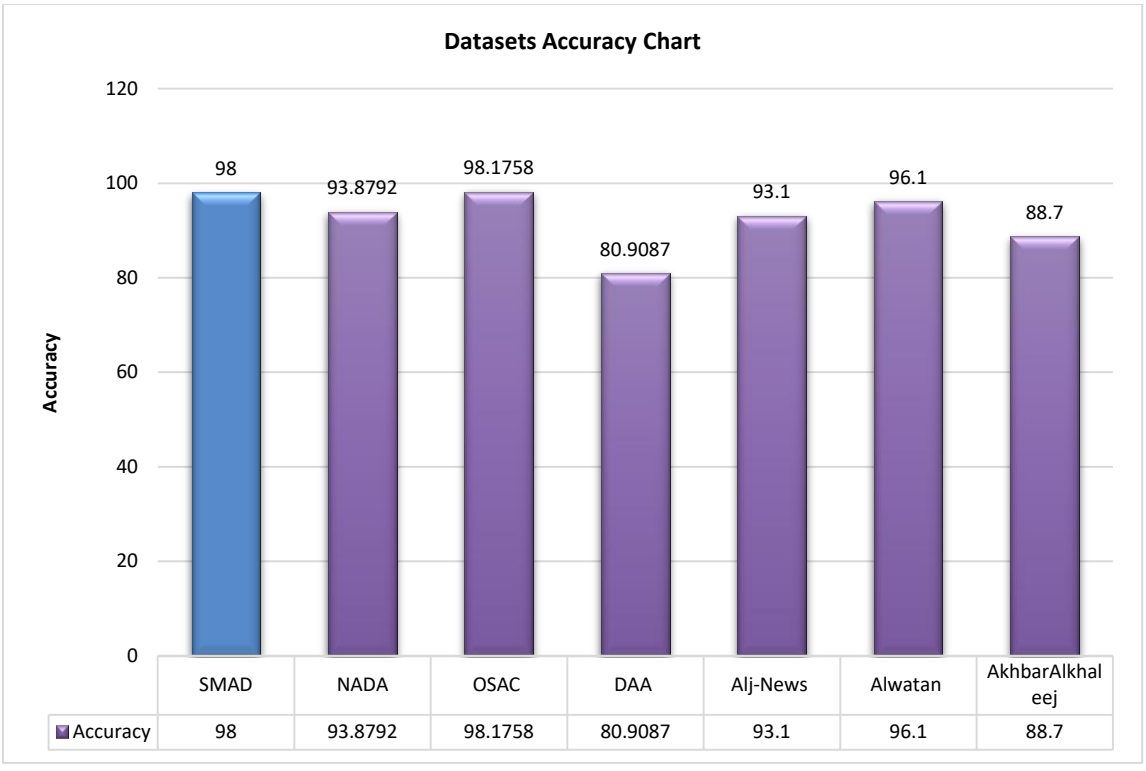
- The news source will generate the news on its web page. Then, it publishes a sample of this news on the Facebook page for the interaction.

- The BOW Scrapper will pre-process posts and extract the bag of words (BOW) to train the classifier to choose the correct domain in the future. This step was completed by using Puppeteer library to scrap Facebook data.
- After the classifier extracted the data and is trained on the BOW, the dataset will be built as a Jason file then transform it into CSV files using unicodecsv library.
- To evaluate the dataset, the sklearn.metrics library will be used to calculate the model confusion_matrix (precision – recall – f-measure) and the accuracy_score for this classifier.



	Aljazeera CORPUS	Akhbar-Alkhalee CORPUS	Alwatan CORPUS	NADA CORPOUS	DAA CORPUS	OSCA CORPUS	SMAD CORPUS
Precision	0.937	0.838	0.966	0.939	0.813	0.982	0.98
Recall	0.93	0.828	0.96	0.93	0.809	0.982	0.98
F-measure	0.931	0.831	0.961	0.932	0.809	0.982	0.98

Fig. 6. Performance Quality Metrics Measurements of different Datasets with SMAD Dataset.



	SMAD	NADA	OSAC	DAA	Alj-News	Alwatan	AkhbarAlkhaleej
Accuracy	98	93.8792	98.1758	80.9087	93.1	96.1	88.7

Fig. 7. Accuracy Comparison between SMAD, NADA, OSAC, DAA, Alj-News, AlWatan and Akhbar-Alkhaleej Datasets.

TABLE I. COMPARISON BETWEEN OSAC, DAA, ALJEZZERA, NADA, AKHBAR-ALKHALEE, ALWATAN AND SMAD CORPUS

CORPORA	CLASS	CLASSIFIER	# OF CLASSES	Precision	Recall	F- Measure
Alj-News CORPUS	SPORT	SVM	5	1	0.983	0.992
	ART			0.934	0.95	0.942
	SCIENCE			1	0.933	0.966
	POLITICAL			0.789	0.933	0.855
	ECONOMIC			0.962	0.85	0.903
	Weighted average			0.937	0.93	0.931
Akhbar-Alkhalee CORPUS	**CLASS**	**CLASSIFIER**	**# OF CLASSES**	**Precision**	**Recall**	**F- Measure**
	Economy	SVM	4	0.821	0.836	0.829
	Int. News			0.98	0.845	0.907
	Local News			0.835	0.917	0.874
	Sport			0.975	0.895	0.933
	Weighted average			0.838	0.828	0.831
Alwatan CORPUS	**CLASS**	**CLASSIFER**	**# OF CLASSES**	**Precision**	**Recall**	**F- Measure**
	Culture	SVM	4	0.838	1	0.912
	Economy			0.892	0.943	0.946
	Religion			1	0.978	0.989
	Sport			0.991	0.972	0.981
	Weighted average			0.966	0.96	0.961
NADA CORPUS	**CLASS**	**CLASSIFIER**	**# OF CLASSES**	**Precision**	**Recall**	**F- Measure**
	Arabic Literature	SMOTE	10	0.920	0.927	0.926
	Social science - economy			0.908	0.884	0.871
	Social science - politics			0.948	0.950	0.944
	Social science - law			0.887	0.896	0.884
	Sport			0.967	0.964	0.959
	Art-General			0.977	0.973	0.970
	General Religions - Islam			0.918	0.933	0.925
	Applied science – computer science			0.912	0.925	0.917
	Applied and health sciences			0.969	0.964	0.960
	Pure Astronomy Science			0.967	0.973	0.925
	Weighted average			0.939	0.939	0.932
DAA CORPUS	**CLASS**	**CLASSIFIER**	**# OF CLASSES**	**Precision**	**Recall**	**F- Measure**
	أدبيات – الادب العربي	TF-IDF	9	0.770	0.760	0.765
	علوم اجتماعية - اقتصاد			0.675	0.856	0.755
	علوم اجتماعي - سياسة			0.485	0.436	0.459
	علوم اجتماعية قانون			0.783	0.720	0.750
	رياضة			0.970	0.953	0.961
	فنون - عام			0.893	0.917	0.905
	ديانات - اسلام			0.861	0.812	0.836
	علوم بحتة – علوم كمبيوتر			0.863	0.805	0.833
	علوم تطبيقية – علوم صحية			0.789	0.723	0.755
	Weighted average			0.813	0.809	0.809
OSCA CORPUS	**CLASS**	**CLASSIFIER**	**# OF CLASSES**	**Precision**	**Recall**	**F- Measure**
	علوم اجتماعية - اقتصاد		6	0.965	0.985	0.984
	علوم اجتماعية قانون			0.970	0.975	0.984
	رياضة			0.966	0.971	0.965
	ديانات - اسلام			0.958	0.959	0.943
	علوم بحتة – علوم فلك			0.999	0.999	0.997
	علوم تطبيقية – علوم صحية			0.996	0.996	0.996
	Weighted average			0.982	0.982	0.982
SMAD CORPUS	**CLASS**	**CLASSIFIER**	**# OF CLASSES**	**Precision**	**Recall**	**F- Measure**
	SPORTS	KNN	5	0.95	1.00	0.98
	EDUCATION			0.98	1.00	0.99
	ART			1.00	0.96	0.98
	HEALTH			0.99	0.97	0.98
	POLITICAL			0.98	0.97	0.98
	Weighted average			0.98	0.98	0.98

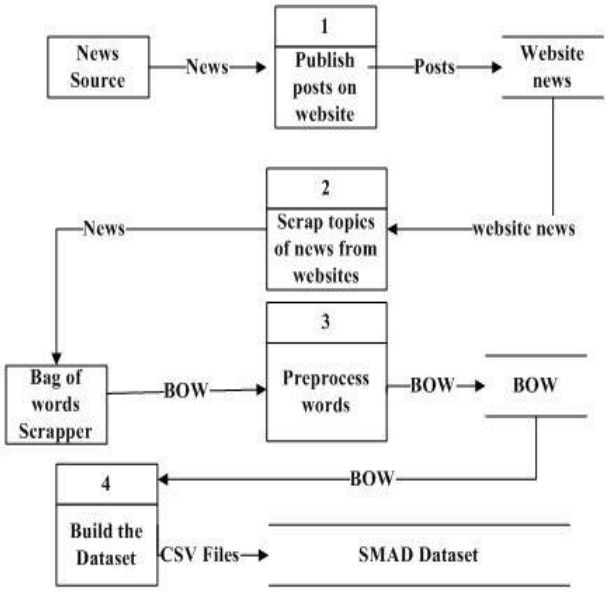


Fig. 8. SMAD Methodology DFD.

VI. Discussion

All the pervious Arabic datasets [10-14] classify the articles using different classifiers for various purposes. All these Arabic datasets didn't take into consideration the fact that news spread rapidly over social media. Facebook is the widest means of social media as news plays an important role in spreading rabidly on it but is not classified into domains. Thus, the purpose of this paper is to classify the news that is widely spread on Facebook by constructing the SMAD dataset in order to save time and effort in recognizing domain news and to improve the search process for specific news at a specific time. The SMAD corpus was compared with other different datasets according to quality measurement metrics (precision, recall, F-measure, and accuracy). In the sports domain, its precision is 0.95, recall is 1 and the F-measure is 0.98. In the education domain, its precision is 0.98, recall is 1 and the F-measure is 0.99. In the arts domain, its precision is 1, recall is 0.96 and F-measure is 0.98. In the health domain, its precision is 0.99, recall is 0.97 and F-measure is 0.98. Finally, the political domain precision is 0.98, recall is 0.97, and F-measure is 0.98. The accuracy of the SMAD dataset is about 98% in five domains while the accuracy of NADA is 93.8792%, accuracy of OSAC is 98.1758 %, accuracy of DAA is 80.9087 %, accuracy of Alj-News is 93.1 %, accuracy of Alwatan is 96.1 % and the accuracy of Akhbar Alkhaleej is 88.7 %.

VII. Further Work

For the future, a news benchmark will be needed for social media from the most credible news sources in a specific domain at a specific period of time to facilitate the searching process to reduce time, effort and checking the veracity of the news from the most credible sources.

VIII. Conclusion

This study was done to construct a new Arabic Dataset corpus built from several websites to classify the news spreads over social media means. Facebook has become one of the news sources. The Facebook social media's source news is not categorized into any domain; this corpus is composed of five domains (Art-Health-Education-Politics-Sports) and can be extended easily by adding new various domains which can be used for several purposes in the Arabic text classifications. This dataset goes through predefined stages of pre-processing and filtering to eliminate the anomalies of the data, then tested and validated using KNN classifier with four evaluation measures: precision, recall, F- measure, and accuracy for each domain. The experiment results deduced that the new corpus is an efficient dataset for the Arabic classification of news with an accuracy of 98%.

References

[1] S.L. Marie-Sainte, and N. Alalyani," Firefly algorithm based feature selection for Arabic text classification", Journal of King Saud University-Computer and Information Sciences, vol:32, No. 3, pp.320-328,2020.

[2] Y. Jaafar, K. Bouzoubaa, "A survey and comparative study of Arabic NLP architectures", In: Intelligent Natural Language Processing: Trends and Applications, Springer, Cham, pp. 585-610, 2018.

[3] MA. Omari, M. Al-Hajj," Classifiers for Arabic NLP: survey"," International Journal of Computational Complexity and Intelligent Algorithms", vol:1, No. 3, pp. 231-58.

[4] R. Al-Shalabi, G. Kanaan, and M. Gharaibeh, "Arabic text categorization using KNN algorithm", In: Proceedings of The 4th International Multiconference on Computer Science and Information Technology, Vol. 4, pp. 5-7, 2006.

[5] M.M. Syiam, Z.T. Fayed, M.B. and Habib, "An intelligent system for Arabic text categorization"," International Journal of Intelligent Computing and Information Sciences", vol. :6, No.1, pp.1-19, 2006.

[6] F. Harrag, E. El-Qawasmeh and P. Pichappan, P.," Improving Arabic text categorization using decision trees", In: 2009 First International Conference on Networked Digital Technologies, IEEE, pp. 110-115, 2009, July.

[7] H.K. Chantar and D.W. Corne," Feature subset selection for Arabic document categorization using BPSO-KNN", In: 2011 Third World Congress on Nature and Biologically Inspired Computing, IEEE, pp. 546-551, 2011, October.

[8] E. Saraç, and S.A. Özel, "Web page classification using firefly optimization", In :2013 IEEE INISTA, IEEE, pp. 1-5, 2013, June.

[9] S.R. Ahmad, N.M.M. Yusop, A.A. Bakar and M.R. Yaakub, "Statistical analysis for validating ACO-KNN algorithm as feature selection in sentiment analysis", In: AIP conference proceedings, AIP Publishing LLC. Vol: 1891, No. 1, p. 020018, 2017, October.

[10] R. Belkebir, and A. Guessoum, "A hybrid BSO-Chi2-SVM approach to Arabic text categorization", In :2013 ACS International Conference on Computer Systems and Applications (AICCSA), IEEE, pp. 1-7, 2013, May.

[11] N. Alalyani and S. L. Marie-Sainte, "NADA: New Arabic dataset for text classification". "International Journal of Advanced Computer Science and Applications", vol.: 9, No.9, 2018.

[12] M. A. Abdeen, S. AlBouq, A. Elmahalawy and S. Shehata," A closer look at arabic text classification"," International Journal Advanced Computer Science Applications", vol:10, No.11, p.p. 677-688, 2019.

[13] B. Hawashin, A. Mansour and S. Aljawarneh, "An efficient feature selection method for Arabic text classification", "International journal of computer applications", Vol:83, No. 17, 2013.

[14] M. M. Al-Tahrawi, "Arabic text categorization using logistic regression", "International Journal of Intelligent Systems and Applications", VOL:7, No.6, p. 71, 2015.

[15] I. Pak and P.L. Teh, P.L," Text segmentation techniques: a critical review"," Innovative Computing, Optimization and Its Applications", pp.167-181,2018.

[16] P. Badjatiya, L.J. Kurisinkel, M. Gupta and V. Varma," Attention-based neural text segmentation", In: European Conference on Information Retrieval, Springer, Cham, pp. 180-193., 2018, March.

[17] R. Mouty and A. Gazdar, “Survey on Steps of Truth Detection on Arabic Tweets”, In: 2018 21st Saudi Computer Society National Computer Conference (NCC), IEEE, p.p. 1-6, 2018, Apr 25.

[18] J.N. Singh and S.K. Dwivedi, “Comparative analysis of IDF methods to determine word relevance in web document”. “International Journal of Computer Science Issues (IJCSI)”, Vol:11, No.1, p.59,2014.

[19] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents", “International Journal of Computer Applications,” vol. 181, No. 1, pp. 25-29, 2018. Available: 10.5120/ijca2018917395.

[20] Al-Shalabi, R., Kanaan, G., & Gharaibeh, M. (2006, April). Arabic text categorization using KNN algorithm. In Proceedings of The 4th International Multiconference on Computer Science and Information Technology (Vol. 4, pp. 5-7).

[21] R. Soleymani, E. Granger and G. Fumera, “F-measure curves: A tool to visualize classifier.