Comparative Analysis of Feature Extraction Techniques for Event Detection from News Channels' Facebook Page

Farzana Kabir Ahmad
School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia.
farzana58@uum.edu.my

***Abstract*—Event detection from the Social Network sites (SNs) has attracted significant attention of many researchers to understand user perceptions and opinions on certain incidents that have occurred. Facebook is the most famous SNs among internet users to express their opinions, emotions and thoughts. Due to its popularity, many news channels such as BBC have created a Facebook page to allow reader to comment on news reported, which has led to an explosion of user-generated data posted on the Internet. Monitoring and analyzing this rich and continuous user-generated content can yield unprecedentedly valuable information, enabling users and organizations to acquire actionable knowledge. Previously, in the context of text mining research, various feature extraction techniques have been proposed to extract relevant key features that could be used to detect the news posts into corresponding event. However, these techniques are separately tested on different data. Moreover, analyzing large number of news posts over a period of time is a challenging task due to its complex properties and unstructured data. Thus, this paper has proposed a comparative analysis on various types of feature extraction techniques on three different classifiers, namely Support Vector Machine (SVM), Naïve Bayes (NB) and K-Nearest Neighbor (kNN). The aim of this research is to discover the appropriate feature extraction technique and classifier that could correct detect event and offer optimal accuracy result. This analysis has been tested on three news channels datasets, namely, BBC, Aljazeera, and Al-Arabiya news channels. The experimental results have shown that Chi-square and SVM has proven to be a better extraction and classifier technique compared to other techniques with optimal accuracy of 92.29%, 87.12%, 87.00% have been observed in BBC, Aljazeera, and Al-Arabiya news channels respectively.**

***Index Terms*—Text mining, Event detection, Feature extraction technique, News channels, Social Network sites**

I. Introduction

Recent years have witness a great attention to the text mining studies due to the availability of large amount of opinion data that have been generated in Social Networks Sites (SNs) such as Facebook and Twitter. Facebook is the most famous and common SNs among internet users with more than 800 million active users. Most of the time, Facebook users used this medium to express their, feelings, opinions emotions and thoughts.

Consequently, many news channels have taken this opportunity to create their own Facebook pages in order to improve the interaction between readers by allowing them to comment on the news posts.

The explosion of user-generated data has led to numerous works in the data mining field. A large number of studies have examined Facebook post data mainly to understand, extract and summarize useful Facebook content to detect events which has become an important emergent research topic. Facebook event mainly correspond to a lot of content generated on Facebook, which is opinions, reactions and information from users. Currently, the most promising events that exist on Facebook are natural disasters such as earthquakes, health epidemics like influenza, trends such as world-cup, terrorism, opinion about products, services or events like political election results.

Generally, in the context of text mining, these news posts can be used as a monitor mechanism to detect the significant events which have happened around the world by extracting key features. Key features are a set of major words in a post that provide a high-level description of its contents to readers. Furthermore, those features are very important to be recognized as it will contribute to the correct detection of post's category. Although many studies have been conducted in this area, detecting relevant key features on a specific event from news posts on Facebook is a difficult task.

This is due to the fact that large numbers of news posts have been released continuously over a period of time [1]. Moreover, the complex properties of news posts have further complicate the analysis process. Generally, news posts may include a large subject domain in which it contains different features that belong to various categories, complex event descriptions or various group of people. Besides that, news posts about a specific topic may grow or vanish in intensity for some periods of time [1]. This leads to another difficulty in order to extract relevant features from large number of news posts.

Previously, various feature extraction techniques such as Chi-squared [2], Pointwise Mutual Information (PMI), Information Gain (IG) [3], Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) [4]-[6] have been used for extracting features. However, most of the studies have tested these techniques separately. Hence, this study aims to identify the appropriate feature extraction method that could lead to the determination of significant key features from large number of news posts and consequently improve classification performance. Thus, in this study, various number of feature extraction techniques namely TF-IDF, TF, Chi-squared, and Information Gain, are examined with different n-gram models (unigram, bigram, and trigram) in order to identify which combination of extraction technique and n-gram models that could offer better classification accuracy. This analysis is tested on prominent classifiers such as SVM, NB, and KNN on three different news channels datasets such as al-Arabiya, Aljazeera, and BBC.

The remainder of this paper is organized as follows. Section II describes previous works on feature extraction techniques. Section III presents the procedure of comparative analysis; meanwhile experimental results are discussed in Section IV. Finally, Section V offers concluding and future direction remarks.

II. Feature Extraction Techniques in Text Mining

The last few decades have witness a tremendous increase in usage for SNs such as Facebook and Twitter as it offers a valuable source for real time news and act as a platform to share thought. Ideally, monitoring and analyzing this rich and continuous user-generated content can yield unprecedentedly valuable information, enabling users and organizations to acquire actionable knowledge. Typically, such news posts can be used as a medium to detect significant events which have happened all around the world such as airplane crash, terrorism, conflict, and so on. However, extracting significant features for those events is a difficult process and extremely time consuming, especially when it is done manually. As a result, various computational works have been done in this area, which mainly used text mining as an approach to analyze the unstructured data. Trend of monitoring users' opinions have been studied lately in Twitter as this subject has attracted the attention of scholars and professionals [6]. In contrast, only relatively a small number of studies have been noticed using Facebook [4], to either identify trends, determine relevant topic classification, or detect posts' sentiments. This may occur due to limitation of data volume and type of available data since it involved Facebook privacy policies, which restrict researchers from collecting personal data.

Previous studies have mostly address the event detection problem usually by using news documents corpus as input and features extraction process is executed to determine the key features that can differentiate diverse events. Feature extraction mainly aims to select suitable set of keywords based on the whole corpus of news documents, and assign weights to those key words. Feature extraction methods usually treat the document as a group or Bag of Words (BOW). Many studies have been carried out on Event Detection and Tracking (EDT) over social media, but mainly focus on analyzing long text discussions from blog posts [7], and Google trends [8]. In contrary, currently there is few existing research related to events trends detection on Facebook, and the number of studies are still relatively small [9]. Additionally, majority of work in EDT field has followed the content based approach, in which content features of a document are selected using unigram or n-grams, then the document is represented as a vector of features in a text stream. Consequently, the new arrival document is classified based on the similarity with the existing pool of events. If the similarity exceeds a specified threshold, the document is classified into the nearest events otherwise, a new event will be formed. Diverse feature extraction methods have been introduced such as lexicon based methods, machine learning techniques, and statistical methods. The main feature extraction methods are described in the following subsection.

A. *Term Frequency-Inverse Document Frequency (Tf-Idf)*

TF-IDF is a popular way to compute terms' weights using the following formula:

$$TF\text{- } IDF = TF * \log N/n \tag{1}$$

where N is the total text of all categories, and n is the number of texts which contain term t. Researchers [6] have converted each token into its TF-IDF weight to find out the most common terms for each eighteen labelled categories. These terms later are used as a training set to build the classifier in order to predict the topic of tweets. On other hand, other previous work [4], [9] have used modified TF-IDF to predict Facebook post's topic since they found that original formula is not applicable of handling the content shared on Facebook because of the limited length of the Facebook posts. The modified formula is as following, where w is a weight for term, ti:

$$W(ti) = 2\, TF(ti) * LOG\ IDF\ (ti) \tag{2}$$

B. *Term Frequency (Tf)*

TF weight is the standard method used in information retrieved since it indicates the most significant relative features which can be used to represent the document. TF has various weighting scheme such as Binary Term Occurrences (BTO) which uses the presence of a term (value 1) or not (value 0). On other hand, Term Occurrences (TO) uses the presence of a term as TRUE or not as FALSE. Although some studies have shown that binary weighting is better compared to frequency scheme for polarity classification task, the feature's frequency is much applicable for event detection. This perhaps due to fact that event detection is highly depend on content features that seem to be repeated. However, this finding is still an open area of research.

C. *Chi-Square (X^2)*

Chi Squared Statistic technique calculates the weight of attributes with respect to the class attribute. The higher the weight of an attribute, the more relevant it is considered. The chi-square statistic is a nonparametric statistical technique, which is used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies. Chi-square statistics use nominal data, thus instead of using means and variances, this test uses frequencies. The value of the chi-square statistic is given by:

$$X^2 = Sigma\ [(O\text{-}E)\ 2/\ E] \tag{3}$$

where X^2 is the chi-square statistic, O is the observed frequency and E is the expected frequency. Furthermore, it is proved to be better in performance than Pointwise Mutual Information since it is a normalized value and its value is more comparable across terms in the same category.

D. *Information Gain (Ig)*

Information Gain computes the weight of attributes regarding to the class attribute by using the information gain. The higher the weight of an attribute, th is defined as follows:

$$IG\ (F,S) = \sum S_v/S\ (Entropy\ (S) - Entropy\ (S_v)) \tag{4}$$

The main problem in information gain is that it may assign high weights to attribute that insignificant mainly when tested with large number of different values.

III. Methodology of Comparative Analysis For Social Event Detection

The research framework of this study is portraits in Figure 1 and Section A to Section E elaborates further each phase in detail.

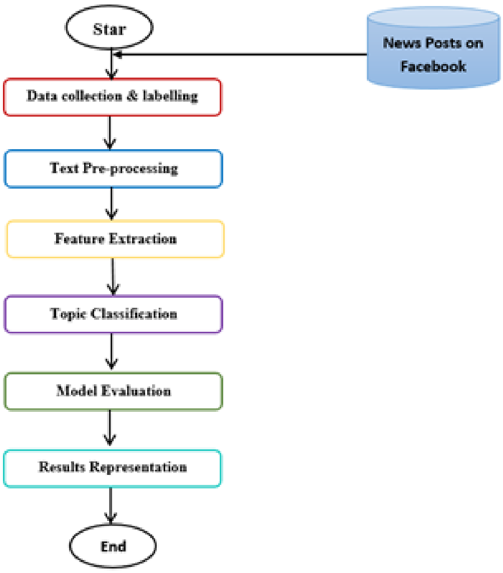

Figure 1: Main steps in detecting social events

A. Data Collection and Labelling Phase

News channels posts on Facebook are used in this study as a tested data. Around 35,858 posts released in 2014 have been extracted from three different news channels namely Aljazeera, BBC and Al-Arabiya. Thus, in order to collect all these news posts, this study has used an application called Netvizz that utilizes Facebook API for gathering purposes. Subsequently, this study has applied the third model supported by Netvizz which focuses on page contents. Additionally, news posts have collected through specifying date from January to December 2014 for each news channel. There are four types of news posts which include link, status, video and photo. However, this study only focused on link, photo and status types since some of news channels have zero video posts (Al-Arabiya channel) as shown in the Table 1.

Table 1
Statistical analysis on three news channels

News channel	Video Posts	Link Posts	Status Posts	Photos Posts	Number of Posts
Aljazeera	2,685	11,987	858	1,099	16,629
BBC	1,230	883	2	4,136	6,251
Al-Arabiya	0	12,308	149	521	12,978
Total	3,915	25,178	1,009	5,756	35,858

Each extracted post contains several details such as type of post, post message, link of post, domain link, post published date and time, number of comments and likes. All those information have been exported and saved in .xls format for next phase. Prior to pre-processing step, a filtering step has been performed on the extracted posts in order to remove any post that does not belong to one of the following categories (airplane crash, disease, natural disaster, conflict, and terrorism). These particular topics have been selected for labelling since these categories were the topics that captured the attention of most people from various media forms like newspaper, and TV. Thus, only 3,985, 3,872, and 1090 news posts from each news channels (Aljazeera, al-Arabiya, BBC) respectively have been used in building event detection model.

B. Pre-Processing Phase

In order to enhance the process of detection and extraction features as well as to increase the performance of event classification for the news posts on Facebook, a pre-processing step is required to be performed on the all three datasets prior to the subsequent phase. The pre-processing process for this study includes several steps as following:

i. Remove manually the URLs that exist within the context of news posts.
ii. Tokenization that splits the post's text into sequence of tokens/words, where each token represented along with its occurrences number in all documents.
iii. Transformation case to change the case of all words in the posts into the lower case letters.
iv. Stop words removal to get rid of all insignificant common English words from the datasets by applying English stop words filter operator.
v. Stemming to return the word to its original root by removing the suffix or prefix parts. This process is executed using the Porter stemmer for English words.
vi. Generate n-Grams terms to create term n-Grams of tokens (unigram, bigram, and trigram), in a document. A term n-Gram is defined as a series of consecutive tokens of length n. Once the data is pre-processed, it is ready for next step which is feature extraction phase.

C. Feature Extraction Phase

Feature extraction method is an essential step that could lead to the determination of robust key features from large number of news posts, where those extracted features can be used as representatives for a specific event to detect the new upcoming news posts. Hence, this research has applied different feature extraction methods such as TF-IDF, TF, TO, BTO, Chi-Squared, and IG. In addition, each feature extraction technique has implemented on all three n-grams (unigram, bi-gram, and trigram).

D. Event Detection Phase

The classification task in this study involves detecting event from news posts into one of their corresponding category (conflict, natural disaster, disease, terrorism, and airplane crash) using different classification techniques. Several classifiers techniques were used in this experiment such as SVM, NB and kNN with k = 1. This study used RapidMiner's default settings for the all classifiers.

E. Evaluation Phase

In order to evaluate the performance of the classifier in detecting event from news posts, this study has used 5-fold cross validation. Each dataset is randomly divided into 5 subsets (each subset has equal number of examples). In this process, five iterations took place and each iteration involved a training model and a testing model. In addition, the proposed model is evaluated by using accuracy metric.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Various feature extraction weighting techniques have been tested in this study along with different types of n-gram such as unigram, bigram, and trigram. Furthermore, well-known machine learning classifiers were used for the classification process such as SVM, NB, and kNN. The key idea in this study is to determine which feature extraction technique and classifier that would lead to the robustness of event detection model.

According to the Figure 2 to 4, BBC dataset is considered as the smallest dataset with only 6,251 posts compared to the other datasets (Aljazeera, and l-Arabiya) which contain almost equal number of posts. However, regardless the large number of news post in Al-Jazeera and Al-Arabiya, the categories in these datasets are not well distributed throughout the year, which is contradict with the BBC dataset. There is also a significant difference in term of number of posts between categories, whereby conflict and terrorism news post are much higher compare with the other three other categories (airplane crash, disease, and natural disaster). This finding has indicated that conflict and terrorism are two events that mostly reported by these three new channels in the year of 2014. Conflict category represents 64% of the total number of posts in both Aljazeera and al-Arabiya news channel, whilst it represents 41% for BBC channel. Terrorism category on the other hands has a ratio of convergence between 20% and 27%. Moreover, the distribution of posts for airplane crash category is considered to be the lowest compared to other categories for all datasets expect for March and April, 2014. This is mainly due the missing Malaysian airplane incident. Therefore, the high weighted mean precision can be observed for airplane crash category during these two months.

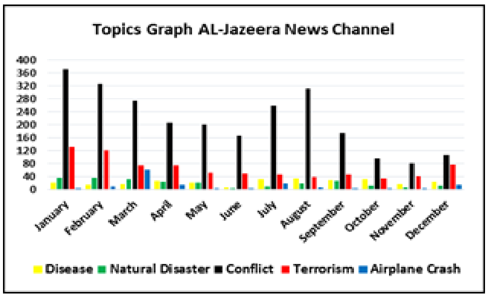

Figure 2: Topics Graph for Al-Jazeera News Channel

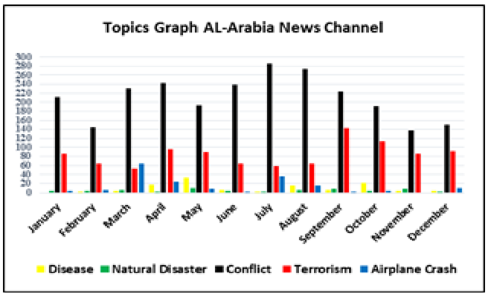

Figure 3: Topics Graph for Al-Arabia News Channel

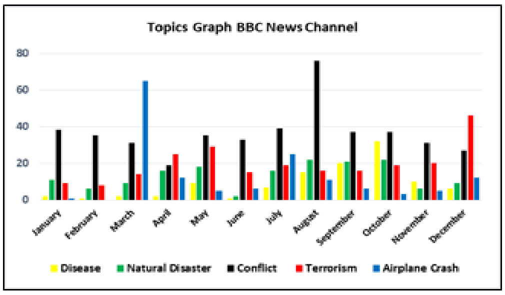

Figure 4: Topics Graph for BBC News Channel

Table 2 on the other hands shows the best accuracy obtained for each news channel dataset along with which features extraction weighting technique and the type of n-gram that have been used in this study. It has been discovered that all datasets, SVM shows better accuracy compared to other classifiers with the used of Chi-square extraction technique with 87.12% in Al-Jazzera, 87.00% in Al-Arabiya and 92.29% in BBC channel. The optimal accuracy has been achieved on BBC dataset as there is no significant difference in the distribution of news posts over all categories. In addition, unigram type is found relevant for both Aljazeera and BBC, meanwhile trigram features are considered significant in searching for keyword features for al-Arabiya dataset. It is also interesting to note that unigram have proved the most effective n-gram type for SVM and KNN classifiers while trigram have performed well with NB classifier as shown in Table 2.

Table 2
Optimal accuracies achieved for each news channel for each classifier along with used feature extraction technique and n-gram type

Channels	Classifier	FE technique	N-grams	Accuracy
	SVM	Chi-square	1	92.29%
BBC	NB	Chi-square	3	89.07%
	KNN	TF	1	87.78%
	SVM	Chi-square	1	87.12%
Al-Jazeera	NB	Chi-square	3	84.03%
	KNN	TF	1	82.19%
	SVM	Chi-square	3	87.00%
Al-Arabia	NB	Chi-square	3	84.23%
	KNN	TF	1	80.79%

V. CONCLUSION

This study has examined different feature extraction methods to detect social event from three Facebook news channels, namely Al-Jazeera, BBC and Al-Arabiya. Six feature extraction techniques which include TF-IDF, TF, BTO, TO, IG, Chi-square have been used for extracting different features types to determine optimal accuracy for topic classification. Additionally, diverse n-gram types such as unigram, bigram, and trigram are used in counting all adjacent n words in a news post. The results of the experiment have showed that Chi-square has proved to be a better extracting technique compared to other techniques as it leads to the highest classification accuracies of 92.29%, 87.12%, 87.00% for BBC, Aljazeera, and Al-Arabiya dataset, respectively. In addition, the experiment has shown that SVM classifier tends to be superior classifier comparing to NB and KNN. Moreover, unigram features have proved to be the most effective feature type which has assist in obtaining high classification accuracy. In future work, sampling technique will be introduced to address the unequal data distribution in these new channel datasets. Moreover, sentiment analysis is another research route that could incorporate readers' emotion on these news posts in order to gain deeper understanding of their perspective.

ACKNOWLEDGMENT

This work was supported by RIMC UUM and KPT FRGS Grant.

REFERENCES

[1] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, & T. By. Sentiment Analysis on Social Media. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (2012), 919–926.

[2] D. Clarke, P. Lane, & P. Hender. Developing Robust Models for Favourability Analysis. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (2011) 44–52. Association for Computational Linguistics.

[3] H. Uğuz. A Two-stage Feature Selection Method for Text Categorization by using Information Gain, Principal Component Analysis and Genetic Algorithm. Knowledge-Based Systems, 24(7) (2011) 1024–1032.

[4] I. P. Cvijikj, & F. Michahelle. Monitoring Trends on Facebook. Proceedings - IEEE 9th International Conference on Dependable, Autonomic and Secure Computing, DASC (2011) 895–902.

[5] L. –W. Ku, Y. –T. Liang, & H. –H. Chen. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. AAAI spring symposium: Computational approaches to analyzing weblogs 100107 (2006).

[6] K. Lee, D. Palsetia, R. Narayanan, M. M. A Patwary, A. Agrawal & A. Choudhary. Twitter Trending Topic Classification. Proceedings - IEEE International Conference on Data Mining, ICDM, (2011) 251–258.

[7] N. Bansal & N. Koudas. BlogScope: A System for Online Analysis of High Volume Text Streams. Proceedings of the 33rd International Conference on Very Large Data Bases, (2007) 1410–1413.

[8] H. Choi & H. Varian. Predicting the Present with Google Trends. Economic Record, 88(s1) (2012) 2–9.

[9] J. Akaichi, Z. Dhouioui, & M. J. Lopez-Huertas Perez. Text Mining Facebook Status Updates for Sentiment Classification. IEEE- 17th International Conference in System Theory, Control and Computing (ICSTCC) (2013) 640–645.