Journal of Theoretical and Applied Information Technology
30th November 2015. Vol.81. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TREND DETECTION IN THE ARABIC SOCIAL MEDIA USING VOTING COMBINATION

[1]ALI SABAH ABDULAMEER, [2]SAIDAH SAAD, AND [3]LAILATUL QADRI ZAKARIA

Center for Artificial Intelligence Technology (CAIT)
Faculty of Information Science and Technology University Kebangsaan Malaysia, 43600 Bangi, Selangor Malaysia.

E-mail: [1]alisabah361@yahoo.com, [2]saidah@ukm.edu.my, [3]lailatul.qadri@ukm.edu.my

ABSTRACT

The amount of information has been increasing tremendously, especially with the use of social media applications, such as Twitter, Facebook, and YouTube. Twitter is a common social application that enables users to share their current thoughts and actions, comment on breaking news, and engage in discussions. Trends are typically driven by emerging events, breaking news, and general topics that attract the attention of a large fraction of Twitter users. Thus, trend detection is highly valuable to news reporters and analysts because they may point to fast-evolving news stories. Researchers have been attempting to detect trends using machine-learning techniques, such as clustering method based on major languages (e.g., English, German, and French). The Arabic language remains in its infancy, but the Arabic social media have been contributing to a large amount of data because of the significant events in the Middle East. The present research aims to detect trends in the Arabic social media. However, this research must overcome several issues such as processing of Arabic user-generated content and lack of resources. To solve these issues, this research presents a voting combination clustering approach, which is divided into six phases, namely, dataset collection from Twitter, text pre-processing, spam filtering using Naïve Bayes, feature selection based on term frequency–inverse document frequency and entropy, statistical analyses, and evaluation. Three statistical approaches for clustering are used, namely, co-occurrence, k-means, and voting combination. The analyses are performed to classify the trends into three categories, namely, Arabic nationality events, personal events, and other events. Experimental results indicate that the voting combination clustering achieved 93%, 87%, and 90% for precision, recall, and f-measure in trend detection, respectively. Finally, trend detection of events is important to companies, governments, national security agencies, and journalists to develop strategies to rectify them.

Keywords: *Trend detection, Arabic social media, Term Clustering, K-means and Voting Combination***.**

1. GENERAL INTRODUCTION

Web 2.0 changed users from traditional passive consumers to active participants who can share opinions, experiences, and information with one another. These individuals are provided with a technology platform such as social networks (SNs), which are products of Web 2.0 technology, where they connect, produce and share content online [1].The public nature of user-generated content (UGC) on social media platforms enables trend monitoring to be used as a tool to gather insight into the topics that attract the attention of a large fraction of users. The value of the content generated on SNs as a recognized information source induces individuals to rely on SNs for opinions and real-time news. Platform providers also support this usage by offering the feasibility of search via a wide range of public status updates to find temporally related information or to monitor content. Moreover, they have also given the opportunity to search the public status updates via their search application programming interfaces (APIs), which have progressed research and commercial efforts to gather knowledge through analysis of shared content. Social media, which are continuously growing given the widely increasing number of users, usually focus not on the headlines of traditional media but on the activity or work of ordinary people. Hence, the identification of trending topics in SNs, such as Twitter, provides a comprehensive overview of the topics and issues that are popular in the online community [2]. The detection of trends reveals a valuable insight into topics that gain the attention of a large volume of social network users. Trending public opinions are interesting for individuals and news reporters because they highlight rapidly evolving stories. Sociologists use such opinions to show the expanding morale of people, whereas marketing professionals use them to identify causes related to monitoring and brand value analysis and benchmarking. Opinion-tracking companies use trending public opinions to collect inferences about elections results. Finally, scholars employ these opinions to elaborate the phenomena of SNs. With the increase in available sources and volume of information, companies and individuals interested in the rapid evolution of trends monitor conversations on social media platforms. These individuals become more dependent on tools that can monitor and automatically detect topics. Thus, research on text analysis and the practical application of existing information have advanced to obtain benefits from the extensive knowledge embedded within UGC.

The attention of scholars and professionals on trend monitoring in English social media sites is not a new concept. Such trend resulted in multifold increases in modification and new algorithms for information extraction and commercial online software. Online SNs in Arabic language have now become one of the major pathways for gathering information and making friends for the young generation. Therefore, the present study tests the provided information and benchmark setters for such media. Considerable research has investigated the evolution and adaptation of trends in western online SNs, but the lack of research on Arabic social media proves that it remains a new concept. Therefore, the present study is directed towards Arabic social media to contribute to the trend monitoring of said social media. Moreover, this research is based on the proposition and evaluation of the trend detection system based on Arabic posts circulating in social media.

2. RELATED WORK

A number of studies have employed trendiness of terms throughout a specific time-span to identify trending topics. Although extensive studies have been conducted in terms of natural language processing (NLP), the trend of using NLP techniques in social media has started only recently, particularly in Twitter because it was only established in 2006. However, in this reasonably brief period, several enlightening studies have been conducted with reference to Twitter. Specifically, numerous recent NLP techniques have been applied to twitter. The enBlogue system, which screens Web 2.0 streams, such as blog postings, tweets, and RSS news feeds, to discover unexpected increase in the level of popularity of tags [3]. A study in the University of Toronto has developed a system known as "Twitter-Monitor," which is capable of tracking trends over Twitter by initially identifying key terms that abruptly show up in tweets at an abnormally higher rate and by categorizing them into trends depending on their co-occurrences [4]. Furthermore, has employed temporal information related to online documents and determined surfacing issues and subjects of interest and their variations by sensing jargon in the documents. A candidate term has been regarded as jargon if its level of concentration surpasses the limit [5].

Goorha et al. [6] have proposed a system that keeps track of news reports, blogs, reviews, and tweets that focus on a subject of interest (i.e., merchandise or organization) and extracts a hundred words across the subject of interest and identify phrases that can initiate a trend. A phrase is considered trending if it has appeared over a particular minimum number of times on any given day or lately appeared over a specific number of times and increased by over a specific percentage above its latest rate of incidence. A phrase is considered substantial if it is regularly stated and fairly distinctive to the referenced product.

Budak et al. [7] have proposed two structure-based trend definitions. They defined coordinated trends as those trends wherein the popularity of a topic is indicated by the number of related users talking about the topic. By contrast, uncoordinated trends are trends in which the score of a topic depends on the range of extraneous individuals interested in it. To facilitate the discovery of a coordinated trend, they offer a high score to subjects that are intensely talked about in a group of closely linked nodes by analyzing the count for each node based on the total count of its neighbors.

Gloor et al. [8] have proposed algorithms for mining the web to determine tendencies and the individuals who start trends. They have considered concepts in the form of associated phrases from a specific area as inputs. The geodesic distribution of the concept in its communication network is computed in the first stage. In the second stage, the social network location of the concept's author is added to the metric to include context-specific components of nodes in the social network. Positive or negative sentiment is assessed in the third stage. The actors employ the concepts in this stage. Recently, Nguyen & Jai. Have detected trending topics from the data stream of specific Twitter accounts. They have employed a method that can detect novel trends by accumulating relevant messages from related groups through content-based approaches and temporary information and by sending information about the messages among the members of a social community in real time. The proposed system enables the user to keep track of sensational topics effortlessly, which are generally talked about lately. However, the effectiveness is rather a concern. Therefore, enhancement of system performance by re-implementing improved algorithms is crucial [9].

3 RESEARCH METHODOLOGY

The research methodology of this study consists of seven major phases; Twitter API, Collection of tweets, Pre-processing, NLP tool, and topic identification, Trend detection (tweet clustering) and evaluation. As shown in Figure 1, all these phases are described and it will be illustrated in the following sections.

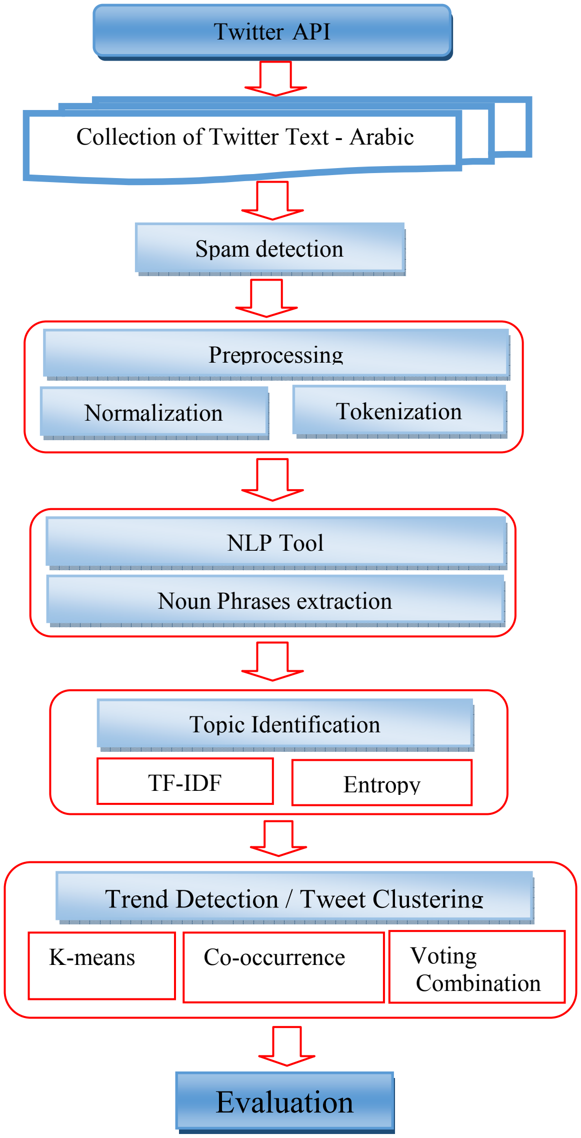


Figure 1: The phases of the research methodology

3.2 Data Collection

The data collection for trend detection in Arabic social media is a challenge itself because the current research supports a fresh concept based on the development of an Arabic body consisting of Arabic tweets and comments. Therefore, a system using the API of Twitter is required to collect tweets in Arabic. The system feeds data from Twitter and extracts comments from Twitter accounts on a 24-hour basis.

To validate the recommended method, we selected Twitter and sourced various datasets from this popular online social networking service. Twitter has more than 500 million active members. These users send more than 340 million tweets every day. Through these tweets, users express their views about movies and dramas, review sites, various forums, products and services, other people, special occasions and blogs among others. The news-event dataset we selected considers each news piece as a Twitter topic. The Twitter trends dataset pertains to topics tagged by Twitter as hot and happening each hour of the day.

In the first dataset, tweets that match all the words in the headlines were sourced. For the second dataset, we utilized Twitter's stream API for sampling tweets, which all have a public status on the site. The data collected from Twitter consist of 1,512 public tweets posted from February 1 to March 26, 2015. These posts are mostly about trending events and related to trending topics such as the death of Saudi Arabia King Abdullah, Islamic State of Iraq and Sham (ISIS), war in Yemen, and Iraqi prime minster. The posts contain several types of information including status, video, link, and photo. The collective distribution of posts over time of day has been investigated, under the presumption that it can possibly affect the distribution of the trending topics. The results obtained are illustrated in Figure 2.

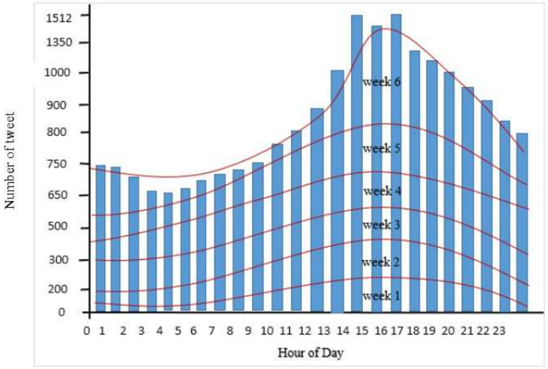


Figure 2: Time Distribution of Tweets

The period of interval for gathering data was centered on two considerations. First, the period should be long to ensure that the trending topics are captured. Second, the period should simultaneously be short enough to facilitate almost real-time monitoring while overpowering the processing challenges presented by the large-sized datasets. The outcome of our system's parameter tuning is based on this rationale.

3.3 Spam Detection

Numerous spam posts or tweets are usually distributed over the social media. Some of these posts advertise a website, social account, or a YouTube video. These spam posts are considered irrelevant data because they do not yield important information. Hence, they have to be removed. To remove spam, this study has used a toolkit introduced by [10], which was built based on the Naïve Bayes (NB) classifier. The NB classifier is a probability-based model composed of joint probabilities of terms of related groups. This classifier is used to estimate the probabilities of categories found in a test document. The two-standard event models used for NB classification of text are the multivariate Bernoulli model and multinomial model [11]. The NB classifier is mostly used for tasks involving text category depending on the Bayes' formula:

$$P(c_i|d) = \frac{p(c_i)p(d|c_i)}{p(d)} \qquad (1)$$

where $p(c_i|d)$ represents the posterior probability of class, *Ci* represents a given new document *d*, and *P* (*Ci*) represents the probability of class *Ci* that may be calculated using:

$$p(c_i) = \frac{Ni}{n} \qquad (2)$$

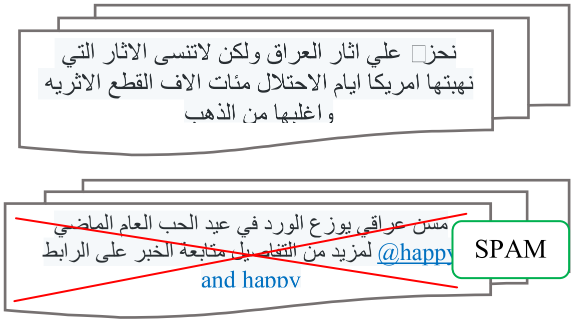
where *Ni* represents the number of documents allocated to class *Ci*, *N* represents the number of classes, $(d|c_i)$ represents the probability of document *d* being given a class *Ci*, and $p(d)$ is the probability of document *d*. Given the independence assumption of NB, the probability of document d may be determined by the following formula:

$$p(c_i|d) = p(c_i)\prod_{k=1}^{n}(t_k|c_i) \qquad (3)$$

where *ti* is a feature that co-occurs with class Ci, and $(t_k|c_i)$ may be determined by

$$p(t_k|c_i) = \frac{1+n_{ki}}{1+\sum_{h=1}^{l} n_{hi}} \qquad (4)$$

where *nhi* is the quantity of documents that contain feature *tk* and are part of class *Ci*, and *l* is the total quantity of unique features in all training documents that belong to class *Ci*. The example of this process is demonstrated in Figure 3.



نحز□ علي اثار العراق ولكن لاتنسى الاثار التي
نهبتها امريكا ايام الاحتلال مئات الاف القطع الاثريه
و اغلبها من الذهب

~~مسن عراقي يوزع الورد في عيد الحب العام الماضي~~
~~لمزيد من التفاصيل متابعة الخبر على الرابط @happy~~
~~and happy~~

Figure 3: Spam and non-Spam Example

3.3 Pre-processing

This phase aims to turn data into a sequence of tokens and eliminate irrelevant data. In this study, pre-processing has three tasks, namely, normalization, tokenization, and spam detection.

3.3.1 Normalization

This task aims to eliminate irrelevant data such as stop-words, digit, and punctuation. These data should be removed to obtain enhanced performance (e.g., @ ,3090 ,الى ,على , من,.)

3.3.2 Tokenization

This task aims to turn the data into a series of tokens to analyze the words and identify their boundaries. This task facilitates the processing of words by eliminating multiple spaces.

3.4 Natural Language Processing Tool

In this section, Arabic tools made by The Stanford Natural Language Processing Group and key phrases abstraction are used for comment analysis to define the topics of discussion and to attract keywords. This phase is one of the most important steps. This phase uses raw Arabic text from Twitter as a main input. In this step, which is part-of-speech (POS) tagging, Arabic named entity recognition and Arabic noun-phrase extraction system are used to recognize and classify noun phrases from Arabic tweets. Noun phrases are then used as topics or candidate keywords. At this stage, we also have to determine the current topics that people frequently refer to in their comments. To determine the named entity, it must be substituted with names of places, people, and institutions, as well as extracted key terms and phrases, e.g., “العراق,” “نوري المالكي,” “سلما□ن بن عبد العزيز,” “ماليزي,” “الملك عبد الله,” etc.) and other keywords. To accomplish our set goal, we must arrange them in a thread form. All of the comments on the subjects must be subsequently compiled. The compilation is then ascertained from the range of topics, which are the focus of contemporary people based on the significance of every topic and recurrent threading trend detection (see Figure 4).

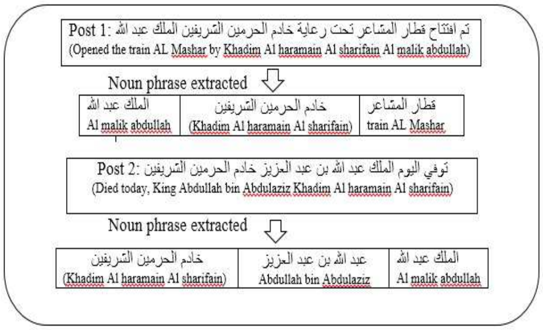

Post 1: تم افتتاح قطار المشاعر تحت رعاية خادم الحرمين الشريفين الملك عبد الله
(Opened the train AL Mashar by Khadim Al haramain Al sharifain Al malik abdullah)
Noun phrase extracted
الملك عبد الله
Al malik abdullah
خادم الحرمين الشريفين
(Khadim Al haramain Al sharifain)
قطار المشاعر
train AL Mashar
Post 2: توفي اليوم الملك عبد الله بن عبد العزيز خادم الحرمين الشريفين
(Died today, King Abdullah bin Abdulaziz Khadim Al haramain Al sharifain)
Noun phrase extracted
خادم الحرمين الشريفين
(Khadim Al haramain Al sharifain)
عبد الله بن عبد العزيز
Abdullah bin Abdulaziz
الملك عبد الله
Al malik abdullah

Figure 4: Keyword Extraction

3.5 Topic Identification

keywords as “entry points” for thread detection. Simply put, hints for new topic discussions are provided by keywords that show a bursty behavior. Additional key terms within the collected public posts may also be noted. In the context of this research, a term is a named entity, an extracted noun, and noun phrase from Arabic tweets if its length is from 1 up to n words belonging to the same sentences in the post. To weigh each term, the “Term Frequency–Inverse Document Frequency” (TF–IDF) is used. This approach was decided because of its simplicity [12].

Under this system, a weight is assigned to a term based on two measures, namely, the frequency of a term within a single document and the quantity of documents in the corpus that contain the given term. The entropy measure is used together with the TF–IDF in which the entropy term can be determined by gathering all tweets consisting of that term. The TF–IDF and entropy measures are effective selection criteria used in topic identification. Additionally, these measures obtain improved results and height accuracy in identifying trends and being real at the same time. Furthermore, the similarity of the IDF formulation to a component noun of entropy has motivated other researchers to establish connections in certain cases. However, these attempts differ from the suggested link. This evidence strongly justifies the use of IDF [13].

3.6 Terms Clustering

The post-topic identification process produces an ordered list of the most important terms in the corpus. Terms are compound nouns consisting of unigram, bigram, and trigram. The distribution of topics is influenced by topic detection. Therefore, accuracy is enhanced. The next step in trend identification is clustering terms that relate to the same topic. To accomplish this step, post clustering is performed. This process is composed of three steps: (1) term clustering by distribution, and co-occurrence, (2) k-means clustering, and (3) clustering by voting combination. The details of these steps are given below.

3.6.1 Term Clustering by distribution

Clustering by distribution is a combination of (1) the comparison of the term weight among various terms, and (2) the intersection of the related documents (e.g., Twitter public posts). The aim is to eliminate multiple occurrences of the parallel n-grams with different lengths from the same posts. Although the terms “الملك عبدالله” and “وفاةالملك عبدالله” are extracted from the same post, they are shown as different terms with the same weight and information, creating redundancy. Upon completing the grouping, the groups are substituted with the longest n-gram because it has the most information about the topic. Algorithm 1 provides a listing of the proposed technique [14].

```
Algorithm 1 : Distribution
for each term in sorted Weight List do
if (term Weight!= previous Term Weight) then
        create New Group(term);
    else
  for each group in topic Groups do
    if (get Posts(group) = get Posts(term)) then
        add Term To Group(term, group)
    else
            create New Group(term);
    end
  end
end
      weight= term Weight
end
```

The weights of all the terms in the list are evidently compared with the weights of the previous terms to validate their equality. If the weights are not equal, a new group for the term is subsequently created. If all the terms from the group are identical to the group, they are incorporated in that group. If that condition is not the true, they are not gathered in one group even if the weights of the terms are equal. Hence, they must not be incorporated in the same group in this step.

3.6.2 Term clustering by co-occurrence

The repeated occurrence of two terms from a text corpus next to each other is a novel approach when clustering co-occurrence data. This method presents an optimization issue in information theory in this framework. Optimal clustering has minimal loss of mutual information. Recently, a divisive clustering algorithm was proposed which monotonically reduces this loss function. After removing redundancy, the next step is to group related terms together through the co-occurrence approach based on the assumption that terms with the same frequency pertain to the same topic. This procedure is conducted to further group terms that are not semantically parallel and those that pertain to different posts but continue to refer to the same topic, such as “الملك عبدالله” and “عبد الله بن عبد العزيز.” The yielded list of topic groups is then ordered by the highest term weight in the group. Algorithm 2 is shown in the listing below [14].

Algorithm 2: Co-occurrence

```
for each term in sorted Weight List do
    for each group in topic Groups do
        if (get Posts(group)  get Posts(term)!= 0) then
                addTermToGroup(term, group)
            else
                createNewGroup(term);
        end
    end
end
```

Algorithm 2 depends on the intersection of the post sets associated with a particular term. Hence, this algorithm initiates from the first term in the list, which is ordered by weight. This algorithm attempts to identify a group in which the intersection of the posts is not a null or empty set. However, if the first group that meets this condition is acknowledged, the term is incorporated and the processing continues with the subsequent terms. These steps ultimately generate a list of topic groups arranged in descending order of term weights in the group. These topic groups are depicted as a set of terms and related posts for each included term.

3.6.3 K-means clustering

K-means is one of the partitioning clustering approaches that aim to divide data into similar groups based on a predefined number of clusters [15]. K-means begins by generating random centroids (i.e., initial solutions). Each point from the data will then be measured with each centroid and will be joined with the most similar one. Subsequently, the centroids are recalculated and the similarity between each point from the data and the new centroids are re-measured. These additional tasks continue until the results no longer change. The five steps of the k-means clustering algorithm form Algorithm 3, which is shown in the listing below.

Algorithm 3: K-means

1. Set the number of clusters k.
2. The first step is selecting a set of K instances as centres of the clusters.
3. The algorithm then examines each instance and puts it in the closest cluster.
4. The cluster centroids are recalculated either after the entire cycle of re-assignment or after each instance assignment.
5. Repeat This process until the centroids do not change any more.

3.6.4 Voting Combination Clustering

The voting rule operates by counting the predictions of component clusters and places test sample x in cluster *I* with the most component prediction equations.

$$o_j = \sum_{k=1}^{D} \max o_{kj} \qquad (4)$$

This approach is equivalent to averaging outputs over clusters (i.e., average rule) [16]. The voting combination in this study aims to apply co-occurrence and k-means clustering. The voting combination then selects the best results between the two. The seven steps of the voting-combination clustering algorithm form algorithm 4, which is shown in the listing below.

```
Algorithm 4: voting combination
1. For each topic in K-means and term clustering by distribution, and Co-occurrence list do.
2. Identify maximum recall value and store the result in cluster called voting combination.
3. Identify maximum precision value and store the result in cluster called voting combination.
4. Identify maximum f-measure value and store the result in cluster called voting combination.
5. End for.
6. Until the number of topic is reached 18 topics.
7. Repeat.
```

3.7 Evaluation Metrics

In this section, we inspect the metrics used to determine the performance of the algorithm. The evaluation framework is composed of extracting ground truth. A given collection of data will contain a large number of topics hidden within the data set. The topics are extracted manually. A ground truth topic is assigned to a time slot according to the time when the topic came out in the mainstream. Topic detection output is compared with ground truth. The experiment was assessed using precision, recall, and F-measure scores to compare with the trending topics identified by Twitter. All three measures determine the number of true positives, which are the items identified as trending topics by the experimental method and Twitter. To determine precision, the number of false positives should be calculated. False positives are the items identified as trending topics by the experimental method that are not shown as trending topics in Twitter.

Determining recall entails a calculation of the number of false negatives. These items are shown as trending topics in Twitter, but they are not shown as trending topics in the experimental method. The equation for F-measure calculation is shown in Figure 5. The topics detected automatically (e.g., the lists of keywords) are compared with the ground truth by using the following three metrics: Topic recall: Percentage of topics detected successfully by a method.

Keyword precision: Percentage of keywords that are correctly detected out of the total number of keywords for topics related to a specific ground-truth topic in the considered time slot. The method's total precision is calculated by micro-averaging individual precision scores over the period of all time slots.

Keyword recall: Percentage of keywords that are correctly detected over the total number of keywords of the ground truth topics that have been matched to a certain candidate topic in the considered time slot.

Mean precision: Number of system-identified correct topics over the total number of system-identified topics.

Mean recall: Number of system-identified correct topics over the total number of human-identified topics

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negaties}}$$

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Figure 5: F-measure calculation

4 Result and Discussion

In this phase, the trend detection model, which was built based on the clustering method, has been evaluated. Three clustering methods have been used, namely, term clustering by distribution and co-occurrence, k-means clustering, and voting combination clustering.

The evaluation concentrated on two aspects: trend distribution and effectiveness of retrieval. Trend distribution refers to the proper distribution of topics, which indicates that several trends have been presented previously without certain trends or duplicating certain trends. Therefore, the number of trends identifies the most accurate representation. Several numbers of trends have been used, ranging from 11 to 18 trends.

The reason for selecting such range is the poor distribution of trends. If less than 11 trends are used, numerous terms are considered as outliers if they do not match any of the topics. Selecting more than 18 trends may cause duplicate trends. Eventually, the effectiveness of retrieval has been evaluated using precision, recall, and f-measure. The following sub-sections describe the process of evaluation in detail.

4.1 Evaluation of Trend Detection Based on term clustering by distribution, and Co-occurrence

This phase aims to evaluate term clustering by distribution and co-occurrence method. Furthermore, evaluation was performed depending on the number of trends (i.e., clusters). Therefore, several numbers of trends have been used within the 11–18 range to seek the best results. Table 2 depicts the results of this phase.

Table 2: Results of Trend Detection Based on term clustering by distribution, and Co-occurrence

Num	Trends	Precision	Recall	F-measure
11	الملك عبدالله	0.896	0.685	0.776
12	ماليزيا	0.917	0.733	0.815
13	سلما□ بن عبدالعزيز	0.923	0.80	0.857
14	خادم الحرمين الشريفين	0.928	0.867	0.896
15	نوري المالكي	0.867	0.867	0.867
16	السعودية	0.867	0.812	0.839
17	داعش	0.768	0.862	0.812
18	الحوثي	0.722	0.867	0.788

Term clustering by distribution and co-occurrence has been evaluated based on the number of trends (see Table 2). The greatest results have been obtained when the number of trends is 14, which achieves 92%, 86%, and 89% precision, recall, and f-measure, respectively. Fourteen trends represent the mean where clusters are not limited (selecting a low number of clusters leads to limited trends). This number of clusters does not cause redundant clusters (selecting a large number of clusters leads to redundant clusters).

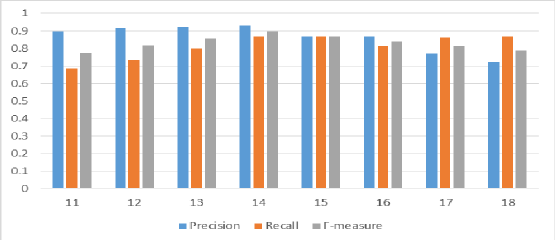


Figure 6: Term Clustering by Distribution, and Co-occurrence Result

4.2 Evaluation of Trend Detection Based on K-means Clustering

This phase aims to evaluate the k-means clustering method. Furthermore, evaluation was conducted depending on the number of trends (i.e., clusters). Therefore, several numbers of trends are within the 11–18 range to seek the best results. Table 3 depicts the results of this phase.

Table 3: Results of Trend Detection Based on K-means Clustering

Num	Trends	Precision	Recall	F-measure
11	السعودية	0.872	0.623	0.726
12	عبد العزيز	0.897	0.792	0.809
13	عبد الله	0.904	0.881	0.835
14	الحرس الثوري الايراني	0.917	0.873	0.889
15	العراق	0.836	0.798	0.809
16	الحوثي	0.857	0.764	0.797
17	داعش	0.805	0.746	0.768
18	العبادي	0.786	0.712	0.743

K-means clustering has been evaluated based on the number of trends (see Table 3). The greatest results were obtained when the number of trends was 14, achieving 91%, 87%, and 88% precision, recall, and f-measure, respectively. Fourteen trends represent the mean where clusters are not limited (selecting a low number of clusters leads to limited trends). This number does not cause redundant clusters (selecting a large number of clusters leads to redundant clusters).

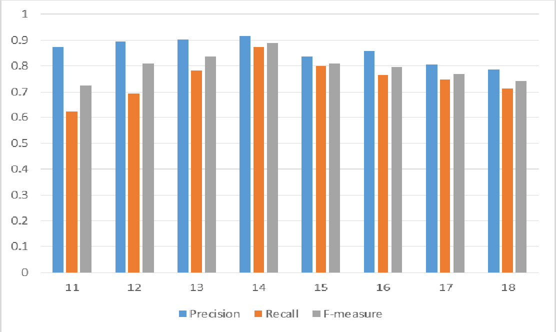


Figure 7: Results of k-means

4.3 Evaluation of Trend Detection based on Voting Combination

Voting combination constructs a comparison between k-means clustering and term clustering by distribution and co-occurrence in terms of trend detection. This approach includes the number of clusters from both methods. Voting selection for the best results is then performed. This phase aims to evaluate the voting combination clustering method. Evaluation was conducted depending on the number of trends (i.e., clusters). Therefore, several numbers of trends have been used within the 11–18 range to seek the best results. Table 4 depicts the results of this phase.

Table 4: Results of Trend Detection Based on Voting Combination

Num	Trends	Precision	Recall	Fmeasure
11	الملك عبدالله	0.906	0.720	0.811
12	ماليزيا	0.909	0.753	0.818
13	العراق	0.912	0.814	0.836
14	الحوثي	0.936	0.874	0.903
15	داعش	0.874	0.846	0.859
16	اليمن	0.891	0.836	0.859
17	سلما□ بن عبدالعزيز	0.824	0.817	0.814
18	خادم الحرمين الشريفين	0.796	0.754	0.769

The voting combination clustering has been evaluated based on the number of trends (see Table 4). The greatest results have been obtained when the number of topics was 14, achieving 93%, 87%, and 90% precision, recall, and f-measure, respectively. Fourteen trends represent the mean where clusters are not limited (selecting a low number of clusters leads to limited topics). This number of clusters does not cause redundant clusters (selecting a large number of clusters leads to redundant clusters).

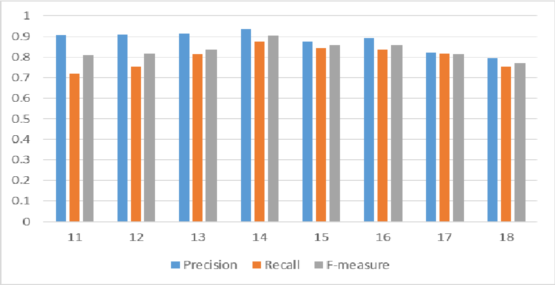


Figure 8: Results of voting combination clustering

5. LIST OF DETECTED TRENDS

Based on the culminated clusters on previous steps, we extracted several trends and calculated the precision, recall, and f-measure of each trend. To obtain enhanced results and high accuracy on trend detection, we compare the two types of clusters presented in Tables 2 and 3 using voting combination. The results of the voting combination represent a group of trends detected through this research (see Table 5). The reason for selecting such range is the poor distribution of trends. Selecting more than 18 trends may cause duplicate trends. Eventually, the effectiveness of retrieval has been evaluated using precision, recall, and f-measure. The calculated precision, recall, and f-measure of each trend are shown in Table 5.

Table 5: List of Trend Detection

Num	Trend detection	Tweets	Correct tweet	False tweet
1.	العراق	213	213	0
2.	اليمن	183	180	3
3.	الحوثي	83	80	3
4.	مصر	67	67	0
5.	الحرس الثوري الإيراني	59	55	4
6.	داعش	58	58	0
7.	العبادي	52	52	0
8.	السعودية	42	42	0
9.	نوري المالكي	39	39	0
10.	خادم الحرمين الشريفين	38	38	0
11.	بغداد	37	33	4
12.	سلما□ بن عبد العزيز	32	32	0
13.	ماليزيا	29	29	0
14.	الملك عبد الله	26	26	0
15.	رئيس الوزراء	25	20	5
16.	المحكمة الدستورية العليا	21	18	3
17.	مساعد الطيار	19	19	0
18.	عبد الله بن عبد العزيز	18	18	0

6. COMPARISON BETWEEN CLUSTERING METHODS

This phase concentrated on establishing a comparison among the three clustering methods, namely, term clustering by distribution and co-occurrence, k-means clustering, and voting combination clustering. The three clustering methods have obtained the best results when number of trend was 14. Therefore, the comparison will be held based on such number of trends. Table 6 presents the results of this phase.

Table 6: Comparison between clustering methods

Clustering method	Precision	Recall	F-measure
term clustering	0.928	0.867	0.896
K-means clustering	0.917	0.873	0.889
Voting combination	0.936	0.874	0.903

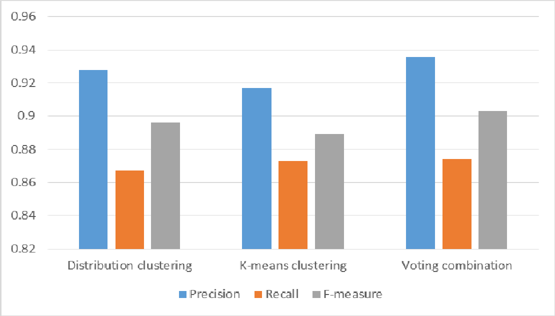


Figure 9: Comparison between Clustering Methods

The voting combination clustering outperformmed other clustering methods by obtaining 93%, 87%, and 90% precision, recall, and f-measure, respectively (see Table 6). Term clustering by distribution and co-occurrence outperformed k-means clustering by achieving 91%, 87%, and 88% precision, recall, and f-measure, respectively. The voting combination aims to select the maximum results of both of k-means and term clustering by distribution and co-occurrence methods, leading to its outperformance. The recall results for all clusters are approximately the same (see Figure 8) because recall is related to the true candidates that are retrieved by the clusters.

7. TREND CATEGORIZATION

A list of all terms serves as input for a given trend. To get the clusters of tweets related to a particular trend, we conduct a single pass of all public tweets and fetch a tweet if it contains any of the topical terms. Based on this adjustment, a categorization task was performed to classify the trends into three categories: (i) Arabic nationality events, (ii) personal events, and (iii) other events. Personal events indicate the trends related to the public community such as the death of Abdullah, the king of Saudi Arabia. Arabic nationality events refer to trends associated with specific operations, such as "explosion made by ISIS," "war on Yemen," and those that are related to the Middle East events. Finally, other events refer to trends that are more global such as "tourism in Malaysia." Therefore, the trend detection phase aims to identify the most trending categories.

The trend categorization phase aims to classify trends into three categories, namely, public trends, operational trends, and others. In this phase, the results of this process have been presented for each category (see Table 7).

Table 7: Trend Categorization

Trends	trend categorization	percentage
داعش العراق الحوثي مصر الحرس الثوري الإيرا□ الشرق الاوسط مدينة الموصل	Arabic nationality events	65 %
اسلما□ بن عبد العزيز الملك عبد الله نوري المالیکي	personal events	20 %
ماليزيا مساعد الطيار الطائرة الامانية	other events	15 %

"داعش" ISIS has recorded the most trending tweets in terms of Arabic nationality events, whereas the death of King Abdullah king Saudi Arabia has obtained the most trending tweets in terms of personal events (see Table 7). Finally, "ماليزيا" Malaysia has achieved the most trending tweets in terms of other events (see Table 7). According to the general sense of the trends, the results have been classified into three event categories (see Table 7). Based on these clusters, the observation of trend distribution indicated that all posts are categorized into three main classes: Arabic nationality events, personal events, and other events. The first category has the highest number of posts at 68% because of the surge of events in the Middle East. The second category accounts for 20% of posts.

Finally, the third category obtains 15% of posts. The events in this category may be positive or negative and receive feedback from the people. Thus, trend detection is important to companies, governments, national security agencies, and marketing groups because it can be used to identify issues that concern the public and develop strategies to solve them.

8. CONCLUSION

Twitter has been selected as the social media trending platform in this study. Research proves that practitioners use other social media platforms for brand-related communication. For example, channels can be created on Twitter platform to which Twitter consumers can subscribe and tune to their favorite brand conversations. Moreover, new social media platforms such as Google, which are at

par with Facebook and Twitter in terms of embracing trends, constantly emerge.
Therefore, the extension of this thesis to other platforms can provide possibilities to specify specific domains where a difference in performance exists. This extension is highly crucial for new companies that aim to join social media because it can aid them in selecting a suitable platform for their goals and needs. To identify the issues facing the public and develop strategies to rectify them, governments, national security agencies, journalists, companies, and marketing groups rely on controversial trend detection. The use of online social media activities has reached an unprecedented peak in the current age of information. A large number of users participate in social awareness streams, such as microblogging and SNs, to inject and share information in the network. Observing Arabic social media is the focus of this research. A system of trend detection is presented based on the features of the Arabic posts shared on social media. This study indicates that the voting combination clustering achieved 93%, 87%, and 90% precision, recall, and f-measure in trend detection, respectively, throughout the time spans. This result reveals the possibility of early controversy trend prediction in social media. Limited work has been conducted on controversial topic detection in Arabic Twitter or Wikipedia. To aid future research, an annotated corpus is formed in controversy trend detection as social media assistants.

9. FUTURE WORK

In this study, we identify several issues that require further investigation. These issues can be expanded to further develop a new methodology for trend detection in Arabic social media based on a hybrid of combined clustering algorithm and statistical topic extraction methods. One of the drawbacks of research on trend detection is the lack of well-annotated data sets. A new and manually annotated dataset that covers more trend categories is required to evaluate used or new methods for Arabic trend detection. Arabic NLP pre-processing tools are needed to improve accuracy and recall. These tools can be useful for reducing errors. The proposed methods can be extended by including advanced clustering algorithms and evaluate them for Arabic trend categorization. The coverage of the designed Arabic trend detection system can be expanded to cover more Arabic blogs, posts, comments, and tweets from blogsites, Facebook and Twitter. An online version of Arabic Twitter monitor can be designed to give audiences the chance to perform in-depth inspection of recent Arabic Twitter trends. Finally, numerous other methods, such as new statistical topic extraction methods and several topic and tweet clustering methods, are available for trend detection. Using evidence from multiple languages has obtained good results, which improves recall. Future work can focus on evaluating these methods and extending them for Arabic social media.

10. ACKNOWLEDGEMENT

This work has been supported by the University Research Grant DPP-2015-015 and GUP-2015-003.

REFERENCES

[1] Ellison, Nicole B. "Social network sites: Definition, history, and scholarship."Journal of Computer-Mediated Communication 13.1 (2007), pp. 210-230.

[2] Benhardus, James, and Jugal Kalita. Streaming trend detection in twitter. "International Journal of Web Based Communities 9.1 (2013), pp. 122-139.

[3] Alvanaki, Foteini, Sebastian Michel, Krithi Ramamritham, and Gerhard Weikum. "See what's enBlogue: real-time emergent topic identification in social media." In Proceedings of the 15th International Conference on Extending Database Technology, 2012, pp. 336-347.

[4] Mathioudakis, Michael, and Nick Koudas. "Twittermonitor: trend detection over the twitter stream." In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010, pp. 1155-1158.

[5] Yi, Jeonghee. "Detecting buzz from sequenced document streams." In e-Technology, e Commerce and e-Service, 2005. EEE'05. Proceedings. The 2005 IEEE International Conference on, 2005, pp. 347-352.

[6] Goorha, Saurabh, and Lyle Ungar. "Discovery of significant emerging trends." In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 57-64.

[7] Budak, Ceren, Divyakant Agrawal, and Amr El Abbadi. "Structural trend analysis for online social networks." Proceedings of the VLDB Endowment 4, no. 10 (2011), pp. 646-656.

[8] Gloor, Peter, Jonas Krauss, Stefan Nann, Kai Fischbach, and Detlef Schoder. "Web science 2.0: Identifying trends through semantic social

network analysis." In Computational Science and Engineering, 2009. CSE'09. International Conference on, vol. 4, 2009, pp. 215-222.

[9] Nguyen, Duc T., and Jai E. Jung. "Privacy-preserving discovery of topic-based events from social sensor signals: An experimental study on twitter." The Scientific World Journal 2014, pp. 5.

[10] Wang, Alex Hai. "Machine learning for the detection of spam in twitter networks." In e-Business and Telecommunications, Springer Berlin Heidelberg, 2012, pp. 319-333.

[11] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." In AAAI-98 workshop on learning for text categorization, vol. 752, 1998 pp. 41-48.

[12] Sparck Jones, Karen. "A statistical interpretation of term specificity and its application in retrieval." Journal of documentation 28, no. 1 (1972), pp. 11-21.

[13] Robertson, Stephen. "Understanding inverse document frequency: on theoretical arguments for IDF." Journal of documentation 60, no. 5 (2004), pp. 503-520.

[14] Cvijikj, Irena Pletikosa, and Florian Michahelles. "Monitoring trends on facebook." In Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on, 2011, pp. 895-902.

[15] Bock, Hans-Hermann. "Clustering methods: a history of k-means algorithms." In Selected contributions in data analysis and classification, Springer Berlin Heidelberg, 2007, pp. 161-172.