N-gram and Word2Vec Feature Engineering Approaches for Spam Recognition on Some Influential Twitter Topics in Saudi Arabia

Ahmed M. Balfagih, Vlado Keselj, and Stacey Taylor
Faculty of Computer Science, Dalhousie University, Halifax, Canada
Email: {ahmed.balfagih, stacey.taylor}@dal.ca, vlado@cs.dal.ca

***Abstract*—Social media platforms, such as Twitter, have become powerful sources of information on people's perception of major events. Many people use Twitter to express their views on various issues and events and use it to develop their opinion on the diverse economic, political, technical, and social occurrences related to their daily lives. Spam and non-relevant tweets are a major challenge for Twitter trend detection. Saudi Arabia is a top ranked country in Twitter usage worldwide, and in recent years has experienced difficulties due to the use and rise of hashtags based on misleading tweets and spam. The goal of this paper is to apply machine learning techniques to identify spam on the Saudi tweets collected to the end of 2020. To date, spam detection on Twitter data has been mostly done in English, leaving other major languages, such as Arabic, insufficiently covered. Additionally, publicly accessible Arabic Twitter datasets are hard to find. For our research, we use eight Twitter datasets on some significant topics in politics, health, national affairs, economy, and sport, to train and evaluate different machine learning algorithms, with a focus on two feature generation techniques based on N-grams and Word2Vec embeddings. One contribution of this paper is providing these new labelled datasets with embeddings. The experimental results show improvement from using embeddings over N-grams in more balanced datasets vs. more unbalanced ones. We also find a superior performance of the Random Forest algorithm over other algorithms in most experiments.**

***Index Terms*—Twitter, spam detection, machine learning, preprocessing, social media**

I. Introduction

Spam is unsolicited and unwanted digital messages. Most spam messages are advertisements for commercial products. With the rise in use of email, service provides have focused a lot of time and attention on filtering and minimizing junk mail in user inboxes. Detecting and addressing spam, however, is still a major challenge for social media platforms. Twitter is one of the premiere microblogging social media site. It is an excellent tool to express short ideas, give brief feedback, and provides a vehicle to engage in activities through retweets and hashtags. Twitter also provides extensive analytics where companies can get information on the number of followers, retweets, and hashtag trends, for example. Yet, these analytics can be undermined by factors such as noise tweets and messages — from both real and fake accounts.

A. Exceptional Role of Twitter in Saudi Arabia

Saudi Arabia, with a population of 35 million people, is the largest market for social media in the Arab Gulf region [1]. Saudi youth often connect through digital media platforms such as Twitter, Snapchat, and YouTube, to express their opinions, follow the news, and engage in entertainment. Communication sites are an arena for dialogue for Saudis on many current issues, especially the Islamic and liberal topics [2]. However, in the last few years, Twitter has become a destination for all parties in Saudi society, starting with young people in the early days, joined later by community leaders, thinkers, politicians, officials, media professionals, and advocates; The accounts of Saudi preachers on the Twitter platform occupy the list of the most followed accounts in the Arab countries. The popularity and widespread social interest in Twitter prompted Saudi Prince Al-Waleed bin Talal to acquire 4.9% of Twitter's shares, making him one of the major investors [3]. This interest was also clearly reflected in government policies and procedures towards the platform and its celebrities, as Saudi government institutions held many conferences and workshops on the best use of the Twitter platform from the government point of view. The Saudi government has given special attention to Twitter through the verified accounts of senior Saudi officials, and more attention to the most famous Saudi influencers. Celebrity accounts on Twitter have been used in preparing the people for important decisions by the Saudi authorities. Thus, the Saudis have invested in the Twitter platform as a media space to communicate with the Saudi public, through their direct accounts, or through celebrities who support the Saudi government trends.

Saudi Arabia is one of the leading countries in its use of the Twitter platform. It is ranked eighth globally in the number of users (around 12.45 million) and is second in the world based on the population's percentage with approximately 35% of the population using Twitter. Fig. 1 gives more details about leading countries based on the

Manuscript received May 25, 2022; revised July 15, 2022; accepted August 2, 2022.

number of Twitter users as of January 2021 [4]. Saudi Twitter trends have shown a powerful influence on leading popular public opinion, and thus can influence the country's decision-makers.

For example, in 2012, Saudi blogger Hamza Kashgari wrote an article in which he claimed that Twitter activists in Saudi Arabia were insulting to the Prophet of Islam [5]. He was then subjected to a broad hashtag campaign (#HamzaKashgari) asking the government to stop him from writing and put him under investigation. Kashgari originally fled to Malaysia, but was repatriated, investigated, and imprisoned on the charge of contempt of religion. In a more recent example (2020), Twitter activists launched a campaign calling for the dismissal of the directory of the Jeddah Downtown Project after tweets resurfaced from 2013 that were considered hostile to the country [6].

However, sometimes trending hashtags are ignored, such as the hashtag "The salary is not enough" [7], or some other human rights claims, because the Saudi government believes that the claims are not based on reasonable justification. It is also believed that these hashtags stem from abroad to fuel criticism and discontent. There is also evidence that the trend of some hashtags have risen due to tweets from fake accounts which have no actual relationship to the topic at all; The sole purpose of the tweets is to gain attention and popularity for the hashtag through a tactic known as "electronic flies phenomena" [8].

These accounts are often anonymous, and some are managed from outside Saudi Arabia. This became very evident after a decision taken by Arab countries, namely Saudi Arabia, Egypt, the United Arab Emirates, and Bahrain, to cut diplomatic ties with the State of Qatar in June 2017. This decision sparked the creation of a lot of fake accounts which sought to influence public opinion, with both sides engaging in electronic warfare [9]. It has also been observed that many hashtags gained prominence due to spam from unknown accounts carrying commercial advertisements, fraud, the promotion of sexual products, as well as ISIS terrorist propaganda. These tweets generally have nothing to do with the name of hashtags, which affects the credibility of the hashtag and its position in the trend.

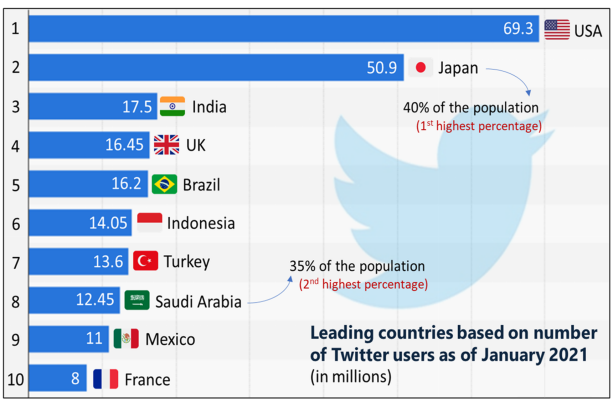

Figure 1. Leading countries based on number of Twitter users.

B. Spam Detection on Arabic Language

Recognizing spam in tweets is an active area of research in the field of Natural Language Processing. There are many advanced studies on spam recognition in tweets in the English language, but significantly less in Arabic. Some of the most important reasons for this are the lack of Arabic content on the Internet compared to English, and the late application of artificial intelligence techniques on Arabic texts. In this study, we apply machine learning experiments on tweets in Arabic, gathered from the Saudi Arabian trends, to detect spam messages that are irrelevant to the hashtag topic, by generating features of the tweet text, then evaluating different approaches and the spam class quality.

The main research questions that we aim to study are the effectiveness of N-gram and embedding-based generating features, different machine learning algorithms and their relationships to different topic domains of tweets, and the balance of datasets in term of Spam vs. Relevant classes.

II. Feature Generation for Arabic Tweets

Position Spam and unrelated tweets can affect the credibility of hashtag trends. Therefore, generating features can help models detect spam tweets. Albadi and Kurdi studied the spreading of religious hatred on Arabic Twitter. They extracted 35 feature categories based on content, tweet, sentiment, and account, from a dataset of 86,346 tweets based on 450 manually labelled accounts. These features were then applied to a bot tweet detection model. Findings indicate that bots were responsible for 11% of hate tweets in their dataset [10]. Almerekhi and Elsayed took the approach of extracting features of Arabic tweets based on attributes such as the formality and structure of the language, as well as temporal aspects, to detect automatically generated tweets [11]. They analyzed 47 features and selected the 16 most significant [12]. With this approach, their model had an accuracy of over 90% using Random Forest.

Feature extraction in spam detection has also been the focus of other research, Boreggah et. al extracted the top 10 features based on tweet content, account behaviour, and account profile. They applied three classification algorithms — Random Forest, Naïve Bayes, and Support Vector Machines. Random Forest returned the highest accuracy of 98.68% [13]. Alorini and Rawat focused on Gulf Dialectical Arabic tweets and extracted only three features and relied on the number of hashtags in tweets, the number of shortened uniform resource locators (URLs), as well as the use of profanity. Naïve Bayes returned the highest accuracy of 86% [14]. AlTwairesh et. al constructed an Arabic Spam Detecting Lexicon (ASDL) containing 108 words [15]. Four features were then identified to be used in the classification model: URL, phone number, number of hashtags, spam lexicon. Both approaches returned comparable results with an average F-measure — 85% (lexicon) and 91.6% (identified features).

Our primary research contribution is to compare different supervised classification methods and evaluate its performance on different hashtags based on two different feature generation approaches: Word2Vec embeddings [16], and the N-gram approach [17]. A further goal is to investigate if grouping the tweets based on hashtag topic (under the same methodology), and combining N-gram and Word2Vec features, has a notable impact on the spam recognition results.

III. Methodology and Data

Our proposed system consists of three major parts: (1) data pre-processing part which includes data collection from Twitter, data cleaning, data modelling process and tweet labelling; (2) feature generation using both N-grams and Word2Vec; and (3) machine learning which includes multiple classification experiments and tests. Fig. 2 shows the system architecture of this study.

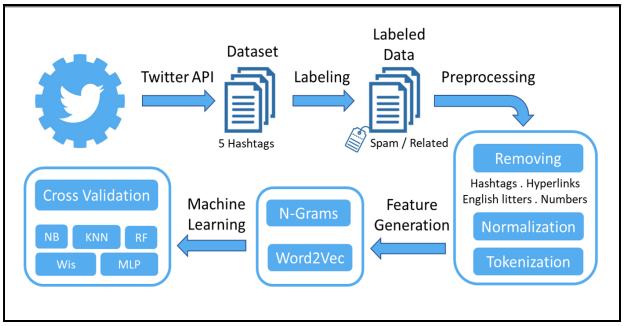

Figure 2. The proposed system architecture.

We created individual datasets for 40,000 Arabic tweets for eight Saudi Arabian trend hashtags. Diverse topics including health (COVID-19, and COVID-10 second wave), politics (boycotting Turkish products, and interview with the Saudi former ambassador to the United States), national affairs (royal orders, Saudi National Day, and news about increasing the Value-Added Tax (VAT)), and sports ("Derby" football match between Al-Hilal and Al-Nasser). Tweets were then manually labelled into two classes: "Relevant" and "Spam". In-line with the common understanding of spam, tweets that are not related to the hashtag being used are labelled as spam. Examples include ads, prayers, and funny jokes. Tweets that relevant to the hashtag topic labelled as relevant. Table I provides details on the datasets of tweets, sorted by spam percentage in descending order.

As our study is focusing on generating features from the text, some tweet features such as the URL, or user were removed. Data cleaning was implemented in Python [18]. All punctuation was removed. Also, the number of words in the tweet were counted and tokenized using a regular expression. For the feature generation phase, we used N-Grams and Word2Vec embedding approaches. N-Grams are continuous sequences of words or symbols or tokens in a document. In technical terms, they can be defined as the neighbouring sequences of items in a document. They come into play when we deal with text data in NLP (Natural Language Processing) tasks [19]. Word embedding, on the other hand, is a term used to describe how words are represented for text analysis in natural language processing. Typically, this representation takes the form of a real-valued vector that encodes the meaning of the word with the expectation that words that are close to one another in the vector space will have similar meanings. [20]. Word2vec is a word embedding technique uses a neural network model to learn word associations from a large corpus of text [16]. Once trained, a model like this may identify terms that are similar or propose new words to complete a phrase. As the name implies, word2vec represents each distinct word with a particular list of numbers called a vector. The vectors are chosen carefully such that a simple mathematical function (cosine similarity between the vectors) indicates the level of semantic similarity between the words represented by those vectors.

The N-gram feature generation was generated by Weka [21], while the Word2Vec was implemented in Python [18]. For the N-grams, applying StringToWordVector filter in Weka, using NGramTokenizer, we generated unigrams, bigrams, and trigrams. Using wordToKeep parameter we set to keep a number ranging between 250 to 300 words to keep, to generate our N-grams, in condition if these words are frequented at least five times using the parameter minTermFreq to limit the generated features to the most 500 frequent N-grams. From the same generation filter, we applied the TF-IDF method [22] using IDFTransform parameter and TFTransform parameter, which evaluates how relevant a word is to a document in a collection of a document. The TFTransform sets whether if the word frequencies should be transformed into log(1+fij), where fij is the frequency of word i in document (instance) j. while IDFTransform sets whether if the word frequencies in a document should be transformed into:

$$fij*\log(\text{num of Docs}/\text{num of Docs with word i})$$

where fij is frequency of word i in document (instance) j.

Using Gensim free open-source Python library [23], we applied Word2Vec algorithm to generate embedding features of the tweet, we trained the whole corpus of tweets datasets using context window size of 5, and 1 for skip-gram model to generate 100 embeddings features foreach tweet. For this model we used the negative sampling method to specify how many "noise words" should be drawn by sitting its value by 10. We published the labeled datasets' embeddings on GitHub [24].

Using 10-fold cross-validation, five machine learning algorithms were applied using Weka [21]: Naïve Bayes (NB) [25], K-Nearest Neighbour (KNN) [26], Random Forest (RF) [27], Multi-Layer Perceptron (MLP) neural network method using Dl4jMlpClassifier function in Weka [28] and WiSARD (Wis) [29]. The MLP network was designed with two dense layers over 32 epochs with a batch size of 100. We also applied this to the WiSARD model as well.

TABLE I. THE TOTAL OF TWEETS IS MORE THAN 40,026 AFTER REMOVING IRRELEVANT TWEETS

Dataset	Description	Topic	#Tweets	Relevant	Spam	Spam%
SA_SND	The Saudi National Day #اليوم_الوطني_السعودي	National Affairs	5,760	2,911	2,849	49%
SA_VAT	Value Added Tax #القيمة_المضافة	National Affairs	1,652	994	658	40%
SA_Cov	Covid-19 #كورونا	Health	7,617	4,861	2,756	36%
SA_Roy	Royal orders from the king #أوامر_ملكية	National Affairs	4,696	3,013	1,683	35%
SA_Tur	Baycotting Turkish products #مقاطعة_المنتجات_التركية	Politics	9,603	7,101	2,502	26%
SA_2Wv	Second Wave of Covid-19 #الموجة_الثانية	Health	2,293	1,760	533	23%
SA_Ban	Prince Bandar Bin Sultan interview #بندر_بن_سلطان	Politics	6,040	5,152	888	14%
SA_Der	Soccer match between Al-Nasser and Al-Hilal #ديربي_الرياض	Sport	2,365	2,206	169	7%
		TOTAL	40,026	27,998	12,038	

For the Multi-Layer Perceptron network, we used an activation layer using ActivationReLU function, then dense layer using ActivationReLU function with 10 outputs, then another dense layer using ActivationReLU function with three outputs, and finally an output layer using Softmax function with two outputs, which is the classification stage. Fig. 3 shows the structure of the Multi-Layer Perceptron Nural Network.

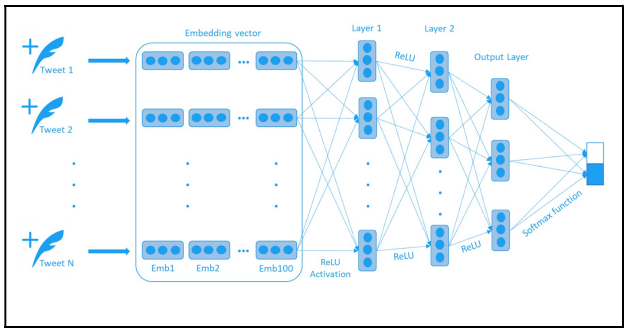

Figure 3. Multi-Layer perceptron network structure.

IV. MACHINE LEARNING EXPERIMENTAL RESULTS AND DISCUSSION

The results of the applied machine learning algorithms show high accuracy for spam recognition for Arabic tweets in general. Random Forest performed the best overall, especially with the N-gram generated features. KNN provided the best accuracy on the SA_Roy dataset which uses features generated by the Word2Vec embedding. The MLP returned the best result on the SA_Tur dataset using Word2Vec. Using N-grams, MLP also delivered the best accuracy on the SA_SND and SA_Der datasets. We did see an improvement of embedding over N-grams using NB in three datasets: SA_SND, SA_Roy, and SA_2Wv. See Table II and Table III for the accuracy details using N-grams and Word2Vev. The highest accuracy results are highlighted in green.

The main outcomes from these experimental results show that both N-grams and Word2Vec outperform the other under certain conditions. When the datasets are unbalanced between Relevant and Spam (where spam accounted for 14% to 26% of the dataset) N-grams returned the stronger result. However, when the datasets were much more balanced (where spam accounted for between 33% and 50%), Word2Vec performed better. Another important finding is that Word2Vec embeddings with 100 generated features has greater accuracy than N-grams with 500 generated features per tweet. See Table IV for the relative difference between N-gram generated features and Word2Vec for all algorithms used in our work. It is noticed that for the N-gram experiments Random Forrest algorithm had the highest accuracy in all datasets but the SA_SND (The most balanced) and SA_Der (The least balanced). See Table II. On the other hand, it is noticed that for the Word2Vec experiments Random Forrest algorithm had the highest accuracy in all datasets except in SA_Roy and SA_Tur, the datasets that are listed in the middle of spam percentage rates. See Table III.

We also applied the same experiments on the grouped datasets by topic. See Table V for the details of the three grouped datasets by topics: national affairs hashtags, COVID-19 related hashtags, and political hashtags. We found that grouped hashtags by topic that have more balanced classes (between Relevant and Spam) have better results using Word2Vec embeddings: the SA_National and SA_Health datasets which have 42% and 33% spam tweets have better accuracy results than N-grams. However, the SA_Politics dataset (21% spam tweets) has better results for most of the algorithms using the N-gram generated features. See Table VI. The time consuming for the experiments on the grouped datasets by topics was longer, especially in the neural networks' algorithms and Random Forrest due to high number of instances.

TABLE II. 10-FOLDS CROSS VALIDATION ACCURACY (NG)

Dataset	Spam %	NB	KNN	RF	MLP	Wis
SA_SND	49%	71.32%	94.79%	95.49%	95.52%	80.89%
SA_VAT	40%	93.28%	93.56%	94.40%	93.28%	92.44%
SA_Cov	36%	94.39%	97.28%	98.04%	96.41%	97.39%
SA_Roy	35%	83.50%	89.56%	90.68%	89.84%	90.21%
SA_Tur	26%	94.88%	96.94%	98.54%	97.74%	96.71%
SA_2Wv	23%	89.18%	96.16%	99.74%	96.55%	99.65%
SA_Ban	14%	91.56%	99.00%	99.00%	98.89%	98.94%
SA_Der	7%	87.02%	94.74%	95.42%	96.42%	96.20%

TABLE III. 10-FOLDS CROSS VALIDATION ACCURACY (W2V)

Dataset	Spam %	NB	KNN	RF	MLP	Wis
SA_SND	49%	86.38%	96.94%	96.99%	95.41%	94.93%
SA_VAT	40%	90.61%	95.39%	96.36%	95.94%	96.06%
SA_Cov	36%	92.84%	98.86%	98.90%	98.62%	98.11%
SA_Roy	35%	85.62%	96.98%	96.55%	95.34%	95.66%
SA_Tur	26%	87.49%	96.99%	97.34%	97.62%	96.14%
SA_2Wv	23%	90.27%	95.38%	96.29%	95.86%	94.46%
SA_Ban	14%	93.82%	96.80%	97.20%	96.93%	95.22%
SA_Der	7%	96.67%	97.98%	98.23%	97.55%	98.18%

TABLE IV. WORD2VEC ACCURACY RELATIVE IMPROVEMENT OVER N-GRAM

Dataset	Spam %	NB	KNN	RF	MLP	Wis
SA_SND	49%	21.12%	2.27%	1.57%	-0.12%	17.36%
SA_VAT	40%	-2.86%	1.96%	2.08%	2.85%	3.92%
SA_Cov	36%	-1.64%	1.62%	0.88%	2.29%	0.74%
SA_Roy	35%	2.54%	8.28%	6.47%	6.12%	6.04%
SA_Tur	26%	-7.79%	0.05%	-1.22%	-0.12%	-0.59%
SA_2Wv	23%	1.22%	-0.81%	-3.46%	-0.71%	-5.21%
SA_Ban	14%	21.12%	2.27%	1.57%	-0.12%	17.36%
SA_Der	7%	11.09%	3.54%	2.60%	1.17%	2.06%

TABLE V. THE GROUPED DATASETS BY THE HASHTAG'S TOPIC

Dataset	Description	#Tweets	Relevant	Spam	Spam %
SA_National	National Affairs hashtags	12,108	6,918	5,190	42%
SA_Health	Covid-19 related hashtags	9,910	6,621	3,289	33%
SA_Politics	Political Hashtags	15,643	12,253	3,390	21%
	TOTAL	37,661	25,792	11,869	

TABLE VI. COMPARISON BETWEEN W2V AND N-GRAM 10-FOLD CROSS VALIDATION ACCURACY RESULTS ON GROUPED DATASETS BY TOPIC

	Dataset	Spam %	NB	KNN	RF	MLP	Wis
NG	SA_National	42%	73.28%	94.69%	95.58%	92.39%	89.34%
	SA_Health	33%	88.40%	97.32%	97.52%	97.07%	96.24%
	SA_Politics	21%	90.56%	97.94%	98.33%	95.42%	95.77%
W2V	Dataset	Spam %	NB	KNN	RF	MLP	Wis
	SA_National	42%	84.18%	96.66%	96.37%	95.42%	94.46%
	SA_Health	33%	89.59%	97.95%	98.06%	97.22%	96.39%
	SA_Politics	21%	74.62%	96.86%	97.13%	93.41%	95.28%

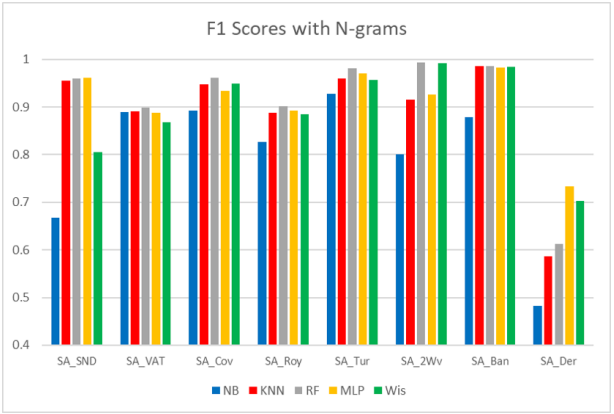

Figure 4. F-Measures of the spam class for N-gram approach.

The F-measure results of the spam class also return higher scores using Word2Vec embeddings on most machine learning experiments, especially for the more balanced datasets, and the least balanced dataset SA_Der. Fig. 4 and Fig. 5 show the F-measures for spam class for all datasets in N-grams and Word2Vec.

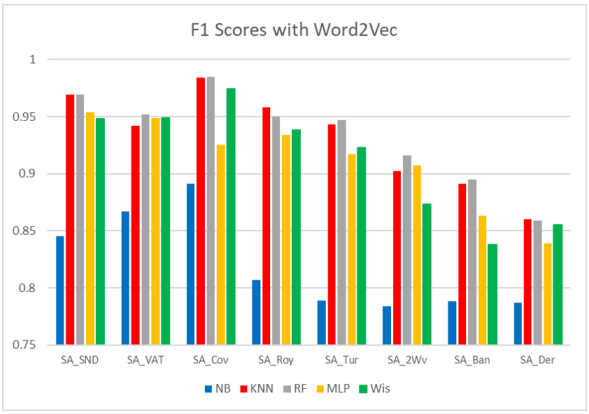

Figure 5. F-Measures of the spam class for Word2Vec approach.

Comparatively, Random Forest returns higher scores F-measure, supporting that Random Forest is the best algorithm for our research. As well, when compared to the neural network algorithms, Random Forest also was the better approach as it took less time, particularly when compared to Multilayer Perceptron, which took the most time. Overall, Naïve Bayes returned the worst results under both Word2Vec and N-grams. It is also noticed that the rate of F-measures decreasing gradually from SA_Cov (36% of spam) to SA_Der (7% of spam) in most algorithms. See Fig. 5.

Another experiment was combining the generated features of both approaches of N-gram and Word2Vec in one data model. Therefore, the total number of generated features will be 600 feature (500 N-grams and 100 Word2Vec embeddings). This is to observe if using both approaches together have a positive impact on classification accuracy. See results in Table VII Using same machine learning methods with 10-folds cross validation.

TABLE VII. 10-FOLDS CROSS VALIDATION ACCURACY (NG+W2V)

Dataset	Spam %	NB	KNN	RF	MLP	Wis
SA_SND	49%	66.08%	96.28%	97.36%	96.67%	90.55%
SA_VAT	40%	90.85%	94.85%	96.49%	95.33%	95.94%
SA_Cov	36%	87.03%	98.27%	98.82%	98.16%	98.63%
SA_Roy	35%	82.60%	96.87%	97.15%	96.65%	96.21%
SA_Tur	26%	87.36%	98.60%	98.96%	98.75%	98.64%
SA_2Wv	23%	88.05%	96.99%	96.90%	97.25%	96.99%
SA_Ban	14%	87.95%	96.37%	97.25%	96.88%	96.89%
SA_Der	7%	90.27%	97.77%	98.44%	98.27%	98.48%

We noticed that Random Forrest perform better than other algorithms in most of these experiments, with overall enhancement in the accuracy in comparison with N-grams datasets and Word2Vec datasets. Five out of eight datasets show a superiority of Random Forest in this experiment in comparison with previous ones. For the other algorithms, combining N-Gram and Word2Vec features is not as improving the f-measure result of spam

class as it did for the Random Forest. For instance, in KNN, five out of eight datasets show better f-measure results in Word2Vec features experiments, and only two datasets that show better f-measure results in combining Word2Vec and N-Gram features. Another example, Naïve Bayes algorithm, that is already having the lowest accuracy results in all experiment comparing to the rest of algorithms, it never showed any improvement in the accuracy results in all datasets.

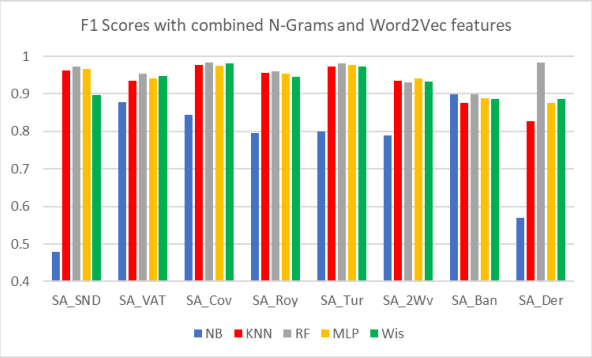

Figure 6. F-Measures of the spam class for combined features of N-gram and Word2Vec approaches.

The F-measure results of the spam class also return higher scores using the combination of Word2Vec embeddings and N-gram features on most of the applied machine learning experiments. Fig. 6 shows the F-measures of the spam class for combining the features of N-gram and Word2Vec Approaches together. It is showing higher quality of the Spam class using Random Forest and it is noticed the higher result in seven datasets out of eight.

Therefore, we can conclude that generating features using Word2Vec embedding is a better approach than using N-grams for most balanced datasets in detecting spam tweets. Moreover, combining Word2Vec with N-Gram features could improve the accuracy results in most algorithms, but we did not find a strong correlation between having the best accuracy and the number of generated features by combining those features, unless in Random Forest algorithm. We can also conclude that Random Forest is the strongest performing algorithm in terms of accuracy, F-measure, and processing time, for the majority of the experiments that used in this study, and Multi-Layer Perceptron, regardless its high performance and results, was the most time-consuming algorithm overall.

V. Conclusion and Future Work

Twitter has become an important and vital source for public opinion in the Arabic speaking world — especially in Saudi Arabia. This has also given rise to Arabic spam and unrelated/unwanted tweets and digital messages, which can have a significant effect on the news and trends on Twitter in Saudi Arabia. To address this challenge, tweet feature generation is an active field of research. In this study, we applied two different feature generation techniques on Arabic tweets — N-grams and Word2Vec — to determine which method returns better results. We determined that using Word2Vec embeddings provides an advantage over using N-grams to generate features on more balanced datasets. Yet, when datasets are unbalanced, using N-grams to generate features provides an advantage over Word2Vec. We also came to the same conclusion in grouping hashtags by topic type first before generating features. Another important finding is that Random Forest outperformed the other algorithms we used in the majority of our experiments — Naïve Bayes, K-Nearest Neighbors, Multi-Layer Perceptron and WiSARD — in accuracy results, spam class evaluation scores, and in processing time. As part of our future work, we plan to apply the same machine learning algorithms on more generated embeddings, to determine if there is a correlation between accuracy and the number of generated features. As well, we plan to apply a resampling method on both approaches to determine if there is improvement in the performance of the model. Moreover, we want to develop Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to see if there are any interesting findings on the results. Finally, we believe it would be extremely valuable to conduct the same study on English datasets to compare the results against those we achieved in Arabic, as will as comparing different Arabic dialects and have tweets from different Arabian countries.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Ahmed Balfagih conducted the research, collected & preprocessed the data, designed the diagrams, and wrote the paper. Vlado Keselj supervised the work and participated in writing. Stacey Taylor participated in writing, arranged tables, and reviewed the language. All authors had approved the final version.

References

[1] V. Godinho, "Two-thirds of Saudi Arabia's population is under the age of 35," *Gulf Business*, vol. 10, 2020
[2] S. Lacroix, "Between Islamists and liberals: Saudi Arabia's new "islamo-liberal" reformists," *Middle East Journal*, vol. 58, no. 3, pp. 345-365, 2004.
[3] Bloomberg, "Why apple, Citigroup and Twitter will be tracking this arrest in Saudi Arabia," *Features*, 2017.
[4] S. Kemp, "Digital 2020: Global digital overview — datareportal – global digital insights," 2020.
[5] D. Keyes, "Saudi writer Hamza Kashgari faces charge of blasphemy after tweets about Muhammad," *The Washington Post*, 2012.
[6] BBC Arabic, "Hisham Malaikah: Who is he? Why did he have to delete his account from Twitter?" Translated from Arabic Source, 2020.
[7] N. Altuwaijri. "Salaries not enough: Saudis shout on social media," A*larabiya News*, 2013.
[8] B. Noureddine, *Electronic Flies and Public Opinion*, University of Mascara, 2017.
[9] P. Wintour, "Qatar given 10 days to meet 13 sweeping demands by Saudi Arabia," *The Guardian*, vol. 23, 2017.
[10] N. Albadi, "Hateful people or hateful bots?" *Proceedings of the ACM on Human-Computer Interaction*, 2019.
[11] H. Almerekhi and T. Elsayed, *Detecting Automatically Generated Arabic Tweets*, Springer, 2015.

[12] A. Alharbi and A. Aljaedi, "Predicting rogue content and Arabic spammers on Twitter," *Future Internet*, 2019.
[13] B. Boreggah, A. Alrazooq, M. Al-Razgan, *et al.*, "Analysis of Arabic bot behaviors," in *Proc. 21st Saudi Computer Society National Computer Conference*, 2018.
[14] D. Alorini and D. Rawat, "Automatic spam detection on gulf dialectical Arabic tweets," in *Proc. International Conference on Computing, Networking and Communications*, 2019.
[15] N. A. Twairesh, M. A. Tuwaijri, A. A. Moammar, *et al.*, "Arabic spam detection in Twitter," in *Proc. the 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*, 2016.
[16] T. Mikolov, K. Chen, G. Corrado, *et al.*, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[17] S. Ching, "N-gram statistics for natural language understanding and text processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 164-172, 1979.
[18] G. V. Rossum and F. L. Drake Jr., *Python Reference Manual*, Centrum voor Wiskunde en Informatica Amsterdam, 1995.
[19] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing," in *Computational Linguistics, and Speech Recognition*, Pearson, 2021.
[20] J. Brownlee, "What are word embeddings for text?" *Machine Learning Mastery*, 2019.
[21] Weka, Weka 3: Machine learning software in Java, 2021
[22] W. Uther and G. Webb, "TF-IDF," *Encyclopedia of Machine Learning*, pp. 986-987, 2011.
[23] R. Rehurek and P. Sojka, "Gensim-Python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.
[24] A. Balfagih. Arabic tweets 100-embeddings datasets. GitHub. DNLB-LAB 2022. [Online]. Available: https://github.com/dnlp-lab/balfagih
[25] S. Russell and P. Norvig, "A modern, agent-oriented approach to introductory artificial intelligence," *ACM SIGART Bulletin*, vol. 6, pp. 24-26, 1995.
[26] N Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, p. 172, 1992.
[27] L. Breiman, "Random forests," *Machine Learning*, pp. 5-32, 2001.
[28] I. Witten, "More data mining with Weka - Simple neural networks," New Zealand: Department of Computer Science University of Waikato, 2019.
[29] WiSARD, "Github - giordamaug/wisard4weka: A supervised classification method for Weka based on weightless neural networks, 2018.

Ahmed M. Balfagih is a data science researcher born in Jeddah, Saudi Arabia, in 1984. He had master's degree in computer science, from Dalhousie University, Halifax, Canada, in 2016.
Currently, he is a Ph.D. candidate in computer science from Dalhousie university, Halifax, Canada. Besides that, he is working as a Teacher Assistant in the faculty of computer science at Dalhousie University, Halifax, Canada. Mr. Balfagih research interests are in natural language processing, machine learning, and social media sentiment analysis.

Vlado Keselj is a Serbian-Canadian computer scientist known for his research in natural language processing and authorship attribution. He is a professor at Dalhousie University, Halifax, Canada. He earned his Ph.D. in 2002 at the University of Waterloo, Waterloo, Canada.
Dr. Vlado is a recipient of the 2019 CAIAC Distinguished Service Award, awarded by the Canadian Artificial Intelligence Association (CAIAC). He is interested in many areas such as Natural Language Processing, Data and Text Mining, Behavioral Analytics.

Stacey Taylor in a Canadian instructor in Faculty of Computer Science in Dalhousie University, Halifax, Canada. She had Master of Business Administration in accounting and finance from Saint Mary's University, Halifax, Canada, in 2010. Currently, she is a Ph.D. candidate in computer science from Dalhousie university, Halifax, Canada. Ms. Taylor is highly experienced finance professional with IT experience with a proven leadership track record.