International Journal on Data Science and Technology
2020; 6(1): 23-36
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20200601.14
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Detection Mechanism for Malicious Messages on KSU Student Social Network

Rawan Almutlaq*, Alaaeldin Hafez

Computer and Information Sciences College, King Saud University, Riyadh, Saudi Arabia

Email address:
437203381@student.ksu.edu.sa (R. Almutlaq), ahafez@ksu.edu.sa (A. Hafez)

*Corresponding author

To cite this article:
Rawan Almutlaq, Alaaeldin Hafez. Detection Mechanism for Malicious Messages on KSU Student Social Network. *International Journal on Data Science and Technology.* Vol. 6, No. 1, 2020, pp. 23-36. doi: 10.11648/j.ijdst.20200601.14

Received: December 8, 2019; **Accepted**: December 26, 2019; **Published**: January 8, 2020

Abstract: The internet has a considerable effect on social relations and connections among people. Social networking platforms have been an enormous medium for establishing relations and connections among different people all over the world. People, organizations and companies use these platforms to communicate and interact with their communities and audience. These platforms have made it easy for people to share information, create content, and communicate and connect with others online; however, online interaction and communication among people have resulted in the creation of many problems. Malicious contents can easily be shared and populated to reach a wider audience than by using the traditional sharing methods. Detection mechanism is a growing area of research that can detect any inappropriateness of data that is more sensitive to malicious behavior. The detection mechanism needs to be involved in the analysis of the abusing messages posted on the Twitter account of King Saud University (KSU). Text mining is one approach that can be used to detect such malicious or abusing messages. Text mining techniques provide the means to perform data classification where messages can be classified into malicious and non-malicious messages. In addition, Sentiment Analysis is used to identify user tendencies, trends, and opinions by classifying a text into positive, negative and neutral. In this paper, we aim to provide a literature review to investigate the current techniques. The study also addresses the detection of malicious messages which identifies the behavior of malicious and abusive messages. Based on the extensive review of the current techniques, our focus is on the analysis of Arabic and English tweets on KSU's Twitter account. First, data was collected from Twitter. This was followed by the preprocessing phase. Then, a corpus was produced applying a machine learning based approach by using Naive Bayes and Random Forest Classifier algorithms. Subsequently, the study focused on comparing the accuracy and performance of the Naive Bayes classifier with Random Forest Classifier algorithms in analyzing Arabic and English texts. In order to ensure reaching accurate results, Arabic and English tweets were analyzed.

Keywords: Social Network, Text Mining, Text Classification, Stemming, Tokenization, Sentiment Analysis

1. Introduction

The internet considerably affects social relations and connections among people. Social media platforms have been an enormous medium for establishing relations and connections among different people all over the world. People, organizations and companies use these platforms to communicate and interact with their communities and audience. People use online platforms to speak about their experiences and express their feelings and thoughts. They may speak in detail about their daily life and their activities [1]. For instance, last year, the number of active users on Twitter per month reached 328 million people who shared 500 million tweets per day [2]. Around 29% of social media users use Twitter [3], and around 83% of the leaders of the world have accounts on this social media platform [4]. According to a report issued by Pew Internet & American Life Project in 2016, the percentage of online adult Twitter users reached approximately the quarter of the total number of users [5]. This percentage of adult users did not significantly change compared to 2015. According to that report [6], Twitter was heavily used by educated people with 29% of its users are holders of higher college degrees [7]. The usage of social media platforms has increased largely

among adults. The percentage of adult users was only 5% in 2005 and it increased to reach 69% in 2016 [5].

Social media websites facilitate communication and engagement; make it easy for people to share and negotiate thoughts; and enable them to establish relations and friendships.

The use of social media has increased largely over the last 10 years. Every minute, Twitter users post 350,000 tweets [9] and Facebook users [8] write 510,000 comments on Facebook. People using these platforms are from different cultures, countries, and education levels and they speak different languages and engage with each other. As a result, much offensive online contents have been posted on these websites, causing the annoyance of many users [10].

This issue is very important and there is a need to develop a tool to detect and remove malicious contents from social media websites.

1.1. Problem Definition

Online information sharing, content creation, communication, and connection establishment among people has become easy. However, online interaction and communication have resulted in the creation of many problems. Malicious content can easily be shared and populated to reach a wider audience than by using the traditional sharing methods [7].

Malicious messages, such as abusive and offensive messages, can result in creating an unhealthy environment that encourages aggression, hatred, and violence among individuals and groups. Malicious online messages can be created anonymously from an anonymous place. This can negatively affect a person's life and his/her communication experience on the social media website [7, 11]. These offensive messages may hurt the feelings of other people and discourage them from continuing to communicate on that online platform. It may also prevent new users from using this particular website so as to avoid being subject to the same bad experience. In addition, the victims of abusive and offensive content may suffer from physical and psychological problems resulting from being subjected to these offences [12].

The terms of service (TOS) regulated by social media websites, such as Twitter, Facebook, etc., prohibit their users from sharing such abusive content. Nevertheless, these social websites filter the content posted by users only partially, which leads to the successful removal of only a small portion of the huge number of offensive tweets and posts while the majority of them remain visible. Some of these social media websites, such as YouTube, give users the opportunity to block offensive and abusive content from appearing to them, but these platforms fail to appropriately detect the abusive content. In some cases, the abusive content may be detected based on an inappropriate reason. For example, a comment can be banned and remain invisible because it received no likes from users. Besides, it is all-consuming to hire people to be responsible for checking every single comment and all contents posted on social media websites [7, 13]. In a 2017 survey report issued by Pew Internet & American Life Project [11], 41% of citizens of the United States have been targeted by abusive content and 66% of them have seen others falling as victims to offensive content. The survey also reported that 27% of the users are not posting anything online.

In a 2016 report [12] that was made in the United States, 70% of people have encountered abusive content, and 25% of them stopped using social media after that experience.

BuzzFeed has made a survey in which 2007 Twitter users participated [13] and this survey showed that the website had much abusive content. Twitter tried to solve the problem but its endeavors were insufficient and unsuccessful. According to the report, 46.4% of Twitter users reported abusive content and the website did not take any action in response. Almost 2.6% out of the users said that Twitter removed the reported abusive content. 18.2% received a reply from Twitter indicating it was not found to be offensive content. 28% received no reply whatsoever from Twitter [13], while most of the users reported that Twitter took a lot of time to respond to their reports. Online social media have become a big threat, especially for young people. A study [11] showed that 80% of the blogs contain offensive and abusive messages. The offensive content has been almost everywhere online.

1.2. Research Questions

The present study seeks to answer the following main questions:

a) What are the features of an abusive message which help identify a message as malicious?
b) How can we apply text mining to classify the abusive message to promote the sentiment analysis?
c) How can we detect the malicious tweet by using the different classification algorithms?
d) What is the benefit of applying the classification algorithms, such as Naïve Base, to a malicious tweet?

1.3. Objectives

Detection mechanism is a growing area of research that can detect any inappropriateness of data that is more sensitive to malicious behavior. The detection mechanism of malicious messages needs to be involved to analyze abusing messages posted on the Twitter account of King Saud University (KSU). Text mining is one approach that can be used to detect such malicious or abusing messages. Text mining techniques provide the means to perform data classification where messages can be classified into malicious and non-malicious messages.

In this paper, we aim to address the detection of malicious messages that identifies the behavior of offensive and abusive messages. We focus on analyzing an Arabic and English text on the Twitter account of King Saud University (KSU). First, data was collected from Twitter. Then came the pre-processing phase in which a corpus was produced employing a machine learning based approach using Naive Bayes and Random Forest classifiers algorithms. Subsequently, the study involved making a comparison

between the accuracy and performance of the Naive Bayes classifier and Random Forest classifier algorithms in Arabic and English texts. In order to ensure reaching accurate results, Arabic and English tweets were analyzed.

2. Research Methodology

In this paper, we used sentiment analysis framework to build a model to identify malicious (abusive) messages in a social media network, namely the Twitter account of KSU. The procedures taken to carry out the research involved more than one phase, namely collection data from Twitter, pre-processing, classification, and finally testing and evaluation. In the classification phase, we classified the tweets by identifying whether they are malicious (abusive) or non-malicious (non-abusive) through using machine learning platforms, such as Vicinitas and Python program language, and using classifier algorithms, such as Naïve Bayes classifier [19], in the classification of the tweets through the use of NLTK [18], TextBlob [17], and NB. Finally, the evaluation phase involved comparing the results reached in the classification phase.

3. Proposed Framework Design

With the increasing amount of data on social networking websites, necessity arose to analyze data and detect data trends. There has been some flexibility when it comes to the collection and analysis of data on social media websites, particularly on Twitter. Twitter provides access to a massive amount of data which attracts researchers to analyze the data available on it. Researchers can find data by using the application programming interface (API).

In this paper, we focus on one type of sentiment analysis, namely the analysis of malicious (i.e. abusive) messages posted on KSU's account on Twitter. We also found honest opinions of visitors. For example, a student wrote an abusive tweet that criticizes the system of services of KSU. We managed to identify the sentiment of the text (i.e., whether it is negative, positive, or neutral).

This part discusses the detection of malicious message approach, as well as the techniques and tools that will be used to the anticipated proposed framework for automatic malicious (abusive) messages classification in social networking websites in general and the Twitter account of King Saud University in particular. There are five reasons why this research has focused on Twitter:

a) The Twitter API is accessible compared with other social network platforms which enables developers to easily create tools to access Twitter data. The software is also available and accessible as an online tool for researchers.
b) Data on Facebook are difficult to find compared with Twitter and are only available in public for marketing objectives.
c) Twitter is a popular platform and can attract more researchers because it contains a massive amount of data.

The framework includes five phases. The first phase is the process of data collection, then comes the phase of tweets cleaning and annotation and data pre-processing. The third phase is the feature extraction process followed by the classification process. The last phase is the testing and evaluation process. Figure 1 shows the proposed framework components. Figure 2 shows the proposed framework implementation. In the proposed framework, we will be using Python language and some packages using it [14-19].

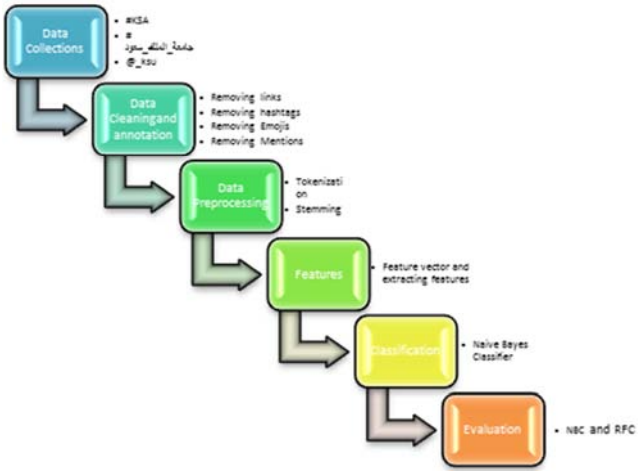

Figure 1. *Proposed Framework Components.*

In Figure 1, (#جامعة_الملك_سعود) it mean hashtag for King Saud University.

Figure 2. *Proposed Framework Implementation.*

3.1. Data Collection

In order to make an automatic malicious messages classification, the first step is to obtain the data by collecting the tweets using a free online API called "Vicinitas" [14]. Vicinitas is an open source software platform for present in-depth analytics and data mining practice. It is also a Twitter analytic tool for tracking hashtags, keywords, accounts, and websites. Vicinitas helps track and analyze real-time and historical tweets of social network websites, especially Twitter.

We used three keywords to collect the tweets, (#KSU, @_ksu, #جامعة_الملك_سعود). The following are the analytical results of the collected data for each keyword (@_ksu). We used this keyword because it is the username of the King Saud University social media account on Twitter, so every tweet that has this keyword is either a reply to a tweet from the university's Twitter account or a mention of it. The following figures show the analytical results of data collected by the keyword (@_ksu):

Figure 3. *Analytical results for @_KSU.*

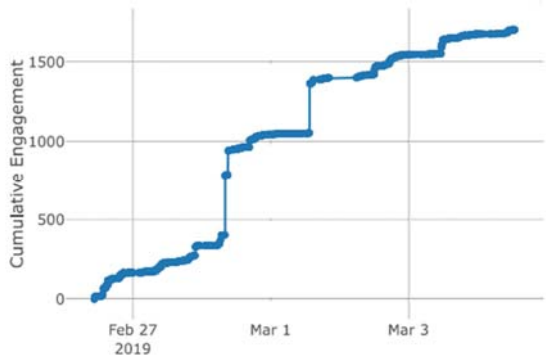

Figure 4. *Engagement timeline for @_KSU.*

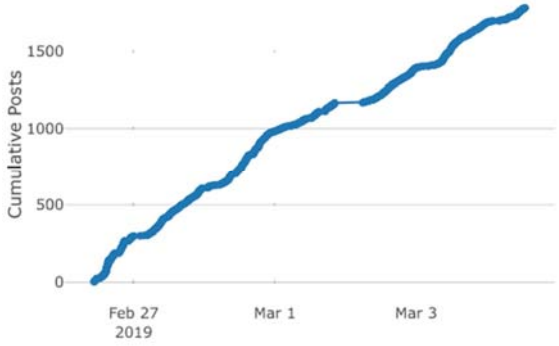

Figure 5. *Posts Timeline for @_KSU.*

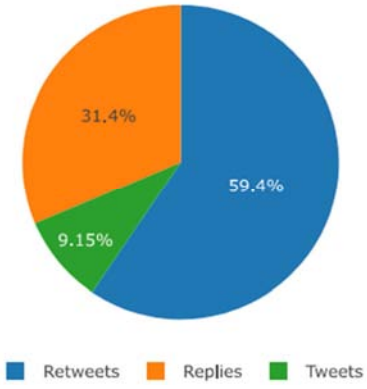

Figure 6. *Types of posts for @_KSU.*

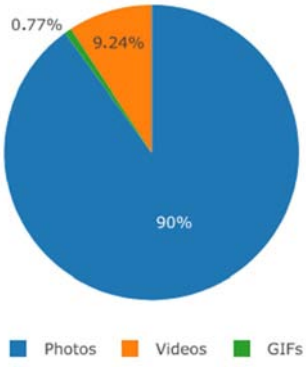

Figure 7. *Types of Rich Media for @_KSU.*

#جامعة_الملك_سعود	390
#الفساد	165
#نزاهة	164
#سعوديات_شامخات_كجبل_طويق	34
#يحدث_الآن	28
#جائزة_رواد_التسويق	27
#tedxksu	24
#خُطى_الاستثمار_المعرفي_وريادة_الاعمال	20
#جائزة_جامعة_الملك_سعود_للتميز_العلمي	17
#ادراج_اللغه_الصينية_في_المناهج	16

Figure 8. *Hashtags for @_KSU.*

We used these two keywords because they are the English and Arabic versions of the hashtags used by students when they are saying something on Twitter about King Saud University. The following are the translations of the Arabic words in all the figures: جامعة الملك سعود means *King Saud University*, الفسادmeans *venality*, نزاهةmeans *integrity*, سعوديات شامخات كجبل طويقmeans *Saudi women as lofty as Mount Tuwaiq*, يحدث الانmeans *happening now*, جائزة رواد التسويقmeans *Marketing Pioneers Award*, خطى الاستثمار المعرفي وريادة الاعمالmeans *cognitive investment and entrepreneurship*, جائزة الملك سعود للتميز العلمي means *King Saud Award for Scientific Excellence*, ادراج اللغة الصينية في المناهج means *Chinese language included in the curriculum*, الجامعات السعوديةmeans *Saudi Universities*, جامعة الملك عبدالعزيزmeans *King Abdulaziz University*, المدينة الطبيةmeans *medical city*, and المتفوقين والموهوبين means *superior and talented*.

Figures 11 to 18 show the analytical results of the data collected by the keyword #KSU. Figures 19 to 26 show the analytical results of data collected by the keyword #جامعة_الملك_سعود.

@_ksu	1.8K
@mohenews	310
@saleh_alnasser_	295
@mohe_sa	225
@minister_moe_sa	219
@nazaha_gov_sa	212
@moisaudiarabia	210
@moi_911	209
@bip_ksa	209
@phds_sa	207

Figure 9. *Mentions for @_KSU.*

Arabic	1.6K
English	101
Undetermined	71
Chinese	25
Catalan	5
Indonesian	2
German	1

Figure 10. *Languages for @_KSU.*

578	785	1.3K	2.1M
Users	Posts	Engagement	Influence

Figure 11. *Analytical results for #KSU.*

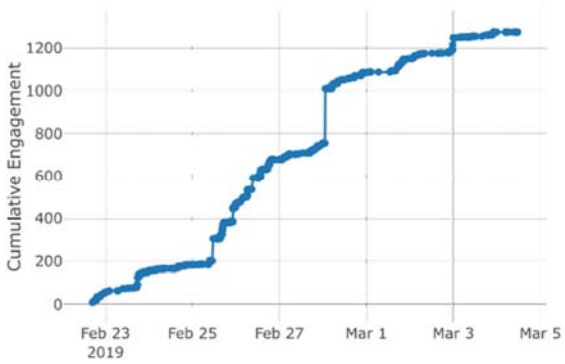

Figure 12. *Engagement timeline for #KSU.*

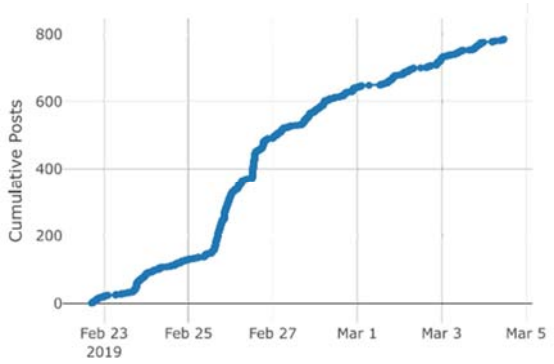

Figure 13. *Posts Timeline for #KSU.*

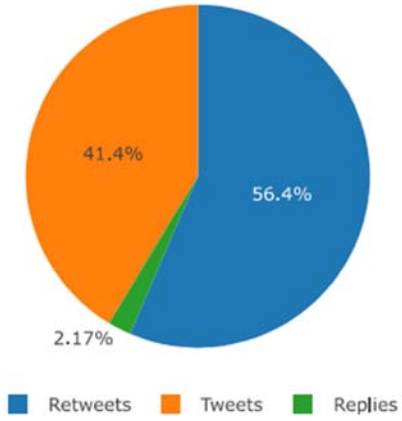

Figure 14. *Types of posts for #KSU.*

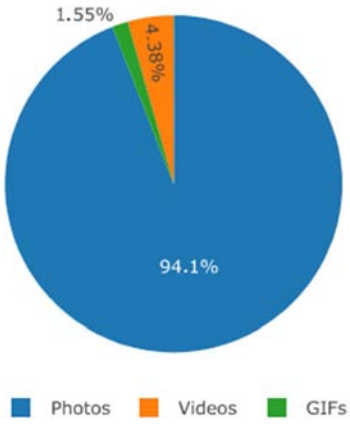

Figure 15. *Types of Rich Media for #KSU.*

#ksu	802
#جامعة_الملك_سعود	206
#جامعة_الملك_عبدالعزيز	180
#الجامعات_السعودية	106
#uwg	67
#pnu	60
#gsu	48
#cau	40
#roadtoksu	36
#asu	36

Figure 16. *Hashtags for #KSU.*

@manaraao	134
@ohaneya1	34
@_luxuryleague	31
@duhhwon	24
@supadopec	24
@woahhtherejay	18
@draftdiamonds	17
@_ksu	16
@theycallme_tanq	15
@thedivinetayj	15

Figure 17. *Mentions for #KSU.*

English	429
Arabic	278
Undetermined	59
Malayalam	7
Japanese	6
Portuguese	2
Turkish	2
Catalan	1
Polish	1

Figure 18. *Languages for #KSU.*

Figure 19. *Analytical results for #جامعة_الملك_سعود.*

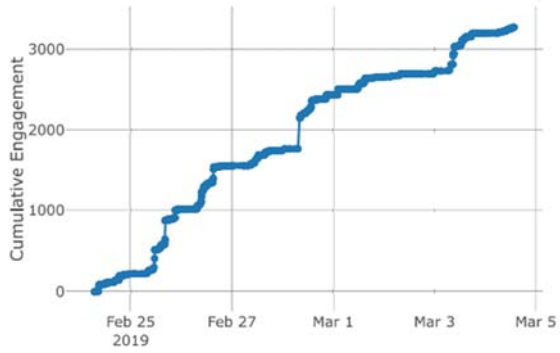

Figure 20. *Engagement timeline for #جامعة_الملك_سعود.*

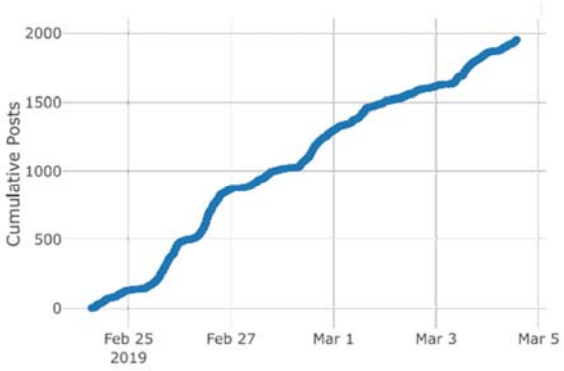

Figure 21. *Posts Timeline for #جامعة_الملك_سعود.*

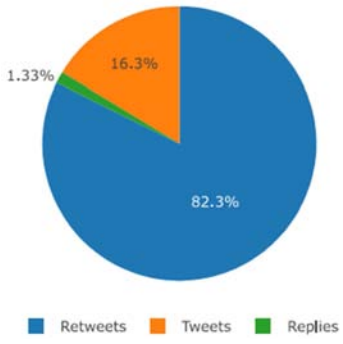

Figure 22. *Types of posts for #جامعة_الملك_سعود.*

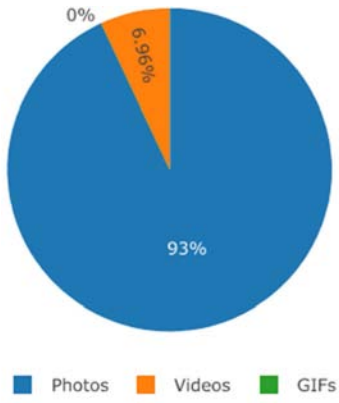

Figure 23. *Types of Rich Media for #جامعة_الملك_سعود.*

#جامعة_الملك_سعود	1.9K
#ksu	209
#جامعة_الملك_عبدالعزيز	201
#نزاهة	191
#الفساد	191
#المدينة_الطبية_الجامعية	166
#المدينة_الطبية_	142
#جسترر	104
#الجامعات_السعودية	99
#المتفوقين_والموهوبين	64

Figure 24. *Hashtags for #جامعة_الملك_سعود.*

@_ksu	528
@ksumedicalcity	292
@phds_sa	227
@minister_moe_sa	194
@nazaha_gov_sa	191
@moi__911	191
@moisaudiarabia	191
@bip_ksa	191
@mohe_sa	191
@manaraao	132

Figure 25. *Mentions for #جامعة_الملك_سعود.*

Arabic	1.7K
Undetermined	219
Catalan	4

Figure 26. *Languages for #جامعة_الملك_سعود.*

Finally, we collected around 4.5 thousand tweets, most of which were written in Arabic. Then, we explained how the tweets were preprocessed and cleaned. There were some issues which we encountered during data collection:

a) In the data collection process, there were few tweets containing malicious terms related on KSU.
b) The data collection process needed to be done using Twitter's API, which was considered to be time-consuming.

3.2. Tweets Cleaning and Annotation

The raw dataset cannot be classified directly; consequently, the second step is to pre-process the data to be suitable for several machine learning approaches. This was done by removing any irrelevant parts of the collected tweets that may produce non-desirable effect during the classification process on the dataset.

3.2.1. Annotation

After collecting data, the corpus of tweets was annotated in order to be labeled into malicious (abusive) messages and

non-malicious (non-abusive) messages. There were 4523 tweets relevant to malicious (abusive) messages, while 2514 appeared to be irrelevant. Table 1 presents a sample of relevant Tweets:

Table 1. *Sample of labeled Tweets.*

Relevant tweet	Arabic	أقول عطونا المكافاة اخلصوا ياجامعة الملك سعود جامعة_الملك ـ سعود@KSU#_
	English	Blackboard is bad @_KSU جامعة_الملك ـ سعود #
Irrelevant tweet	Arabic	تفتح جامعة الملك سعود باب القبول في برنامج يمنح درجة البكالوريوس في إدارة موارد التراث والإرشاد السياحي للطالبات ابتداء من العام الجامعي ١٤٤٠/١٤٤١ #جامعة_الملك ـ سعود
	English	Calm Before the storm: Assessing Climate Change and Sustainability in Saudi Arabia Universities #_KSU

3.2.2. Removing Unwanted Patterns

Cleaning the data can be done using Preprocessor [15] package using Python programming language. Preprocessor is a preprocessing library for tweet information written in Python. When building Machine Learning frameworks dependent on tweet information, a preprocessing is required. This library makes it easy to clean, parse, or tokenize the tweets. Using the package in our code, we were able to clean the data by removing:

a) URLs.
b) Hashtags.
c) Mentions.
d) Reserved words (RT, FAV).

After removing the unwanted patterns, we created a code for removing retweets from the tweets. Then, we manually removed the tweets which were written neither in Arabic nor English.

The data were then manually annotated and classified into three labels; namely: "-1" as a label for an offensive tweet, "0" for a neutral tweet, and "1" for a positive tweet.

Some issues were encountered during the process of data cleaning and pre-processing:

a) During cleaning data, we found most of the tweets to be advertisements.
b) It was not easy to recognize the malicious message in a tweet. The researcher had to detect it independently and to distinguish which tweets were malicious and which of them were non-malicious.

The following table shows the statistics of the collected data after cleaning them:

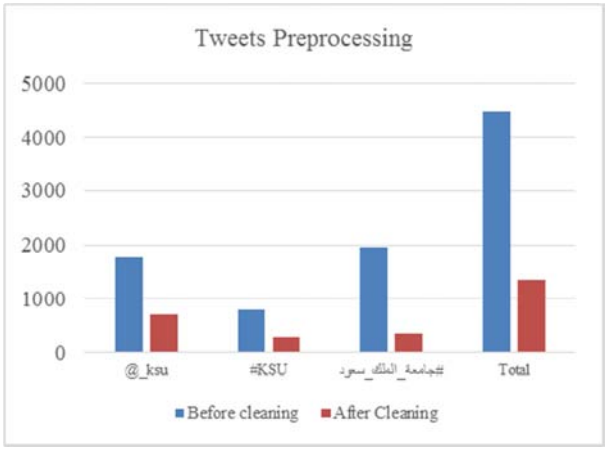

Figure 27. *Tweets Preprocessing Results.*

Table 2. *Number of data tweets before and after cleaning.*

	Before Cleaning	After Cleaning
@_ksu	1783	714
#KSU	786	294
#جامعة_الملك_سعود	1954	349

3.3. Preprocessing Phase

The preprocessing step should fulfill some conditions, that is, it shall not lose any information which is valuable to the task. This step is divided into subtasks. These subtasks are explained in the next subsections. The preprocessing step involves the removal of punctuation marks, the normalization of the annotation tweets, tokenization, stemming, and the removal of stop words.

3.3.1. Removing Punctuations

Punctuation marks such as (!) and (?) are often related to malicious messages [18]. Consequently, we remove the punctuation marks which are single or double that are not important.

3.3.2. Normalizing the Annotation Tweets

Users can write various forms of texts on Twitter such as the word " أنا" which means "me" or write it various other forms such as " انا", "آنا", " إنا", etc. which makes it is difficult to extract individual user attribute.

In addition, diacritic marks can be removed. A diacritic mark is an additional mark attached to a letter either as superscript (ﹶ) or subscript (ﹺ) in different types like Fathah " ً", Kasrah " ٍ", Dammah " ٌ", Sukoon " ْ", Tanween Fath " ً", Tanween Kasr " ٍ", Tanween Damm " ٌ", and Shaddah " ّ". A diacritic mark can change the meaning of a word.

3.3.3. Tokenization and Stemming

In order to carry out the process of tokenization and stemming, we used Tashaphyne [16] Package using python language. Tahsphyne is a finite state automation stemmed based which extracts affixes from a predefined list. It separates all conceivable attachments to a word and refers to all conceivable setup stemming of a given word. This involves extracting the stem of an Arabic word. This capacity indicates the stemming positions (left, right), and at this point, it becomes conceivable to get other determined characteristics like: root, stem, suffixes, and prefixes. It can be used in the following applications: Sentiment Analysis, Text Classification and Categorization, Text Stemming, and Named Entities Recognition. "Tashaphyne: Arabic Light Stemmer" is a library that can perform the following tasks:

Word Segmentation, Arabic word Light Stemming, Word Normalization, Root Extraction, adaptable Light stemmer: plausibility of progress stemmer options and information, Data autonomous stemmer and Default Arabic Affixes list.

Tokenization is the process of splitting the text into a list of separated words. Stemming is the process of getting the root of the word such as stemming “done” to “do”. It is an important step because when extracting features, if the words are not stemmed, there would be unnecessary features. This can be achieved through the process of porter stemming algorithm in Tashaphyne [16]. The following figures present an example of tokenizing and stemming tweets from the collected data. For English tweets, we also used porter stemmer [70] available through NLTK [18].

```
دعوة للطلاب والطالبات للتسجيل والترشح لجائزة خادم الحرمين الشريفين لتكريم
المغتربين والموهوبين.
Tokens-----
['دعوة', 'للطلاب', 'والطالبات', 'للتسجيل', 'والترشح', 'لجائزة', 'خادم',
'الحرمين', 'الشريفين', 'لتكريم', 'المغتربين', 'والموهوبين.']
```

Figure 28. *Example Arabic tweet tokenization.*

```
Join us TONIGHT for our skate party: 's edition!! Wear your best s fit, door
prizes every half hour, food, and more!
Tokens-----
['Join', 'us', 'TONIGHT', 'for', 'our', 'skate', 'party:', "'s", 'edition!!',
'Wear', 'your', 'best', 's', 'fit,', 'door', 'prizes', 'every', 'half', 'hour,',
'food,', 'and', 'more!']
```

Figure 29. *Examples of English tweet tokenization.*

```
Tokens-----
['دعوة', 'للطلاب', 'والطالبات', 'للتسجيل', 'والترشح', 'لجائزة', 'خادم',
'الحرمين', 'الشريفين', 'لتكريم', 'المغتربين', 'والموهوبين.']
Stems-----
['دعو', 'طلاب', 'طالبا', 'سجيل', 'رشح', 'جائز', 'خادم', 'حرم', 'شريف',
'كريم', 'مغترب', 'وموهبين.']
```

Figure 30. *Examples of English tweet stemming.*

3.3.4. Removing Stop Words

We removed the stop words from the preprocessed tweets after the previous steps were taken. As for the Arabic tweets, we used a list of 750 words [21]. As for the English tweets, we used the list of stop words provided by the NLTK [18] package.

3.4. Creating Vector and Feature Selection

In this part, we will specify the main features that help carrying out the implementation phase to detect the malicious messages. The preprocessed tweets we had after taking the previous steps were then translated into word vector. This subtask can be done using the trained vector in [5]. In the created vector, the most important features are the existence of an abusive word in the preprocessed string. We need to find abusive words in both Arabic and English. As for Arabic, we have a list that contains words like (قاطعوهم- سامجة- الله ياخذكم- ذليتو امنا- غثيتونا- ما تستاهلون-معفن- نظامكم خياس-ياصبرنا عليكم- ما بغيتو- ريحونا). The list of abusive words in English is similar to the lists in [17]. In addition, we used the bag-of-words approach to build and carry out the feature extraction process. The wordlist (lexicon) is reached by the straightforward tally of occurrences of each word in the dataset. Then we created a feature vector from that bag of words. We also used the word2vec word embeddings [22] that make a vector space for the Arabic data. As for English data, we used the word2vec embedding that is available via Keras [23].

We added other features in the vector such as emoticons and we used the Emoticon Detector [24] package which provided us with the textual version of all emoticons and the information about how they can contribute to the text features. Furthermore, the exclamation marks in the tweets were also added as another feature.

3.5. Classification

We used Naïve Bayes classifier [19] in the classification of the tweets through the use of NLTK [18] and TextBlob [17]. The code is written using Python on Jupyter Notebook [25]. NLTK is a main stage for structure Python projects to work with NLP. It gives simple to utilize interfaces to more than 50 corpora and lexical assets, such as WordNet [20], alongside a suite of content handling libraries for characterization, tokenization, stemming, labeling, parsing, semantic thinking, and wrappers for mechanical quality NLP libraries. TextBlob is a Python library for handling string information. It gives a basic API to jumping into NLP undertakings.

We used Naïve Bayes classifier in the classification of the tweets. The following results were extracted. Figure 31 shows the results for #KSA while Figure 32 shows the results for #جامعة_الملك_سعود, and Figure 33 shows the results for @_ksu. The overall classification results are shown in Figure 34:

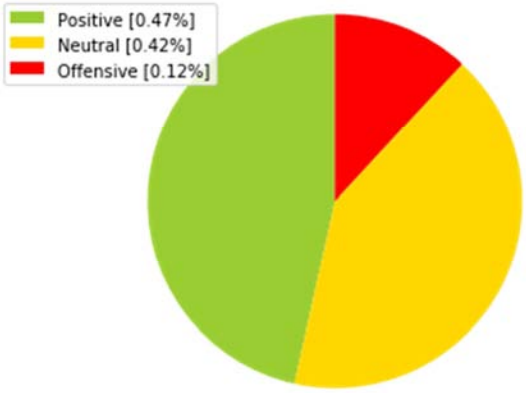

Figure 31. *#KSU Classification Results.*

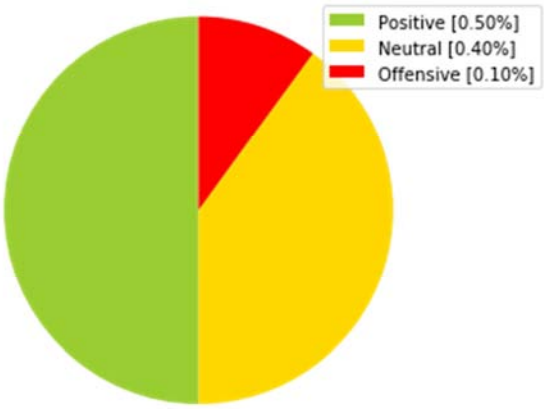

Figure 32. *Classification Results in #جامعة_الملك_سعود.*

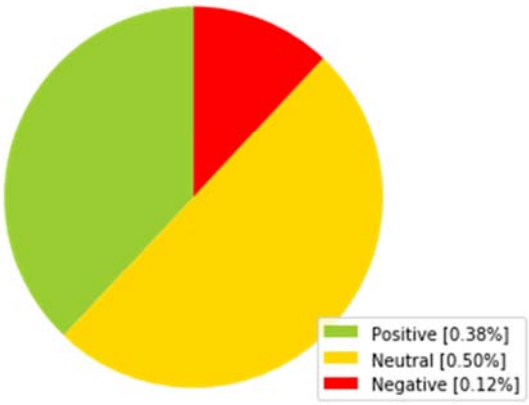

Figure 33. *@_ksu Classification Results.*

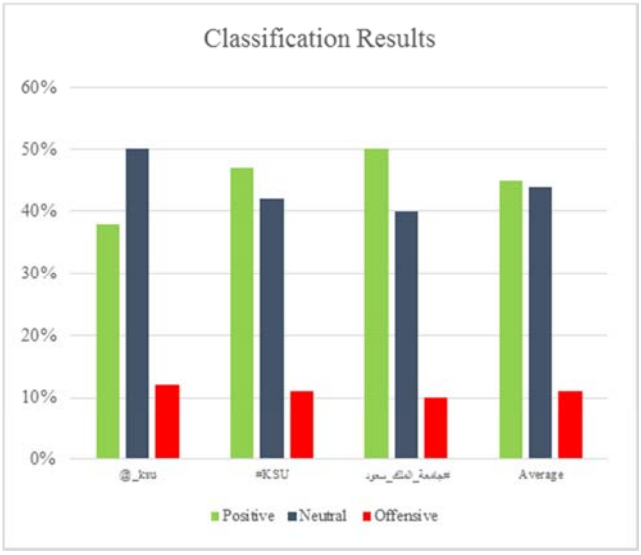

Figure 34. *The overall classification Results.*

4. Analysis and Evaluation

In this part, we aim to examine and compare the performance of two machine learning classification techniques, namely Bernoulli Naïve Bayes and Random Forest, using different extraction features, such as the bag-of-words approach. To achieve this, we will analyze the classifiers' performance using various experiments on the collected dataset.

4.1. Method and Results

We manually annotated the dataset in the three excel files for the three keywords (@_ksu, #KSU, #جامعة_الملك_سعود) and we used -1 as a label for offensive tweets, 0 for neutral ones, and 1 for positive ones. The tweets were analyzed. Figures 35, 36, and 37 show the distribution of tweets over the three classes in @_ksu, #KSU, and #جامعة_الملك_سعودrespectively.

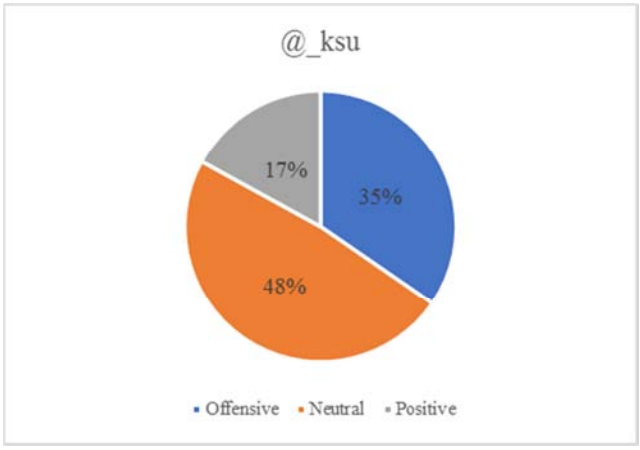

Figure 35. *Class Distribution in @_ksu.*

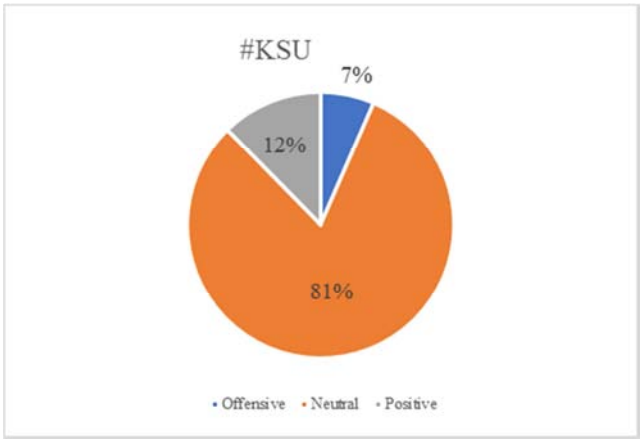

Figure 36. *Class Distribution in #KSU.*

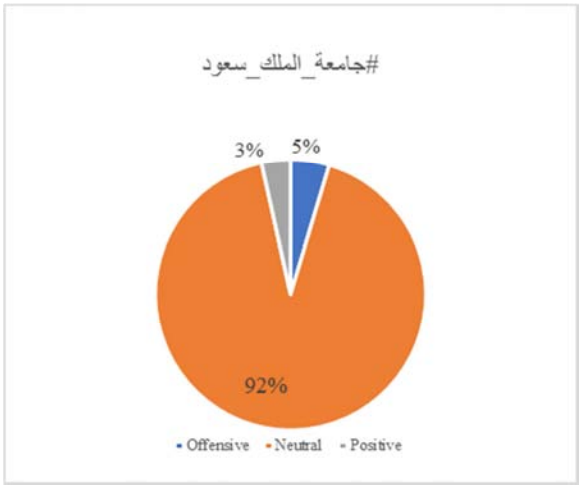

Figure 37. *Class Distribution in* #جامعة_الملك_سعود.

We then analyzed the total distribution of classes on the whole dataset which is represented in figure 38.

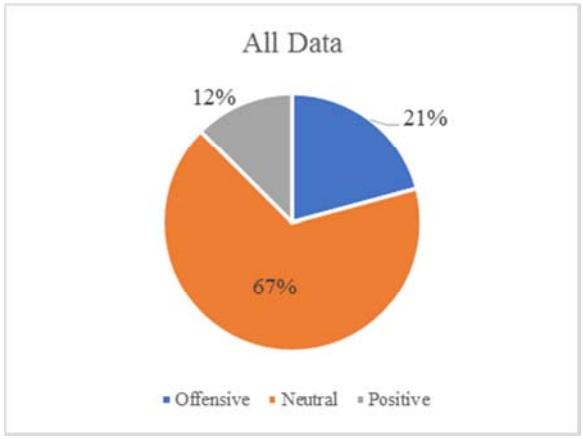

Figure 38. *Class Distribution for all data set.*

Then, we separated the data files of the tweets into Arabic tweets and English ones and we also collected all Arabic tweets in one file. In addition, we did the same with the English tweets. Consequently, we analyzed two separated datasets. Figures 39 and 40 show the analysis results for Arabic and English datasets respectively:

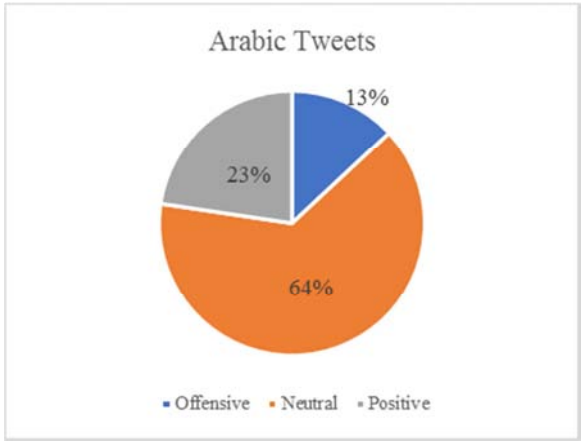

Figure 39. *Class Distribution for all Arabic dataset.*

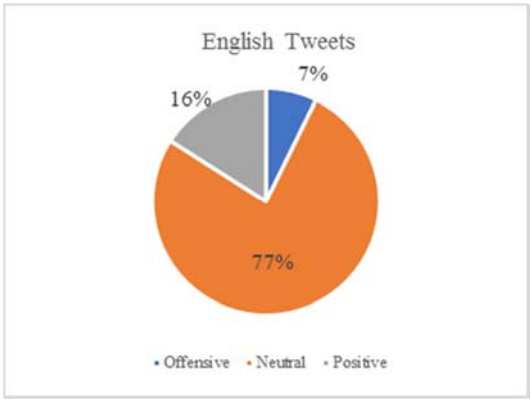

Figure 40. *Class Distribution for all English data set.*

4.2. Experiment1: Using Bernoulli Naïve Bayes Classifier on English Tweets

The purpose of this experiment is to test Bernoulli Naïve Bayes classifier on the English dataset and evaluate its performance. First, the English data were preprocessed, cleaned and tokenized. Figure 41 shows a sample of the tokenized cleaned data:

	text	emotion	tokenized_text
0	[im, at, king, saud, univers, ksu, in, riyadh]	1.0	[Im, at, King, Saud, University, KSU, in, Riyadh]
1	[im, at, colleg, of, medicin, ksu, in, riyadh]	1.0	[Im, at, College, of, Medicine, KSU, in, Riyadh]
2	[rip, king, saud, univers, ksu, in, riyadh]	0.0	[Rip, King, Saud, University, KSU, in, Riyadh]
3	[regist, now, student]	0.0	[Register, now, students]
4	[perfect]	1.0	[Perfect]

Figure 41. *Sample of tokenized cleaned data from English data set.*

Then, we tried to create a bag of words by finding the most frequent word occurrences. Figure 42 shows the most frequent occurrences of words. But these words included stop words, so we had to filter them first before creating the bag of words. Figure 43 shows the most frequent word occurrences after the filtration of stop words. Figure 44 shows the top words that build bag words. Figure 45 shows the most common words across different sentiments (-1: offensive, 0: neutral, and 1: positive.)

```
[('the', 124), ('to', 107), ('in', 88), ('and', 78), ('of', 64)]
```

Figure 42. *Most frequent words in the English dataset.*

```
[('thi', 34), ('univers', 31), ('ksu', 30), ('riyadh', 24), ('amp', 24)]
```

Figure 43. *Most frequent words in the English data set after filtering stop words.*

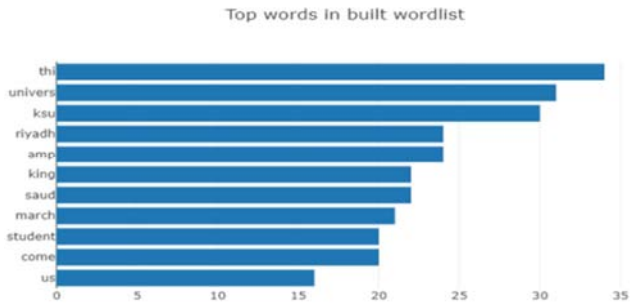

Figure 44. *Most frequent words in the English data set in the built wordlist.*

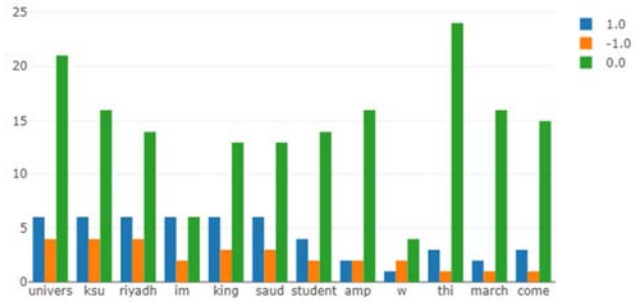

Figure 45. *Most frequent words across different sentiments.*

There are two diverse ways to set up a NB classifier. The first way is the multinomial model that creates one term from the vocabulary in each situation of the record, where we expect a generative model.

The second way is the multivariate Bernoulli model or

Bernoulli model. It creates a pointer for each term of the vocabulary, either 1 demonstrating nearness of the term in the report or 0 showing nonappearance. Figure 46 shows the preparing and testing calculations of the Bernoulli NB algorithm. The time efficiency of both Bernoulli model the multinomial model is similar.

In order to train our classifier, we first separated the dataset into a training set and a testing set. The training set is 70% from the dataset and the testing is 30% from the dataset. The results of the running time and the evaluation of the classifier are shown in Figure 47.

```
TRAINBERNOULLINB(C, D)
1  V ← EXTRACTVOCABULARY(D)
2  N ← COUNTDOCS(D)
3  for each c ∈ C
4  do N_c ← COUNTDOCSINCLASS(D, c)
5     prior[c] ← N_c / N
6     for each t ∈ V
7     do N_ct ← COUNTDOCSINCLASSCONTAININGTERM(D, c, t)
8        condprob[t][c] ← (N_ct + 1) / (N_c + 2)
9  return V, prior, condprob

APPLYBERNOULLINB(C, V, prior, condprob, d)
1  V_d ← EXTRACTTERMSFROMDOC(V, d)
2  for each c ∈ C
3  do score[c] ← log prior[c]
4     for each t ∈ V
5     do if t ∈ V_d
6           then score[c] += log condprob[t][c]
7           else score[c] += log(1 − condprob[t][c])
8  return arg max_{c∈C} score[c]
```

Figure 46. *Bernoulli NB algorithm.*

```
------------------------------------------------
Crossvalidating BernoulliNB...
Crosvalidation completed in 18.72588539123535s
Accuracy: [0.47222222 0.69444444 0.69444444 0.57142857 0.72727273 0.78787879
 0.6969697  0.78787879]
Average accuracy: 0.6790674603174602
------------------------------------------------
```

Figure 47. *Bernoulli NB classifier results on English dataset.*

4.3. Experiment 2: Using Random Forest Classifier with Additional Features on English Tweets and Comparing It with Bernoulli NB

The purpose of this experiment is to test Random Forest classifier on the English dataset and evaluate its performance. Then, its performance will be compared with that of the Bernoulli NB. We used the same schematic technique of tokenization and a feature of vectors was created depending on the bag-of-words approach. But we added the features of emoticons as an additional feature to the feature vector. The exclamation marks were also added to the feature vector.

In Random Forests, we created a choice tree for various Bootstrap tests. When developing the tree, we selected an irregular example of m < p indicators to consider in each progression. This aimed to prompt altogether different (or "uncorrelated") trees from each example. At long last, the expectation of each tree was normal. Random forest algorithm is shown in Figure 49.

The dataset is also divided into training and testing the dataset as in the similar classifier into 70% and 30% respectively. The results of the classifier are shown in Figure 48:

```
------------------------------------------------
Crossvalidating RandomForestClassifier...
Crosvalidation completed in 6.973997354507446s
Accuracy: [0.38888889 0.66666667 0.69444444 0.65714286 0.72727273 0.78787879
 0.75757576 0.66666667]
Average accuracy: 0.6683170995670996
------------------------------------------------
```

Figure 48. *Bernoulli NB classifier results on English dataset.*

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

Figure 49. *Random Forest Classifier.*

The English dataset is so narrow and the effect on the dataset is very modest. The two classifiers have different accuracies but they are almost the same. Figure 50 shows the comparison between both classifiers in terms of accuracy:

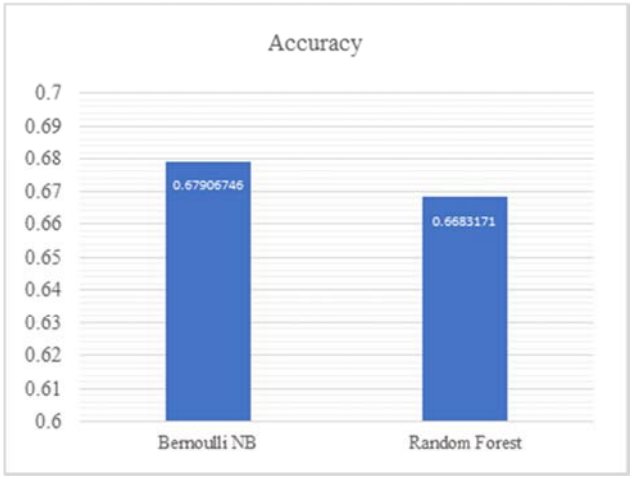

Figure 50. *Comparison of accuracy of both classifiers on English dataset.*

4.4. Experiment 3: Using Bernoulli Naïve Bayes Classifier on Arabic Tweets

The purpose of this experiment is to test Bernoulli Naïve Bayes classifier on the Arabic dataset and evaluate its performance. First, the English data was preprocessed, cleaned, tokenized, and stemmed then we created the bag of words. We used the same algorithm of Bernoulli Naïve Bayes used in the classification before for the English tweets. We separated the data in the same approach into 70% and 30% for the training and testing set respectively.

The results of using that classifier on the Arabic dataset are shown in Figure 51:

```
================================================
Testing BernoulliNB
Learing time 0.12911629676818848s
Predicting time 0.011012554168701172s
================== Results ==================
            Offensive     Neutral     Positive
F1        [0.0952381  0.73318872 0.18965517]
Precision[0.14285714 0.6627451  0.25       ]
Recall    [0.07142857 0.82038835 0.15277778]
Accuracy 0.571875
================================================
```

Figure 51. *Bernoulli NB classifier results on Arabic dataset.*

4.5. Experiment 4: Using Random Forest Classifier with Additional Features on English Tweets and Comparing It with Bernoulli NB

The purpose of this experiment is to test Random Forest classifier on the Arabic dataset and evaluate its performance. Then, its performance was compared with that of the Bernoulli NB. We used the same schematic technique of tokenization and a feature of vectors was created depending on the bag of words. But we added the features of emoticons as an additional feature to the feature vector. The exclamation marks were also added to the feature vector.

The dataset was also divided into training and testing as in the similar classifier (into 70% and 30% respectively). The results of using this classifier on the Arabic dataset is shown in Figure 52:

```
================================================
Testing RandomForestClassifier
Learing time 1.750579833984375s
Predicting time 0.10509467124938965s
================== Results ==================
            Offensive     Neutral     Positive
F1        [0.07017544 0.74213836 0.18867925]
Precision[0.13333333 0.65313653 0.29411765]
Recall    [0.04761905 0.8592233  0.13888889]
Accuracy 0.590625
================================================
```

Figure 52. *Random Forest classifier results on Arabic dataset.*

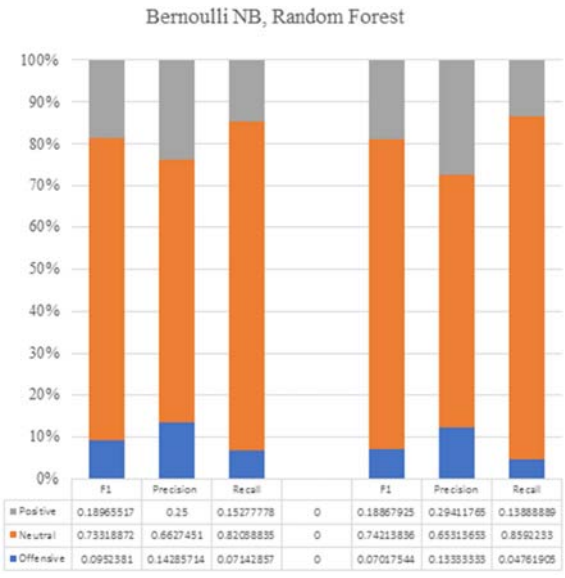

Figure 53. *Stacked representation of evaluation measures for the two classifiers on Arabic dataset.*

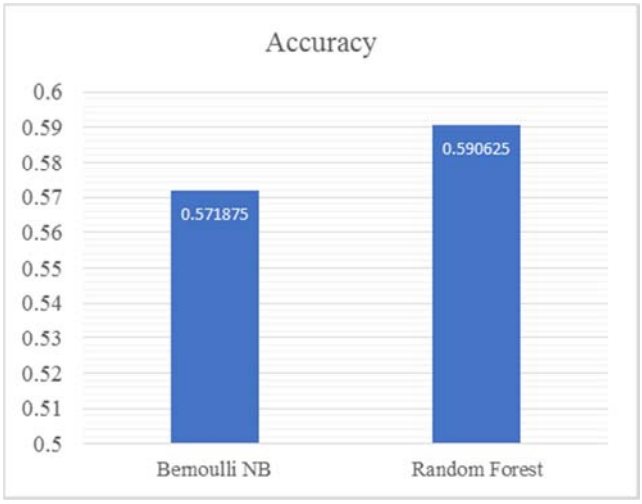

Figure 54. *Comparison of accuracy of both classifiers on Arabic dataset.*

The Arabic dataset is not narrow and the effect on the dataset is explicitly marked in the results. The two classifiers have different accuracies and the effect of the new features has improved the accuracy of the Random Forest classifier. A comparison of the two classifiers is show in Figure 53 which shows a stacked representation of the two classifiers evaluation measures. The Bernoulli NB is shown on the left side while Random forest is on the right side. A comparison between the two classifiers in terms of overall accuracy is shown in Figure 54:

5. Summery

We applied two types of classifier algorithms in our study; namely, the Bernoulli NB and the Random Forest Classifier. Then, the two classifier algorithms were compared in terms of accuracy on Arabic dataset. The best results were achieved using Random Forest classifier with an accuracy rate of 0.590625%.

Afterwards, the same procedures were taken and the accuracy of both classifiers was tested when applied on the English dataset. The best results were achieved using the NB Classifier with an accuracy rate of 0.679067%.

On the one hand, when a comparison was made between the performance of both classifiers on the Arabic dataset, the Bernoulli Naïve Bayes Classifier gave a higher performance, during 0.01101S.

On the other hand, when a comparison was made between the performance of both classifiers on the English dataset, Random Forest Classifier gave a higher performance, during 6.9739.

6. Conclusion

Social networking platforms have been an enormous medium for establishing relations and connections among different people all over the world. People, organizations and companies use these platforms to communicate and interact with their communities and audience. These platforms have

made it easy for people to share information, create content, and communicate and connect with others online; however, online interaction and communication among people have resulted in the creation of many problems. Among problems is the problem of easily publishing malicious content. The first step to avoid the spread of malicious content must involve carrying out a sentiment analysis by using text mining with the aim to reach useful results and conclusions that help in the decision making to decrease these malicious messages.

Detection mechanism is a growing area of research that can detect any inappropriateness of data that is more sensitive to malicious behavior. The detection mechanism needs to be involved in the analysis of the abusing messages posted on the Twitter account of King Saud University (KSU). Text mining is one approach that can be used to detect such malicious or abusing messages. Text mining techniques provide the means to perform data classification where messages can be classified into positive, negative, or neutral. This can be carried out using Machine Learning algorithms such as NB, SVM, Logistic regression, K-nn, and decision tree learning. These algorithms can be carried out on various natural languages such as English and Arabic. However, some challenges may be encountered when analyzing some languages such as Arabic, because the process involves computational linguistics.

There were many challenges that we faced while carrying out the analysis. This includes the necessity to have tools which can deal with the complex Arabic Tweets. In addition, there was a lack of available tools.

In this paper, we aimed to present a classification prototype to detect malicious messages using data mining and text preprocessing techniques. In addition, the impact of text preprocessing techniques on the classifications accuracy and performance in both Arabic collected tweets and English collected tweets was examined. As an important, preprocessing phase, a corpus was maintained and text stemming was carried out for use by the proposed prototype. In the last phase of the research, Machine Learning algorithms, such as NB and Random Forest Classifier were applied in Vicinitas and Python language platforms and the performance of both classifiers was also investigated.

7. Recommendations

Following are the recommendations which the study proposes for future researches:

a) The approaches applied in this study need to be adopted and implemented on all universities' social media accounts. They also need to be applied on video and audio social networking websites.
b) Oher text mining approaches need to be applied to reach and propose the best techniques of malicious text detection.
c) A malicious tweet detection system or mechanism needs to be proposed to be added to Twitter. This can check the tweets before sharing them. It can be applied to an academic account as a beginning.

References

[1] Chapin, J. (2016). Adolescents and cyber bullying: The precaution adoption process model. In *Education and information technologies*, *21* (4), 719-728.

[2] Twitter Inc. (2017). Selected Company Metrics and Financials. Retrieved from https://investor.twitterinc.com/static-files/73896e27-c138-4519-b63f-2cd4b80b568c (Last Accessed, 11 Nov 2018).

[3] Omnicore. (2018, Oct). Twitter by the Numbers: Stats, Demographics & Fun Facts. Retrieved from https://www.omnicoreagency.com/twitter-statistics/ (Last Accessed, 11 Nov 2018).

[4] Smith, K. (2017, Dec). 45 Incredible and Interesting Twitter Statistics. Brandwatch. Retrieved from https://www.brandwatch.com/blog/44-twitter-stats/ (Last Accessed, 11 Nov 2018).

[5] Moss, K. (2017). Results of a Survey of Social Media Use in NYS Libraries. *JLAMS*, *13* (1), 2.

[6] Pew Internet. (2018, Feb). Social Media Fact Sheet. Retrieved from http://www.pewinternet.org/fact-sheet/social-media/ (Last Accessed, 11 Nov 2018).

[7] Edupuganti, V. (2017). *Harassment Detection on Twitter using Conversations* (Doctoral dissertation, Wright State University).

[8] Vandersmissen, B. (2012). Automated detection of offensive language behavior on social networking sites. In *IEEE Transaction*.

[9] Zephoria. (2018). Top 15 valuable Facebook statistics. Retrieved from https://zephoria.com/top-15-valuable-facebook-statistics/. (Last Accessed, 29 Nov 2018).

[10] Internetlivestats. (2018). Twitter Usage Statistics - Internet Live Stats. Retrieved from http://www.internetlivestats.com/twitterstatistics/. (Last Accessed, 29 Nov 2018).

[11] Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. In *Archives of suicide research*, *14* (3), 206-221.

[12] Pew Internet. (2017, Jul). Online Harassment 2017. Retrieved from http://www.pewinternet.org/2017/07/11/online-harassment-2017/ (Last Accessed, 11 Nov 2018).

[13] Lenhart, A., Ybarra, M., Zickuhr, K., & Price-Feeney, M. (2016). In *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute.

[14] Cornaz, N. (2019). An analysis of the# AidToo movement on Twitter: What impacts can a hashtag achieve on sexual exploitation and abuse in the aid sector?

[15] White, G., Wimmer, H., Rebman, C., & Nwankwo, C. (2018). Using Twitter Sentiment Analysis to Analyze Self-Sentiment of the POTUS. In *Proceedings of the Conference on Information Systems Applied Research ISSN* (Vol. 2167, p. 1508).

[16] T. Zerrouki, Tashaphyne, Arabic light stemmer, retrieved from https://pypi.python.org/pypi/Tashaphyne/0.2.

[17] Bharti, S. K., Babu, K. S., & Jena, S. K. (2015, August). Parsing-based sarcasm sentiment recognition in twitter data. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (pp. 1373-1380). ACM.

[18] Garg, P., & Bassi, V. G. (2016). *Sentiment analysis of Twitter data using NLTK in python* (Doctoral dissertation).

[19] Karim, M., & Rahman, R. M. (2013). Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing. In *Journal of Software Engineering and Applications*, *6* (04), 196.

[20] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2014). Ranked wordnet graph for sentiment polarity classification in twitter. In *Computer Speech & Language*, *28* (1), 93-107.

[21] Altowayan, A. A., & Tao, L. (2016, December). Word embeddings for Arabic sentiment analysis. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 3820-3825). IEEE.

[22] Gulli, A., & Pal, S. (2017). *Deep Learning with Keras*. Packt Publishing Ltd.

[23] Pla, F., & Hurtado, L. F. (2017). Language identification of multilingual posts from Twitter: a case study. In *Knowledge and Information Systems*, *51* (3), 965-989.

[24] Perez, F., & Granger, B. E. (2015). Project Jupyter: Computational narratives as the engine of collaborative data science. In *Retrieved September*, *11* (207), 108.

[25] Nabil, M., Atiya, A. F., & Aly, M. (2015, April). New approaches for extracting Arabic keyphrases. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)* (pp. 133-137). IEEE.