An Integrated Framework for Relevance Classification of Trending Topics in Arabic Tweets

Abdullah M. Alkadri*, Abeer ElKorany, Cherry A. Ezzat
Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt

***Abstract*—Social media platforms such as Twitter are a valuable source of information about current events and trends. Trending topics aim to promote public events such as political events, market changes, and other types of breaking news. However, with so much data being generated, it would be difficult to identify relevant tweets that are related to a particular trending topic. Therefore, in this paper, an integrated framework is proposed for the detection of the degree of relevance between Arabic tweets and trending topics. This framework integrates natural language processing, data augmentation, and machine learning techniques to identify text that is likely to be relevant to a given trending topic. The proposed framework was evaluated using a real-life dataset of Arabic tweets that was collected and labeled. The results of the evaluation showed that the proposed framework achieved the highest macro F1 score of 82% in binary classification (relevant/irrelevant) and 77% in categorical classification (degree of relevance), which outperforms the current state of the art.**

***Keywords*—Trending topics; social media platforms; machine learning; Arabic relevance classification; data augmentation**

I. Introduction

In recent years, social media platforms have become a significant source of information for real-time events and trending topics [1]. The vast amount of user-generated content on these platforms, especially Twitter, provides a wealth of information for various applications, including sentiment analysis, event detection, and trend analysis [2].Twitter is now a real-time information distribution channel that is utilized for news, politics, and advertising [3]. Users can use the trending topics feature or popular hashtags to discover current popular news. However, trending topics are often irrelevant to the content being discussed. Therefore, trending topics relevance classification is considered as an important text analysis task for Twitter data. Trending topics relevance aims to identify relevant tweets that are related to a particular trending topic. However, this task is challenging due to the noisy and unstructured nature of social media text, misspellings, slang, and various other factors that can affect the relevance of a tweet to a trending topic [4], [5].

Currently, the hashtag trending topic feature is used by users to find out what topics are currently popular on Twitter [6], [7]. Some users employ trending topics, to get more attention to their tweets. However, the trending topic is irrelevant to the topic being discussed by the tweet itself. For example, some people include political trending topics that are popular during the election period in their tweets, but the tweet content does not include any political topics.

Irrelevant content for trending topics reduces the level of communication offered by social media networks. They pollute social networks and affect how people perceive the contents of the internet. The user experience will significantly decrease if someone is excessively exposed to irrelevant information, which would result in user losses for the social service provider [8]. Therefore, social platforms must develop algorithms to filter unwanted information and determine the relevance of content to trending topics.

Several approaches have been proposed for automatic learning and detection including trending topics detection on social media [9]. Previous research has extracted and utilized a wide range of features, from simple to complex, as well as a wide range of learning and classification algorithms, from traditional machine learning techniques to deep learning [9], [10].

However, these approaches face several challenges specific to Arabic language processing, such as the absence of diacritics, the presence of multiple dialects, and the use of Arabic script, which can affect the accuracy of the model. Additionally, the availability of labeled datasets for Arabic text relevance classification in trending topics is limited, making it difficult to train and evaluate models on this task.

This paper focuses on the problem of detecting the relevance level of Arabic texts related to trending topics on Twitter. Specifically, we aim to identify tweets that are relevant to a given trending topic in order to be able to filter out irrelevant content. A novel framework that integrates several techniques, including text preprocessing, feature extraction, and machine learning algorithms is proposed to classify tweets associated to trending topics into relevant/irrelevant. Furthermore, this framework is able to categorize the relevant content into low, medium, and high based on its degree of relevancy. This work is based on the trending topics and tweets related to Yemeni politics written in Arabic.

The main contributions of this paper are the following:

- Build a public and available dataset of Arabic tweets related to trending topics in both binary and categorical classes.
- Develop a trending topics relevance text classification framework using machine learning algorithms with two scenarios: binary and categorical classes.
- Apply data augmentation to enhance the performance of the framework.

*Corresponding Authors

As will be explained later, to overcome the challenge of limited labeled datasets, data augmentation techniques [11] were applied to improve the performance of the model. Several techniques will be applied including word embedding-based techniques, to generate additional training data, which helps the model better generalize to unseen data and handle noisy, misspelled tweets and improve the overall performance of the model.

The proposed framework is beneficial for researchers, journalists, and businesses who require analysis of trending topics on social media platforms. It can also be used to extract insights about public opinions, monitor the impact of events, and discover emerging trends. The rest of the paper is organized as follows: Section II discusses the background and related work; Section III explains the methodology used in our approach; Section IV presents the experimental results; Discussion are in section V; Finally, section VI concludes the paper and outlines possible directions for future work.

II. Background and Related Work

A. Trending Topics

Trending topics refer to popular and widely discussed topics on social media platforms such as Twitter, Facebook, and Instagram. These topics are characterized by a large volume of posts or tweets that use a specific hashtag or keyword, and they often reflect current events, news, and opinions that are of interest to the public [6].Trending topics are important because they provide valuable insights into public opinion and social trends. They allow individuals and organizations to track and monitor the conversation around a particular topic, and they can be used to identify emerging trends and issues in real-time [12]. In recent years, the analysis of trending topics has become an important field of research. Researchers have explored various approaches and techniques to identify and analyze these topics based on different domains and languages. For example, some studies have focused on identifying trending topics related to politics, sports, entertainment, or health, while others have looked at trending topics in different languages such as English, or Arabic. Some of these approaches include [13], [14]:

- Keyword-based approach: This approach involves using a set of pre-defined keywords to identify the trending topics.
- Text classification approach: This approach involves training a classifier on a labeled dataset to identify the trending topics.
- Topic modeling approach: This approach involves using topic modeling techniques such as Latent Dirichlet Allocation (LDA) to identify the trending topics.
- Hybrid approach: This approach combines multiple methods to improve trending topic detection accuracy. For example, a hybrid approach may use both keyword-based and text classification methods to identify trending topics.

Overall, the choice of the approach depends on the specific requirements of the task, such as the available resources, the amount of labeled data, and the desired level of accuracy.

Trending topics analysis has many applications, including social media monitoring, reputation management, crisis management, and market research. It can be used by businesses, governments, and other organizations to gain insights into public opinion, track the effectiveness of their social media campaigns, and respond to emerging trends and issues in real time.

B. Related Work

1) Topic detection: There is limited research on trending topics relevance text classification in social media. However, significant research has focused on related areas such as topic detection and sentiment analysis [15]. Here, we only provide a brief overview of a few of the most pertinent studies for topic detection. There is a wide range of techniques employed for topic analysis on Twitter. Studies that utilize machine learning techniques generally rely on supervised learning [16], [17], [7], while others adopt a hybrid approach that incorporates latent Dirichlet allocation (LDA) [18]. The combination of sentiment analysis and topic detection has been employed to analyze content related to COVID-19 in Brazil and the USA [3]. Lee et al. [7] categorized Twitter Trending Topics into 18 broad categories, including sports, politics, and technology, using a text-based classification approach with a Bag-of-Words and a network-based classification.

There have been several significant works related to topic detection on social media for the Arabic language. An assimilated model was introduced to identify events from Arabic Twitter data by Alsaedi et al. [19]. Their main objective was to distinguish disruptive events from other events in social media data streams. The model they developed relies on the frequency of terms occurring together over time. In a different study [20], a comprehensive framework for event detection was introduced. The authors emphasized the importance of temporal, spatial, and textual characteristics of each cluster in event detection. They compared the effectiveness of their proposed framework with LDA and demonstrated that LDA was not suitable for analyzing short messages like tweets. In [21], a feature-pivot method was employed to identify bursty features of terms from Arabic Tweets. The approach employed TFIDF, entropy, and stream chunking to capture bursty terms that were highly relevant to a particular event during a given time interval. The document-pivot method was introduced in [22] to extract trending topics for Arabic Twitter users. These works serve as examples of the diverse range of approaches that can be utilized to detect topics on Twitter.

2) Relevance text classification of trending topics: Despite the abundance of research on topic detection on Twitter, comparatively less work has focused on classifying relevant text within trending topics, which is the primary focus of our proposed approach. A framework model known as TORHID (Topic Relevant Hashtag Identification) was introduced in a research paper [23] to identify and retrieve hashtags relevant to a particular topic on Twitter. The model utilized small tweets of a hashtag as seeds and employed a Support Vector Machine to classify new tweets as relevant or irrelevant. According to the reported results, the TORHID model achieved an accuracy of 67.25%. Cahyaniet et al. [4] conducted research on the relevance classification of tweet content and trending topics on social media. The study focused on political tweet data related

to Indonesian trending topics and employed Support Vector Machine (SVM) for classifying tweets as either relevant or irrelevant. The study reported an F1 measure of 70% for the applied model. However, the absence of research on trending topics relevance text classification in Arabic represents a noteworthy gap in the existing literature, which we aim to address.

III. Proposed Framework for Relevance Classification of Trending Topics in Arabic Tweets

As shown in Fig. 1, an integrated framework for Relevance Classification of Trending Topics in Arabic Tweets (RCTAT) is proposed. This framework consists of two main components: data preparation and augmentation and trending topics relevance text classification. In the following subsections, each component will be described in details.

A. Data Preparation and Augmentation

Data preparation involves collecting and cleaning the data, while data augmentation involves generating more data from the existing data to improve the model's generalization and reduce overfitting [24]. The following steps were taken in data preparation and augmentation for Relevance Classification of Trending Topics in Arabic Tweets:

1) Data collection: The first step was to collect Arabic tweets from Twitter. The tweets were collected using the Twitter Streaming API starting December 2019 to April 2020, using the query "lang: ar" (language is Arabic) and the trending hashtag (#YEMEN).

2) Data cleaning: To obtain distinct tweets, several cleaning steps were applied. These steps involved removing diacritics, repeated characters, and punctuations from the tweets. Additionally, both Arabic and non-Arabic alphabets were normalized to ensure consistency. Furthermore, the Python NLTK library* was utilized to perform Arabic light stemming (ARLSTem) and remove Arabic stop words. These cleaning techniques enhanced the quality and readability of the tweets, making them more suitable for further analysis.

3) Relevant terms extraction: To build a list of commonly used terms related to the situation in Yemen, the 313k distinct tweets are analyzed. The frequency analysis technique [25] is used to identify the most frequently used and important terms by counting the occurrence of words in the tweets. As shown in Fig. 2, a list of terms was extracted and used to construct our dataset. These terms cover a variety of topics associated with the situation in Yemen.

4) Manual data labeling: From the 313 k unique tweets, a random sample of 5,000 was chosen for manual labeling. To construct our relevant/irrelevant dataset, the extracted tweets have been manually annotated into six categories: irrelevant (1), 30% relevant (2), 50% relevant (3), 65% relevant (4), 85% relevant (5), and fully Relevant (6) to the situation in Yemen. To assist in the labeling process, a website† was developed and 15 Arabic annotators helped in this process. The annotators had diverse educational qualifications, including advanced educational degrees such as B.C, M.S., or Ph.D. with an age range spanning from 25 to 40 years old. This combination of advanced education and diverse age range allowed the annotators to bring a wealth of knowledge and perspectives to the annotation task. Each annotator was provided with our definition of relevant, relevant ranges, and irrelevant content, along with relevant examples, before commencing the labeling process. The criteria for definition are as follows:

- If one or more relevant terms are discussed, a tweet is typically considered fully relevant.
- If no relevant term is included, a tweet is considered fully irrelevant.
- If a tweet contains a mixture of relevant and irrelevant terms, it is considered partially relevant.

The 5,000 tweets were divided into five groups of 1,000 tweets each. Three different annotators were assigned to work independently on each set of 1,000 tweets. This means that each tweet was annotated by three different annotators, resulting in three values for each tweet based on the six categories, with each category from a different expert. This expedited the process and enabled multiple experts to label the same tweet. After the annotations were completed, a tweet-relevant ratio (r) was calculated for each tweet by adding the three annotation results and dividing the total by the highest summation result (N).

$$r = (\sum_{i=1}^{3} annotatorResult_i)/N \quad (1)$$

Where :

$annotatorResult_i$ is value from 1 to 6

N is the highest summation result = 18

After calculating the tweet-relevant ratio (r) for each tweet, the tweets were further categorized into binary (relevant or irrelevant) and categorical (low, medium, high) datasets. This categorization was likely based on a threshold value determined by the tweet-relevant ratio (r). In order to determine the optimal threshold for tweet categorization, we adopted an experimentation approach. A series of experiments were conducted by systematically varying the threshold values and assessing their impact on the performance of our classification model. A representative dataset of tweets was collected, and their tweet-relevant ratios (r) were computed. Through iterative adjustments of the threshold values and comprehensive analysis of evaluation metrics including accuracy, precision, recall, and F1-score, we successfully identified the threshold that yielded the best performance. This meticulous experimentation and analysis process allowed us to select the threshold value that maximized our chosen evaluation metric, ensuring the accurate differentiation of relevant and irrelevant tweets within our tweet categorization framework. Furthermore, these tweets were appropriately assigned to their respective relevance categories.

For example, if the threshold value was set at 0.34, tweets with a tweet-relevant ratio (r) higher 0.34 would be considered relevant, and those with a ratio of 0.34 or below 0.34 would be considered irrelevant. The binary dataset would then consist of two categories: relevant and irrelevant. Similarly, the

*https://www.nltk.org

†https://tweettag.000webhostapp.com/login.php

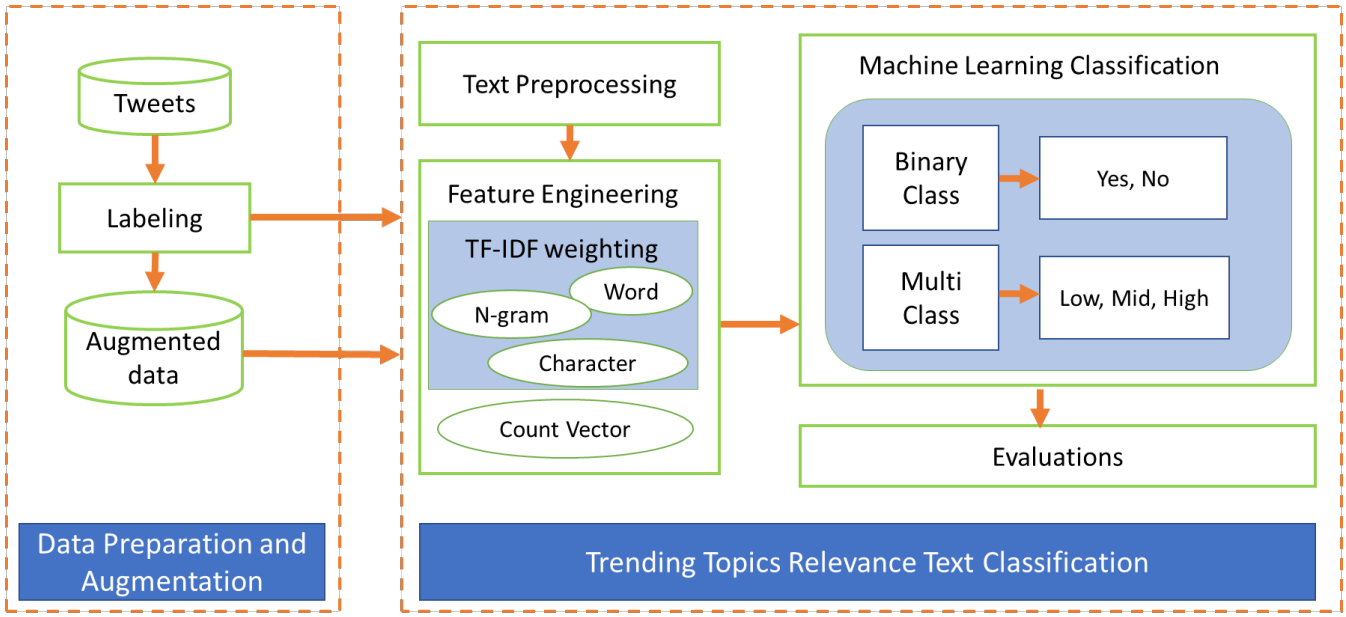

Fig. 1. An integrated framework for Relevance Classification of Trending Topics in Arabic Tweets (RCTAT).

Term count	Term	English Translation	Term count	Term	English Translation
335826	اليمن	Yemen	21992	الامارات	UAE
33083	الحوثي	Houthi	37398	السعودية	Saudi Arabia
18881	التحالف العربي	Arab coalition	8371	مارب	Marib
14981	الشرعية	legitimacy	14405	الجنوب	The south
10090	عدن	Aden	2851	نهم	Nihm
12589	صنعاء	Sana'a	6085	الجوف	Al-Jawf
4530	الرئيس هادي	President Hadi	10054	الجيش	The army
4949	الانتقالي	Transitional	3515	الإرهاب	Terrorism
24516	إيران	Iran	19887	الحرب	The war

Fig. 2. List of terms used to label the dataset.

categorical dataset would be created by dividing the relevant tweets into three categories based on their tweet-relevant ratio (r), such as low (0 to 0.17), medium (0.17 to 0.34), and high (0.34 and above). This categorization was used to evaluate the performance of the framework in identifying relevant tweets and to compare it against human annotations. Tables I and II shows the number of tweets in both manual binary (relevant or irrelevant) and categorical (low, medium, high) datasets.

TABLE I. BINARY MANUAL ANNOTATED DATASET

Binary Class	Ratio	Number of tweets
Irrelevant	r<=34 %	1589
Relevant	r>34 %	3411

TABLE II. CATEGORICAL MANUAL ANNOTATED DATASET

Categorical Class	Ratio	Number of tweets
Low	r<=17 %	791
Medium	r<=34 %	798
High	r>34 %	3411

5) Expanding the annotated dataset: In order to increase the size of the manually labeled dataset, which can be expensive and time-consuming to create manually, we utilized text data augmentation techniques to automatically generate a labeled dataset from the existing one. Data augmentation is employed to improve the classification of relevant tweets in trending topics.

Data augmentation aims to tackle overfitting at the data level, address class imbalance, and enhance the model's generalization [24]. Increasing the diversity of training samples through data augmentation can help the model learn more fundamental features of the data, leading to a higher quality classifier. Tables III and IV show how the size of the binary (irrelevant) and categorical (low, medium) samples changed after the data augmentation technique was employed. In our previous work [26], data augmentation techniques were applied to increase the size of our datasets. This approach aimed to address the class imbalance, avoid overfitting, and enhance the generalization of the model. The authors utilized word embedding techniques [27], specifically the AraVec word vectors trained on Arabic content from Wikipedia and Twitter. They replaced words in the dataset with synonyms based on similarity scores obtained from the word vectors. By applying a random ratio of 50% to 70% for token replacement, they effectively increased the diversity and quantity of training samples. This data augmentation process resulted in a substantial increase in the size of our datasets, which can potentially improve the accuracy and performance of our model by providing a more comprehensive representation of the data.

Table III shows that the size of the irrelevant binary class increased from 1589 to 3236 tweets after data augmentation. Similarly, the low categorical class increased from 791 to 2439 tweets, and the mid categorical class increased from 798 to 2388 tweets as shown in Table IV. This suggests that the data augmentation technique was successful in increasing the size of the datasets, which can help improve the performance of the model by increasing the diversity of the training samples.

TABLE III. BINARY AUGMENTED DATASET

Binary Class	Ratio	Number of tweets
Irrelevant	r<=34 %	3236
Relevant	r>34 %	3411

TABLE IV. CATEGORICAL AUGMENTED DATASET

Categorical Class	Ratio	Number of tweets
Low	r<=17 %	2439
Medium	r<=34 %	2388
High	r>34 %	3411

B. Trending Topics Relevance Text Classification

This section describes the Trending topics relevance text classification process by applying different machine learning techniques. This process includes the following steps: Data pre-processing, feature extraction, and finally, applying various machine learning classifiers.

1) Data pre-processing: In order to ensure data cleansing and eliminate noise that could impact the accuracy of the system, various techniques were employed on the dataset. These techniques include tokenization, normalization, removal of diacritics, removal of repeated characters, removal of punctuations, removal of stop words, removal of non-Arabic alphabets, and light stemming.

2) Content features extraction: Once the data is prepared and augmented, features need to be extracted from the text to represent it into a numerical representation that can be used by the machine learning models. The features listed in Table V were employed for this purpose.

TABLE V. RELEVANCE DETECTION FEATURES

Feature	Description
Word-Level	Each word in the dataset is represented using the TF-IDF matrix.
N-gram-Level	Unigram, bigram, and trigram models are used and represented using the TF-IDF matrix.
Character-Level	TF-IDF character scores for each tweet are represented in the dataset.
Count Vector	The text in the dataset is represented as a vector of term counts.

3) Machine learning classification: Several machine learning classifiers have been utilized to classify relevance text in Trending topics. Three classifiers were trained based on the extracted features to assess their effectiveness in classifying relevant content. The classifiers used were Naive Bayes, Logistic Regression, and LinearSVC, which were selected because they have been widely used as a baseline in previous works on Arabic classification. The experiments' outcomes are discussed in Section IV.

IV. EXPERIMENTS

Experiments were conducted to evaluate the quality of manually labeled and augmented datasets using the features extracted in Section III-B2. Two types of Arabic relevance classification were explored: binary classification (relevant/irrelevant) and categorical classification (low, medium, high). The dataset comprised 5000 tweets in the non-augmented labeled dataset and up to 8000 tweets in the augmented dataset, as shown in Sections III-A4 and III-A5. The Arabic relevance classification was performed using LinearSVC (SVC), Naive Bayes (NB), and Logistic Regression (LR) classifiers with 10-fold cross-validation on both datasets.

The results of binary classification and categorical classification for the Arabic relevance classification are presented in the following subsections:

A. Binary Classification

In this section, we present the results of our experiments for Arabic trending topic relevance text classification using binary classification, as shown in Table VI and Fig. 3.

The results showed that the Logistic Regression classifier with the n-gram TF-IDF feature achieved superior classification performance. Specifically, the classifier attained a macro F1 (M-F1) score of 72% for the non-augmented dataset. On the other hand, applying the same feature with the SVC classifier resulted in the best classification performance for the augmented dataset. The classifier obtained a macro F1-score of 82%.

Likewise, the n-gram TF-IDF feature with the LR classifier and the word TF-IDF feature with the SVC classifier produced the highest precision (P) values of 71% on the non-augmented dataset. On the other hand, the n-gram TF-IDF feature with the SVC classifier achieved the highest precision value of 82% on the augmented dataset.

In terms of recall (R), the SVC classifier with the n-gram TF-IDF and character TF-IDF features yielded the highest recall of 74% on the non-augmented dataset, while the SVC classifier with the n-gram TF-IDF feature achieved the highest recall of 82% on the augmented dataset. Finally, the n-gram TF-IDF feature with the Logistic Regression classifier yielded the highest accuracy (A) of 76% on the non-augmented dataset, while the n-gram TF-IDF feature with the SVC classifier produced the highest accuracy of 82% on the augmented dataset .

B. Categorical Classification

In this section, we present the results of our experiments for Arabic trending topic relevance text classification using categorical classification, as shown in Table VII and Fig. 4.

The results showed that, using the Logistic Regression classifier with n-gram TF-IDF and word count features, as well as the SVC classifier with word TF-IDF, led to significantly better classification performance. These classifiers achieved a macro F1-score of 51% with the non-augmented dataset. On the other hand, the SVC classifier that utilized word, n-gram, and character TF-IDF features obtained the best classification performance on the augmented dataset. This classifier achieved a macro F1-score of 77%.

Likewise, the highest precision value of 51% was achieved with the SVC classifier using the word TF-IDF feature, and the

TABLE VI. BINARY DATASET - EXPERIMENT 1 CONFUSION MATRIX

Classifier		SVC				NB				LR			
Dataset	**Feature**	P	R	A	M-F1	P	R	A	M-F1	P	R	A	M-F1
Non-Augmented	Word Count	0.699	0.699	0.739	0.698	0.625	0.637	0.692	0.629	0.7	0.71	0.74	0.7
	TF-IDF (word-level)	0.709	0.708	0.746	0.707	0.63	0.64	0.694	0.633	0.69	0.7	0.74	0.7
	TF-IDF (n-gram-level)	0.654	0.736	0.753	0.665	0.64	0.64	0.69	0.63	0.71	0.72	0.76	0.72
	TF-IDF (character-level)	0.654	0.739	0.755	0.666	0.63	0.64	0.69	0.63	0.68	0.68	0.72	0.68
Augmented	Word Count	0.784	0.784	0.783	0.783	0.738	0.738	0.737	0.737	0.795	0.795	0.795	0.795
	TF-IDF (word-level)	0.789	0.789	0.788	0.787	0.734	0.734	0.734	0.734	0.788	0.788	0.788	0.787
	TF-IDF (n-gram-level)	0.82	0.82	0.82	0.82	0.74	0.74	0.74	0.74	0.807	0.807	0.807	0.807
	TF-IDF (character-level)	0.813	0.814	0.814	0.813	0.736	0.736	0.735	0.735	0.769	0.769	0.769	0.768

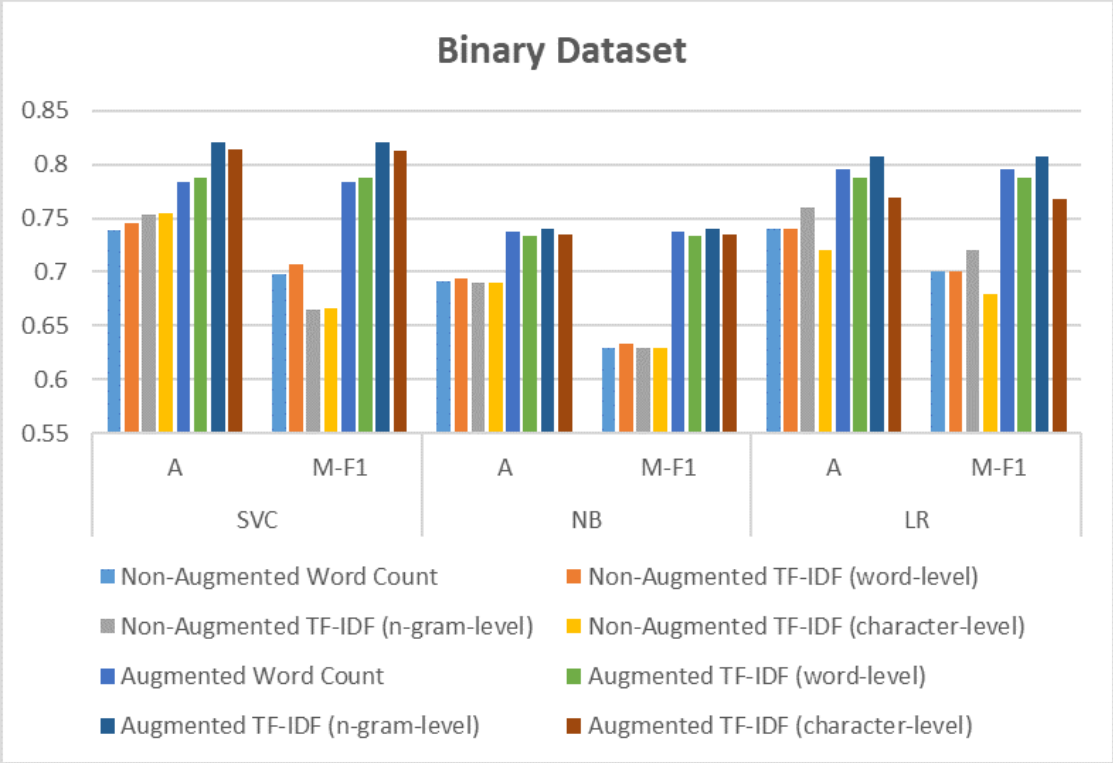

Fig. 3. Plot of binary dataset results.

TABLE VII. CATEGORICAL DATASET - EXPERIMENT 2 CONFUSION MATRIX

Classifier		SVC				NB				LR			
Dataset	**Feature**	P	R	A	M-F1	P	R	A	M-F1	P	R	A	M-F1
Non-Augmented	Word Count	0.5	0.51	0.68	0.5	0.45	0.48	0.67	0.45	0.51	0.51	0.68	0.51
	TF-IDF (word-level)	0.51	0.53	0.69	0.51	0.45	0.48	0.66	0.45	0.48	0.51	0.68	0.49
	TF-IDF (n-gram-level)	0.42	0.59	0.71	0.43	0.44	0.48	0.66	0.45	0.51	0.52	0.69	0.51
	TF-IDF (character-level)	0.42	0.61	0.71	0.43	0.44	0.48	0.67	0.45	0.47	0.49	0.66	0.48
Augmented	Word Count	0.717	0.717	0.726	0.716	0.55	0.56	0.68	0.55	0.59	0.61	0.73	0.59
	TF-IDF (word-level)	0.769	0.779	0.783	0.772	0.55	0.55	0.67	0.55	0.59	0.6	0.73	0.59
	TF-IDF (n-gram-level)	0.767	0.776	0.78	0.769	0.55	0.56	0.67	0.55	0.59	0.6	0.73	0.58
	TF-IDF (character-level)	0.767	0.778	0.78	0.77	0.55	0.55	0.68	0.55	0.68	0.68	0.69	0.68

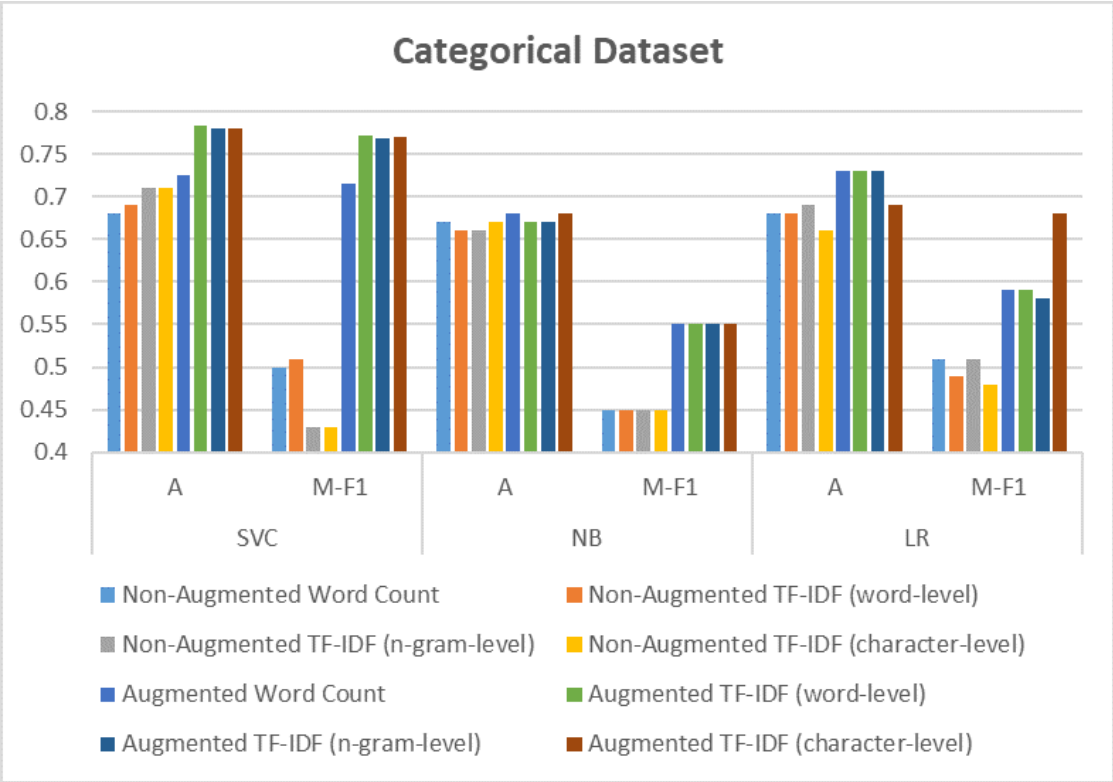

Fig. 4. Plot of categorical dataset results.

N-gram TF-IDF and word count features. On the other hand, the SVC classifier with word, n-gram, and character TF-IDF features obtained the highest precision value of 77% on the augmented dataset.

Likewise, the highest recall value of 61% was achieved using the character TF-IDF feature with the SVC classifier on non-augmented datasets, whereas, on augmented datasets, the highest recall value of 78% was obtained by the same classifier with word, n-gram, and character TF-IDF features. Additionally, the highest accuracy was attained using the SVC classifier with n-gram and character TF-IDF features of 71% on the non-augmented dataset, and the same classifier with word, n-gram, and character TF-IDF features of 78% on the augmented dataset.

V. Discussion

The main goal of this study was to create a benchmark dataset of tweets related to popular topics in Arabic social media. The dataset includes Relevance tweets in Arabic for both binary and categorical classifications. Based on the experimental outcomes, the manually annotated dataset can be utilized as a baseline for future research on Relevance Classification of Trending Topics in Arabic Tweets. As no benchmark dataset exists for classifying Arabic trending topic Relevance tweets, this dataset will prove beneficial to the research community once it becomes publicly accessible.

According to the statistical analysis, the non-augmented dataset had lower values for macro F1, accuracy, recall, and precision in its classification compared to the augmented dataset, which exhibited better results. Based on the findings of the previous section, it can be concluded that the use of data augmentation techniques has improved the classification results, leading to the highest macro F1 score of 82% in binary classification and 77% in categorical classification. The results achieved in the binary classification outperform the work in [6], This work is closest to our work as it also aimed to classify relevant tweets on Indonesian trending topics, and they achieved an F1 measure of 70%.

Machine learning techniques that employ N-gram features provide better results in classifying Relevance tweets within trending topic datasets than other features. Moreover, binary classification achieved superior results compared to categorical classification. Learning in categorical classification is comparatively less accurate than binary classification, as it is a more complex process.

VI. Conclusion and Future Work

This paper presents a novel Arabic dataset consisting of Relevance tweets for binary and categorical classifications, which will be available for public research. The process of tweet collection, manual labeling, and data augmentation for the dataset is described in details. We employed three classifiers, namely Naive Bayes, Logistic Regression, and Support Vector Machine, for Relevance Classification of Trending Topics in Arabic Tweets. The classifiers were trained using four types of features: count vector and TF-IDF (word-level, n-gram-level, and character-level). The study found that the performance varied depending on the classifiers and features used and that higher performance could be achieved with

more annotated data. The Support Vector Machine approach was found to perform well for Relevance classification of Twitter content, with average macro F1 scores of 82% and 77% obtained in the binary and categorical datasets, respectively.

In future work, it would be valuable to explore the effectiveness of advanced deep learning techniques like convolutional neural networks, recurrent neural networks, and BERT for Relevance Classification of Trending Topics in Arabic Tweets. Moreover, extending the dataset to include a wider range of topics and domains and evaluating the generalizability of the proposed classification models across diverse datasets would be of interest.

REFERENCES

[1] K. Morabia, N. L. B. Murthy, A. Malapati, and S. Samant, "Sedtwik: segmentation-based event detection from tweets using wikipedia," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 77–85.

[2] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook, "Mapping the global twitter heartbeat: The geography of twitter," *First Monday*, 2013.

[3] K. Garcia and L. Berton, "Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa," *Applied soft computing*, vol. 101, p. 107057, 2021.

[4] D. E. Cahyani and A. W. Putra, "Relevance classification of trending topic and twitter content using support vector machine," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, 2021, pp. 87–90.

[5] J. Yong, "A cross-topic method for supervised relevance classification," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 2019, pp. 147–152.

[6] S. Nilekar, S. Rawat, R. Verma, and P. Rahate, "Twitter trend analysis," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, 2020.

[7] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *2011 IEEE 11th international conference on data mining workshops*. IEEE, 2011, pp. 251–258.

[8] F. Masood, A. Almogren, A. Abbas, H. A. Khattak, I. U. Din, M. Guizani, and M. Zuair, "Spammer detection and fake user identification on social networks," *IEEE Access*, vol. 7, pp. 68 140–68 152, 2019.

[9] I. Sarker, "Machine learning: algorithms, real-world applications and research directions. sn comput sci 2: 160," 2021.

[10] D. T. N. Huy, T.-H. Le, N. T. Hang, S. Gwoździewicz, N. D. Trung, and P. Van Tuan, "Further researches and discussion on machine learning meanings-and methods of classifying and recognizing users gender on internet," *Advances in Mechanics*, vol. 9, no. 3, pp. 1190–1204, 2021.

[11] C. Wong, "Analyzing easy data augmentation techniques for text classification," Ph.D. dissertation, Harvard College Cambridge, MA, USA, 2021.

[12] A. Zubiaga, D. Spina, V. Fresno, and R. Martínez, "Classifying trending topics: a typology of conversation triggers on twitter," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2461–2464.

[13] S. C. Han, H. Chung, D. H. Kim, S. Lee, and B. H. Kang, "Twitter trending topics meaning disambiguation," in *Knowledge Management and Acquisition for Smart Systems and Services: 13th Pacific Rim Knowledge Acquisition Workshop, PKAW 2014, Gold Cost, Qld, Australia, December 1-2, 2014. Proceedings 13*. Springer, 2014, pp. 126–137.

[14] A. Rafea and N. A. GabAllah, "Topic detection approaches in identifying topics and events from arabic corpora," *Procedia computer science*, vol. 142, pp. 270–277, 2018.

[15] M. Hernandez-Mendoza, A. Aguilera, I. Dongo, J. Cornejo-Lupa, and Y. Cardinale, "Credibility analysis on twitter considering topic detection," *Applied Sciences*, vol. 12, no. 18, p. 9081, 2022.

[16] Z. Mottaghinia, M.-R. Feizi-Derakhshi, L. Farzinvash, and P. Salehpour, "A review of approaches for topic detection in twitter," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 33, no. 5, pp. 747–773, 2021.

[17] E. Verasakulvong, P. Vateekul, A. Piyatumrong, and C. Sangkeettrakarn, "Online emerging topic detection on twitter using random forest with stock indicator features," in *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 2018, pp. 1–6.

[18] C. Zhang, S. Lu, C. Zhang, X. Xiao, Q. Wang, and G. Chen, "A novel hot topic detection framework with integration of image and short text information from twitter," *IEEE Access*, vol. 7, pp. 9225–9231, 2018.

[19] N. Alsaedi and P. Burnap, "Arabic event detection in social media," in *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I 16*. Springer, 2015, pp. 384–401.

[20] N. Alsaedi, P. Burnap, and O. Rana, "Sensing real-world events using arabic twitter posts," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 1, 2016, pp. 515–518.

[21] M. Hammad and S. R. El-Beltagy, "Towards efficient online topic detection through automated bursty feature detection from arabic twitter streams," *Procedia Computer Science*, vol. 117, pp. 248–255, 2017.

[22] A. Rafea and N. A. Gaballah, "Trending topic extraction from twitter for arabic speaking user," in *The 33rd International Conference on Computers and Their Applications (CATA 2018), Las Vegas, Nevada, USA*, 2018, pp. 214–219.

[23] F. Figueiredo and A. Jorge, "Identifying topic relevant hashtags in twitter streams," *Information Sciences*, vol. 505, pp. 65–83, 2019.

[24] J. Gao, "Data augmentation in solving data imbalance problems," 2020.

[25] J. Valaski, S. Reinehr, and A. Malucelli, "Approaches and strategies to extract relevant terms: How are they being applied?" in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer ..., 2015, p. 478.

[26] A. M. Alkadri, A. Elkorany, and C. Ahmed, "Enhancing detection of arabic social spam using data augmentation and machine learning," *Applied Sciences*, vol. 12, no. 22, p. 11388, 2022.

[27] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–39, 2022.