A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning

Zaheer Abbass
Department of Computer Science
University of Lahore, Gujrat Campus, Pakistan
Zaheerabbaass@outlook.com

Zain Ali
Department of Computer Science
University of Lahore, Gujrat Campus, Pakistan
Zain.ali_02@Nadra.gov.pk

Mubashir Ali
Department of Computer Science
University of Lahore, Gujrat Campus, Pakistan
mubbashircheema@gmail.com

Bilal Akbar
Department of Mechanical Eng.
MUST
10250, AJK, Pakistan
Bilal.akbar@must.edu.pk

Ahsan Saleem
Department of Computer Science
COMSATS University, Wah Campus, Pakistan
Ahsansaleem111@yahoo.com

Abstract: An increasing amount of data and information coming from social networks that can be used to generate a variety of data patterns for different types of investigation such as human social behavior, system security, criminology etc. A framework is developed to predict major types of social media crimes (Cyber stalking, Cyber bullying, Cyber Hacking, Cyber Harassment, and Cyber Scam) using the data obtained from social media website. The proposed framework is consist of three modules; data (tweet) pre-processing, classifying model builder and prediction. To build the prediction model Multinomial Naïve Bayes (MNB), K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) is used that classify given data into different classes of crime. Further N-Gram language model is used with these machine learning algorithms to identify the best value of n and measure the accuracy of the system at different levels such as Unigram, Bigram, Trigram, and 4-gram. Results shows that all three algorithm attain the precision, Recall and F-measure above than 0.9 however Support vector machine performed slightly better. The proposed system produced better accuracy result as compared to existing network-based feature selection approach.

Keywords: Natural Language Processing (NLP), Twitter, Supervised Machine learning, Information Extraction, Topic Modeling, Social Media Crime.

I. INTRODUCTION

Now a day's new wave of social media technologies such as Facebook, blogs, wikis, microblogging, twitters plays a vital role in formal and informal communications [6]. Microblogging site (twitter) is an electronic platform where users share their ideas, thoughts, and news in under 280 characters of text. Twitter is a unique way of following friends and sending tweets (Twitter messages) unlike any other social media networks because twitter friendship is not mutual. For example, you can follow the celebrities without requiring them to follow you back [1]. Twitter plays a virtual online world for its users. Virtual world interacts like a real world where location act as an intermediate connection. Commonly used GPS feature in Smartphone and tablet-enabled social media users to attach real-time locations when sending out tweets [2]. For example, a tweet related to an event like a tornado might be written in a very short time after a user witnessed a tornado was formed. The information could be spread faster than any other electronic media (TV, news or website). Secondly, tweets contain information that could help to evaluate the actual situation of the event [3].

In 2018 monthly active Twitter user was 330million that were posting 500 million tweets per day [4]. These massive tweets having diverse dimensions of data, used by researchers for different types of inquiries to predict future trends like future marketing outcomes, forecasting box-office movies revenues, flu spreading diseases, disaster response, crime prediction, forecasting election result [5] etc.

Crimes occur everywhere in the world, for increase rate of crimes law enforcement agencies are demanding advanced information systems that can help to reduce the crimes and protect the society [8]. Criminology is the scientific study of crime to find out the causes of crimes by collecting and investigating data [7,8]. That way Natural Language Processing is a good approach for text analysis.

II. RELATED WORK

Monitoring unethical behavior and preventing cybercrimes is the interest of all the countries of the world to reduce risk or danger. Individuals or group activities on a computer generate a massive amount of data that can be used to get information about individuals or group activities. Crime information is not easily available to the public. A framework is proposed to demonstrate how theft information extracted from online news articles from three

978-1-7281-6332-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICSC.2020.00073

different countries [9]. This research only focuses on theft crime only. Name entity relation algorithm (NER) was used to identify the location where the crime had happened. NER along with conditional random filed (CRF) is used to identify. It is a statistical modeling approach to classify whether a sentence in news articles is crime location or not. The main steps proposed framework are building corpus related theft crime, data preprocessing, location identification, feature selection, label assignment, training conditional random field model and at last crime sentence classification.

A survey report of world health organization is found that there were 788,000 people all around the world who committed suicide in 2015. It is an alarming condition for the world because the rate of the ratio in suicide is increasing day by day where the average ages of victims were between 20 to 30 years [11]. The use of social networking site Twitter to make suicide pacts by people with each other was also studied. The research study is focused to South Korea where the rate of suicide is highest in the world that is (36.1/100000) in 2015. To implement the proposed framework tweets related to the term suicide pacts are extracted. All twitter posts who contain the word suicide pacts between 16, October 2017 to November 2017 are included in dataset. Total data crawling duration was 45 days where the total number of tweets extracted is 489, but only 1702 twitter tweets contain the world suicide pacts, which was posted by 551 Twitter users. Six variables (name of the city where the user lived, gender, age, preferred methods of suicide pacts, preferred gender, and information about how to contact the user) about the user to make suicide pacts with each other is analyzed. The result of research indicates that Twitter is an attractive platform where social media users make suicide pacts.

A research work done in twitter text analysis is to detect the hacking behavior and different communication forms or patterns used by social media users [12]. There is a certain level of skills motivation and knowledge used by computer hackers to trap the social media users easily. In this research study, a set of indicators is used to perform analysis of social media communication, discussion patterns, threats, user activities, positive and negative sentiment to examine the hacking behavior and communication pattern of hackers. This study used a SKRAM (Skills, knowledge, Resources, Access to targets, and last one motivation) model to discover unethical behavior of criminal users. SKRAM model is used by different computer experts to access possible threats to any organizational information system. SKRAM stand for that means different types of skills extensive knowledge of hardware and software use different resources to trap the target or social media users or use different types of motivation. SKRAM model is implemented to examine the hacking behavior and social media communication patterns of hackers. To perform the experiment, they extracted 12M twitter posts to examine the hacking behavior and communication patterns of hackers.

III. PROPOSED METHODOLOGY

The proposed system is composed of three modules. First module is tweet pre-processing that is used as an input to the second module for generating a trained model and final is Prediction as. shown in Figure 1.

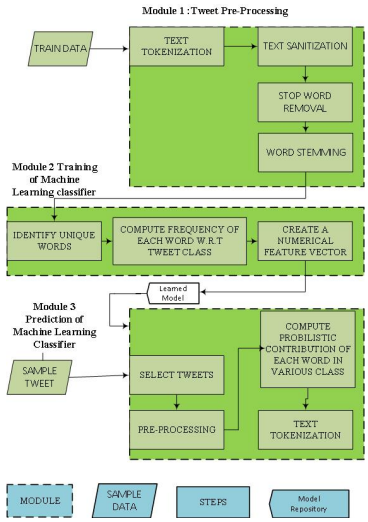

Figure 1. A three stage proposed frame work for cybercrime prediction

A) Data Collection

Researchers and application developers get a huge number of Tweets for different data analysis via Streaming API [7].150k tweets collected for proposed framework according to predefined classes which is illustrated in table 1.

TABLE I. NUMBER OF TWEET RELATED TO DIFFERENT CATEGORIES OF KEYWORDS.

Categories	Number of tweets
Stalking	35k
Harassment	30k
Bullying	30k
Hacking	35k
Scam	20k
Total Number of tweets	150k

1) Cyber-Harassment

Harassment can consist of warning, worrisome, emotionally harmful, or sexual messages delivered via an electronic medium that lead the fear or distress much like real-world harassment [13, 14]. data

collected data for harassment class by using popular hashtags like #harassment, #Metoo, #Sexual harassment and #Sexualassualt for further crime analysis.

2) Cyber-Bullying

Cyberbullying or suicide is a worldwide public health problem and major causes of death among youngster. According to the Pan American Health Organization (PAHO), 1 million people commit suicide each year [15]. Hinduja and Patchin define cyberbullying as a deliberate, repeated and hurtful activity using the computer, mobile phone and other electronic devices [16]. Some other causes of cyberbullying are demographic, confounding hygiene and class differences in the society [17]. Popular hashtags like #cyberbullying, #victim, and #suicide are used to collect data for cyberbullying class.

3) Cyber-Stalking

Most famous cases involve the stalking of celebrities and college students than the general public [18,19]. Data collected for this by using popular hashtags like #stalking, #love, #someone, and #kill.

4) Cyber-Scam

Cybercrimes mostly occur due to the collapse of technology but we cannot neglect the human elements involved in it. Phishes use social engineering and psychological tactics with the help of social media like Twitter, Facebook to influence individual [20,21]. Following hashtag used to collect data #scam, #hoax, #scammer, and #fraud.

5) Cyber-Hacking

Professional IT expert prepare cyber-attack on web [22,23]

B. Data Preprocessing

The data gathered from Twitter is unstructured and noisy Following preprocessing steps are performed to remove noise form data.

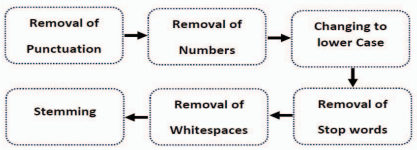

Figure 2. Detail of data cleaning steps.

1) Tweets Tokenization

The first step in data preprocessing is tweets segmentation or tweets tokenization. Word tokenizing is a basic unit for text analysis [31

2) Removal of Punctual Marks

The second phase in tweets processing is the removal of punctual marks such as period, semicolon, comma, question mark, ellipsis, exclamation point, quotation marks, parentheses and apostrophe from the dataset.

3) Stop Word Elimination

The most frequently used words in the text are called stop words. The words which come frequently in the text has very low worth.

4) Lower-case conversion

Text analysis treats both upper and lower case letters equally [33]. The size of feature words is increased if we deal with both upper case and lower-case letter in our training corpus

5) Stemming

Another important step in data preprocessing is stemming. Stemming reduces the word to its base form by removing affixes that save time and space

B) Feature Selection

The process of selecting an appropriate feature form a huge collection of processed data. [24]. There are several objectives of feature selection for instance, it reduces the training time, easier to interpret, improve the accuracy and enhance the performance of model if effective attributes are selected from the corpus [26]

Important feature selection techniques are Document Frequency (DF), Information Gain (IG), CHI-test (X2 statistic), Mutual Information (MI), Odds Ratio, Gini Index and term frequency and inverse document frequency (TF-IDF) [1,2,3].

1) TF-IDF

The TF-IDF approach is used to select feature words from Social crime dataset. TF-IDF stand for Term frequency- inverse Document Frequency which means that if a term comes frequently in a particular document it has less importance and less weighting in that document [27,28].For example, if t represent the term frequency or word frequency in the document d, and N represent the number of complete document and n is the number of documents containing the term or word frequency t.

Then the formula for IDF is as follows:

$$DF = log\left(\frac{N}{n}\right) \quad (1)$$

The formula for Calculating TFIDF as:

$$w_k(d) = \frac{tf_k(d) log(\frac{N}{n_k} + 0.01)}{\sqrt{\sum_{k=1}^{n}(tf_k(d))^2 \times log2(\frac{N}{n_k} + 0.01)}} \quad (2)$$

n_k Mean that number of documents containing the term or word frequencyt_i.

Scikit is a python library used to extract features from the text in the numerical form. Vectorization is the process which transformed a collection of text documents into numerical feature vectors [28].

C. Algorithms

These classifiers are Multinomial Naïve Bayes, Support Vector machine; K Nearest Neighbors:

1) Multinomial Naïve Bayes

NAÏVE Bayes classifier is a simple, faster, efficient and easy to implement [28].

In plain English, the major classifier (Multinomial Naïve Bayes) can be written as

$$Posterior = \frac{Class\ Prior\ probability\ \times likelihood}{predictor\ prior\ probability} \quad (3)$$

To classify a document using Naïve Bayes, if d_k is given document, the probability of the class C_j is calculated as

$$p(C_j|d_k) = \frac{P(d_k|c_j)\cdot p(c_j)}{P(d_k)} \quad (4)$$

The class label d_i can be determined as if $P(d_k)$ is the same for all class.

$$label(d_k) = argMax_{C_j}\{P(C_j|d_k)\}$$

$$= argMax_{cj}\{P(d_k|C_j).p(C_j)\} \quad (5)$$

Calculating the probability of $P(d_k|C_j)$ for Multinomial Naïve Bayes classifier. The occurrence of each feature word in a document is independent of its position and the frequency of each feature word is counted.

If the number of times a word W_q occur in the document di asn_{kq}. The probability of equation 6 can be calculated as:

$$P(d_k|C_j) = P(|d_k|)|d_k|!\prod_{k=1}^{|V|} P(\frac{W_q|C_j}{n_{kq}!})^{n_{kq}} \quad (6)$$

The number of words in a document is d_k denoted as $|d_k|$, now the probability of likelihood $P(C_j)$from equation 2 is estimated as:

$$p(C_j) = \frac{1+n_j}{l+n_{all}} \quad (7)$$

The number of documents is denoted by $n_j$in the classC_j, where $l$the number of class is, n_{all} is the number of the entire document in the training dataset.

Now calculate the probability of$P(W_k|C_j)$

$$P(W_{qk}|C_j) = \frac{1+N_{cjk}}{N_{all}+N_j} \quad (8)$$

N_j Number of words in classC_j,$N_{cjk}$is the number of words $W_q$in classC_j, N_{all} the number of all word in the training dataset.

2) K- Nearest Neighbors

KNN stand for K-Nearest Neighbors and is a well-known supervised machine learning classifier used for tweets classification.KNN can be defined mathematically as follow to measure the cosine similarity.

$$\frac{\sum_{k-1}^{t}(d_{jk}.t_k)}{\sqrt{\sum_k^t d\ jk^2}.\sum_{k=1}^{t} t\ k^2} \quad (9)$$

d_{jk}Is the weight of term K in training corpus i.

k_t Is the weight of the term K in test corpus t.

3) Support Vector Machine

SVM is widely used for short text classification based on the principle of structured [36]. Mathematically SVM is defined as following:

$$f(x) = sign(wy - b) \quad (10)$$

Where w is the weight vector. Support vector machine find the hyperplane as follow

$$y = wy + b \quad (11)$$

IV. Results and Discussions

Pre-processed dataset is divided into two sets as 70% training and 30% testing Effectiveness and accuracy of each machine learning algorithm are measured by using standard evaluation criteria such as

precision, recall, and F-measure. Precision, recall, and f-measure can evaluate by using following equation 12, 13 and 14.

$$Precision = \frac{True\ Positves}{True\ positives + False\ Positves} \quad (12)$$

$$Recall = \frac{True\ Positves}{True\ positives + False\ Negatives} \quad (13)$$

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (14)$$

Table 2, 3 and 4 gives the Precision, Recall and f1 score of three supervised algorithms.

TABLE II. MULTINOMIAL NAÏVE BAYES RESULTS.

Social Crime Category	Precision	Recall	F-measure
Cyber-bullying	0.79	0.99	0.88
Cyber-Harassment	0.95	0.89	0.92
Cyber-Hacking	0.98	0.92	0.95
Cyber-Stalking	1.00	0.91	0.96
Cyber-Scam	1.00	0.78	0.88

TABLE III. K NEAREST NEIGHBOR RESULTS.

Social Crime Category	Precision	Recall	F-measure
Cyber-bullying	0.79	0.99	0.88
Cyber-Harassment	0.95	0.89	0.92
Cyber-Hacking	0.98	0.92	0.95
Cyber-Stalking	1.00	0.91	0.96
Cyber-Scam	1.00	0.78	0.88

TABLE IV. SUPPORT VECTOR MACHINE RESULTS.

Social Crime Category	Precision	Recall	F-measure
Cyber-bullying	0.79	0.99	0.88
Cyber-Harassment	0.95	0.89	0.92
Cyber-Hacking	0.98	0.92	0.95
Cyber-Stalking	1.00	0.91	0.96
Cyber-Scam	1.00	0.78	0.88

A. Comparative analysis of Proposed Model

Comparative analysis has been performed between three supervised machine learning classifiers (MNB, SVM, and KNN)

TABLE V. COMPARATIVE ANALYSIS OF SUPERVISED MACHINE LEARNING CLASSIFIERS.

Categories	Precision	Recall	F-measure	Accuracy
MNB	0.93	0.92	0.92	91.5%
KNN	0.93	0.92	0.92	91.5%
SVM	0.94	0.93	0.93	92.0%

B) Comparative analysis with competitor

The result of proposed framework compared with existing state of art which improves better accuracy.

TABLE VII COMPARATIVE ANALYSIS WITH COMPETITOR

Techniques Adopted	*Accuracy*	
Proposed Approach	BOW	92%
Competitor approach	Network-based feature selection	90%

B. Evaluation of proposed System using N-Gram Language Model

N-gram language model. N-gram is a statistical language model [34] In our case system shows the best result with bigram where vale of n=2.

TABLE VIII COMPARATIVE ANALYSIS OF SUPERVISED MACHINE LEARNING CLASSIFIERS USING N-GRAM LANGUAGE MODEL

Classifiers	*Unigram*	*Bigram*	*Trigram*	*Quad gram*
MNB	0.774%	0.915%	0.627%	0.426%
KNN	0.774%	0.915%	0.627%	0.426%
SVM	0.774%	0.920%	0.649%	0.420%

V. CONCLUSION AND FUTURE WORK

The aim of this research study to predict social media crimes by suing twitter data. We use three ML classifier with bag of word model. The study proves better result with existing state of art. The proposed model is currently offline in future work it can be extended for real-time Twitter data streaming to predict further crimes. More crime classes can be added to make the system efficient and robust.

REFERENCES

[1] Axel Bruns and Jean Burgess, "The use of Twitter hashtags in the formation of ad hoc publics. In Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference, University of Iceland, Reykjavik, 2011.

[2] Hernandez-Suarez; Sanchez-Perez; Toscano-Medina; Martinez-Hernandez; Sanchez: Perez-Meana. A Web Scraping Methodology for Bypassing Twitter API Restrictions. CS IR: 2018; 1.

[3] Shih-Feng Yang; Julia Taylor Rayz. An Event Detection Approach Based On Twitter Hashtags. CS,SI April 2018, 1.

[4] Danah Boyd, Scott Golder, Gilad Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter", Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010.

[5] Andranik Tumasjanl; Timm O. Sprenger1; Philipp G. Sandner1; Isabell M. Welpel. Election Forecasts with Twitter: How 140 Characters Reflect the Political Landscape. SS CR 2011, 4, 402-418.

[6] Xuefei (Nancy) Deng; Yibai Li; K. D. Joshi. Introduction to HICCS-51 Minitrack on Digital and Social Media in Enterprise. Hawaii International Conference on System Sciences, Hawai, 2018.

[7] Haewoon Kwak; Changhyun Lee; Hosung Park; Sue Moon. What is Twitter, a Social Network or a News Media, 2010, 591-600.

[8] Xiaforg Wang; Donald E. Brown; Mathew S. Gerber. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information, in: Intelligence and Security Informatics. Springer 2012.

[9] Rexy Arulanandam; Bastin Tony Roy Savarimuthu; Maryam A. Purvis. "Extracting crime information from online Newspaper articles" Proceeding AWC '14 Proceedings of the Second Australasian Web Conference, Volume 155, Pages 31-38, Auckland, New Zealand — January 20 - 23, 2014.

[10] Xiaofeng Wang; Matthew S. Gerber, Donald E. Brown. Automatic Crime Prediction using Events Extracted from Twitter Posts. Proceedings of the 5th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction, University of Virginia, April 2012.

[11] S.Y. Lee, Y. Kwon, "Twitter as a Place Where people meet to make suicide Pacts (2018), Public Health 2018, page21-26.

[12] Olga Babko-Malaya, Rebecca Cathey, Steve Hinton, David Maimon, Taissa Gladkova, "Detection of Hacking Behaviors and Communication Patterns on Social Media", IEEE International Conference on Big Data, Boston, MA, USA , June 2017.

[13] Vivian E. von Gruenigen, Beth Y. Karlan," Sexual harassment in the work place: Its impact on gynecologic oncology and women's health", Gynecologic Oncology, 2018, 149(2), 227-229.

[14] Filipa Pereira.; Brian H. Spitzberg.; Marlene Matos.; Cyber-harassment victimization in Portugal: Prevalence, fear and help-seeking among adolescents. Computers in Human Behavior, (2016), 62, 136-146.

[15] Juliana Escobar Chavarria, Laura Elisa Montoya González*, Diana Restrepo Bernal, David Mejía Rodríguez, "Cyberbullying and suicidal behavior: What is the connection", Universidad CES, Medellin, Colombia, 18 October 2017.

[16] Despoina Chatzakouy, Nicolas Kourtellisz, Jeremy Blackburnz Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakaliy, "Detecting Aggression and Bullying on Twitter", Aristotle University of Thessaloniki Telefonica Research University College London, 12 May 2017.

[17] Nadine Shaanta Murshid "Poor hygiene and bullying victimization in Pakistan", University at Buffalo School of Social Work, 685 Baldy Hall, University at Buffalo, Buffalo, NY 14260, United States, 2018.

[18] Patrick Q. Brady.; Matt R. Nobles.; Leana A. Bouffard. Are college students really at a higher risk for stalking Exploring the generalizability of student samples in victimization research. Journal of Criminal Justice, 2017, 5, 12-21.

[19] Brett Eterovic-Soric, Kim-Kwang Raymond Choo, Helen Ashman, Sameera Mubarak "Stalking the stalkers detecting and deterring stalking behaviours using technology A review". Computers & Security, 2017, 70, 278-289.

[20] Nicole L. Muscanell, Rosanna E. Guadagno, Shannon Murphy, "Weapons of Influence Misused: A Social Influence Analysis of Why People Fall Prey to Internet Scams", Journal of Social and Personality Psychology Compass, 2014, 8, 388-396.

[21] Anupama Aggarwal, Ashwin Rajadesingan, Ponnurangam Kumaraguruy, "Automatic Real-time Phishing Detection on Twitter", 2012 eCrime Researchers Summit, Las Croabas, Puerto Rico, 23-24 Oct. 2012.

[22] Saad Alsunbul, Phu Le, Jefferson Tan, Bala Srinivasan, "A Network Defense System for Detecting and Preventing Potential Hacking Attempts", International Conference on Information Networking (ICOIN), Kota Kinabalu, Malaysia, 13-15 Jan. 2016.

[23] Daniel Nguyen, "State Sponsored Cyber Hacking and Espionage", APRIL 13, 2015.

[24] Jian Sun, Xiang Zhang, Dan Liao, Victor Chang, "Efficient Method for Feature Selection in Text Classification". International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21-23 Aug. 2017.

[25] Mohamed Abdel Fattah, "A Novel Statistical Feature Selection Approach for Text Categorization", Journal of information processing systems, 2017, 13(5), 1397-1409.

[26] George Forman," An Extensive Empirical Study of Feature Selection Metrics for Text Classification". Journal of Machine Learning Research, 2003, 3, 1289-1305.

[27] Guifen Zhao, Yanjun Liu, Wei Zhang, Yiou Wang, "TFIDF based Feature Words Extraction and Topic Modeling for Short Text, 2nd International Conference on Management Engineering, Software Engineering and Service Sciences, Wuhan, China, January 13 - 15, 2018.

[28] Mingyong Liu1 and Jiangang Yang, "An improvement of TFIDF weighting in text categorization, International Conference on Computer Technology and Science, Singapore 2012.

[29] Hetal Doshi, Maruti Zalte, "Performance of Naïve Bayes Classifier Multinomial Model on Different Categories of Documents", National Conference on Emerging Trends in Computer Science and Information Technology (ETCSIT), 2011.

[30] Jingnian Chen, Houkuan Huang, Shengfeng Tian, Youli Qu, "Feature selection for text classification with Naïve Bayes", Expert Systems with Applications, 2009, 36(3), Pages 5432-5435.

[31] Jonathan J. Webster & Chunyu Kit, "TOKENIZATION AS THE INITIAL PHASE IN NLP". Proceedings of the 14th conference on Computational linguistics - Volume 4; 1106-1110, Nantes, France - August 23 - 28, 1992.

[32] Mazhar Iqbal Rana, Shehzad Khalid, Muhammad Usman Akbar, "News Classification Based On Their Headlines: A Review". 17th IEEE International Multi Topic Conference, Karachi, Pakistan, 8-10 Dec. 2014.

[33] Alper Kursat Uysal, Serkan Gunal, "The impact of preprocessing on text classification", "Information Processing and Management", International Journal of Information Processing and Management, 2014, 50(1), 104 - 112.

[34] Riyad Al-Shalabi Rasha Obeidat, "Improving KNN Arabic Text Classification with N-Grams Based Document Indexing", "Proceedings of the 6th International Conference on Informatics and Systems INFOS, Cairo-Egypt, March 27-29, 2008.