*Research Article*

# Exploiting Language Models to Classify Events from Twitter

## Duc-Thuan Vo,[1] Vo Thuan Hai,[2] and Cheol-Young Ock[1]

[1] *School of Electrical Engineering, University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan 680-749, Republic of Korea*
[2] *Soongsil University, 369 Sangdo-ro, Dongjak-gu, Seoul 156-743, Republic of Korea*

Correspondence should be addressed to Cheol-Young Ock; okcy@ulsan.ac.kr

Classifying events is challenging in Twitter because tweets texts have a large amount of temporal data with a lot of noise and various kinds of topics. In this paper, we propose a method to classify events from Twitter. We firstly find the distinguishing terms between tweets in events and measure their similarities with learning language models such as ConceptNet and a latent Dirichlet allocation method for selectional preferences (LDA-SP), which have been widely studied based on large text corpora within computational linguistic relations. The relationship of term words in tweets will be discovered by checking them under each model. We then proposed a method to compute the similarity between tweets based on tweets' features including common term words and relationships among their distinguishing term words. It will be explicit and convenient for applying to k-nearest neighbor techniques for classification. We carefully applied experiments on the Edinburgh Twitter Corpus to show that our method achieves competitive results for classifying events.

## 1. Introduction

Twitter (https://twitter.com/) is a social networking application that allows people to microblog about a broad range of topics. Users of Twitter post short text, called "tweets" (about 140 characters), on a variety of topics as news events and pop culture, to mundane daily events and spam. Recently, Twitter has grown over 200 million active users producing over 200 million tweets per day. Twitter is a popular microblogging and social networking service that presents many opportunities for researches in natural language processing (NLP) and machine learning [1–6]. Locke and Martin [5] and Liu et al. [4] train a classifier to recognized entities based on annotated Twitter data for Named Entity Recognition (NER). Some research has explored Part of Speech (PoS) tagging [3], geographical variation in language found on Twitter [2], modeling informal conversations [1], and also applying NLP techniques to help crisis workers with the flood of information following natural disasters [6]. Benson et al. [7] applied distant supervision to train a relation extractor to recognize artists and venues mentioned within tweets of users who list their location.

Classifying events in Twitter is a difficult task that focuses on the automatic identification and classification of various types of events in tweet texts. In Twitter, events are topics that often draw public attention, for example, football matches or natural disasters. Several approaches have been proposed to classify events for detection such as wave analysis [8, 9], topic model approach based on latent Dirichlet allocation [10], hierarchical Dirichlet processes [11], and text classification and clustering [12]. Kireyev et al. [8] explored the use of topics models for analysis of disaster-related Twitter data. Sakaki et al. [12] investigated the real-time interaction of events such as earthquakes in Twitter and proposed an algorithm to monitor tweets and to detect target events. However, existing approaches encounter failures from in either latent topics detection or analyzing terms relationships. Because topic model techniques [13–15] have only focused on how to list set of relevant words into a group (called topic) it is missed on analyzing relations between topics. Considering tweets have been discussed in two events shown in Table 1, we are easy to recognize that $T_1$ and $T_2$ are discussed in event 1 and $T_4$ and $T_5$ are discussed in event 2. However, if using topic models the system will group $T_1$, $T_2$, and $T_3$ in the same event category

TABLE 1: Some samples of discussed tweets in two events.

| Category | Tweets | Relatedness with event |
|---|---|---|
| Event 1 | $T_1$: Amy Winehouse has passed away aged 27. | Yes |
| | $T_2$: Amy Winehouse found dead at her home in North London. | Yes |
| | $T_3$: Nelson Mandela, who led the peaceful transition from white-only rule, has died aged 95. | No |
| Event 2 | $T_4$: plane crash kills majority of KHL team Lokomotiv. | Yes |
| | $T_5$: plane crash in Russia kills 36 or 37 assumed to be hockey player. | Yes |
| | $T_6$: plane crash, helicopter, was in Moscow with 2 dead. | No |

even $T_3$ does not belong to the event because set of relation words as <"*passed away*," "*dead*," "*died*"> in these tweets is in the same topic model. Likewise, $T_6$ will be grouped into event 2 with $T_4$ and $T_5$ together even if $T_6$ does not belong to this event because sets of relation words as <"*plane*," "*crash*," "*helicopter*">, <"*Russia*," "*KHL team*," "*Lokomotiv*," "*hockey*">, and <"*kills*", "*dead*"> in these tweets are within the same topic models, respectively. Due to limitations in using topic models, we therefore propose the method to exploit language models having relations reference to not only analyze topics but also analyze relatedness of event in tweets to overcome these problems.

In this paper, we investigate the use of generative and discriminate models for identifying the relationship of objects in tweets that describe one or more instance of a specified event type. We adapt language modeling approaches that capture how descriptions of event instances in text are likely to be generated. Our method will find the distinguishing term words between tweets and examining them with a series of relationships, extracted by language models such as Concept-Net [16] and LDA-SP [17]. These language models have been widely studied based on large text corpora within computational linguistic relations. Hence the relationship among distinguishing terms and common terms between tweets becomes clear to measure their similarity by examining them under each model. Measuring similarity between tweets is explicit and convenient to apply it in the classifier algorithms, such as SVM and k-nearest neighbor (kNN), to classify events in Twitter.

The rest of this paper is structured as follows. Section 2 presents related work that refers to research on event detection. In Section 3, we discuss exploiting language models. In addition, we present a method to calculate the similarity between tweets for event classification. In the next following section, experiments that are applied to the Edinburgh Twitter Corpus for event classification are presented and discussed. Section 5 ends with conclusions and future work.

## 2. Related Work

Several applications have detected events in Web to apply to weblogs [18–20], news stories [21, 22], or scientific journal collections [23]. Glance et al. [19] presented the application of data mining, information extraction, and NLP algorithms for event detection across a subset of approximately 100,000 weblogs. They implemented a trend searching system that provides a way to estimate the relative buzz of word of mouth for given topics over time. Nallapati et al. [22] attempted to capture the rich structure of events and their dependencies on a news topic through event models by recognizing events and their dependencies on event threading. Besides the standard word for based features, their approaches took into account novel features such as the temporal locality of stories for event recognition. Besides that, some researches [24–27] have analyzed social network to search or detect emergency events on the internet. Dai et al. [25] presented a cycle model to describe the internet spreading process of emergency events which applied the Tobit model by analyzing social psychological impacts. Hu et al. [27] analyzed historical attributes then combined with HowNet polarity and sentiment words on microblog which has network information transmission of social emergency events. And, they then provided the important guidance in the analysis of microblog information dissemination that has relatedness with social emergency events on internet. Meanwhile, Dai et al. [24] proposed a method to search the shortest paths of emergency events through IBF algorithm by analyzing social network.

Some research has focused on summarizing Twitter posts for detecting events [28–31]. Harabagiu and Hickl [28] focused on the summarization of microblog posts relating to complex world events. To summarize, they captured event structure information from tweets and user behavior information relevant to a topic. Takamura et al. [31] summarized Japanese Twitter posts on soccer games during the time when people provide comments and expressed opinions on the timeline of a game's progress. They represented user actions in terms of retweets, responses, and quoted tweets. In particular, Sharifi et al. [30] detected events in Twitter by summarizing trending topics using a collection of a large number of posts on a topic. They created summaries in various ways and evaluate those using metrics for automatic summary evaluation.

Recently, several approaches have been proposed to detect events from tweets using topic model approach [8, 10, 12]. Kireyev et al. [8] explored the use of topic models for the analysis of disaster-related Twitter data. Becker et al. [32] and Popescu et al. [33] investigate discovering clusters of related words or tweets which correspond to events in progress. Sakaki et al. [12] investigated the real-time interaction of events in Twitter such as earthquakes and propose an algorithm to monitor tweets and to detect a target event. Diao et al. [10] attempted to find topics with bursty patterns on microblogs; they proposed a topic model that simultaneously captures two observations such as posts published around the same time and posts published by the same user. However, existing approaches have still met with failure in either
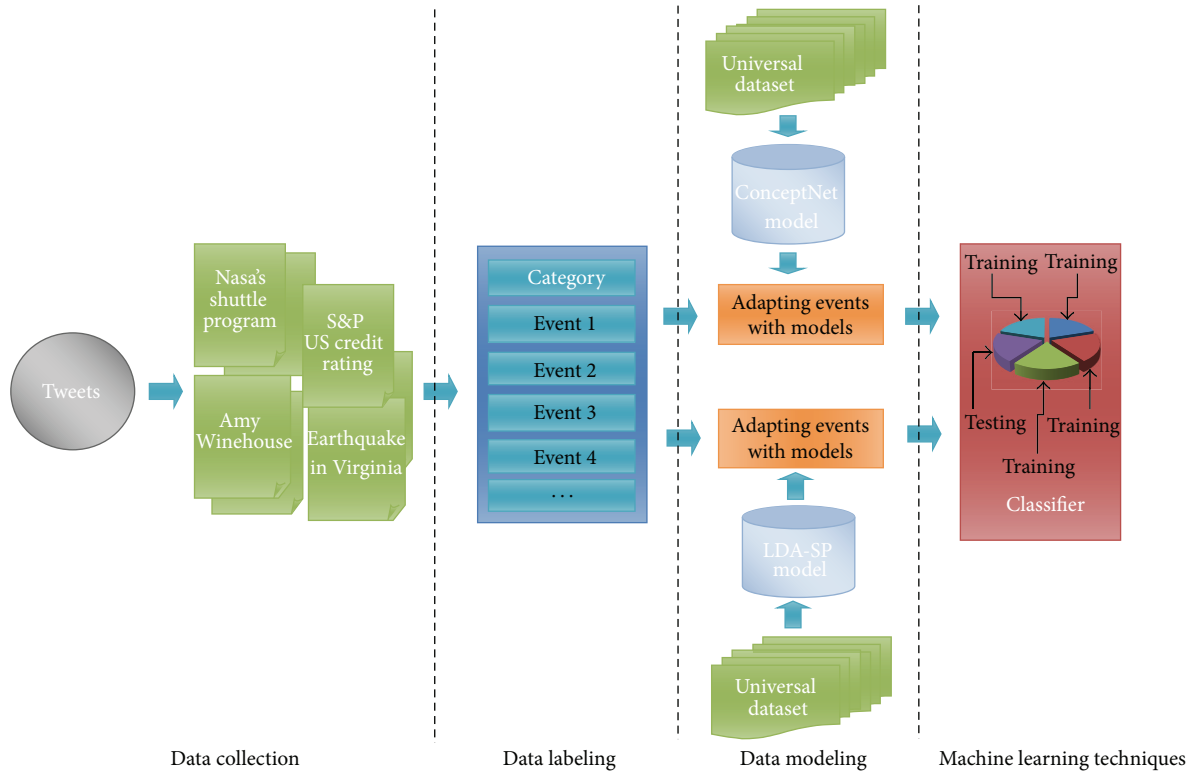
Figure 1: Proposed method.

latent topic detection or analyzing relationship terms, because tweets messages usually contain very limited common words in topics. Therefore, in this paper we propose a method to discover the relationship of objects in tweets by exploiting language models used to compare each of the snippets indirectly for classifying events in Twitter.

## 3. Exploiting Language Models to Classify Events

In this paper, we investigate the use of generative and discriminate models for identifying the relationship among objects in tweets that describe one or more instances of a specified event type. We adapt language modeling approaches that capture how descriptions of event instances in text are likely to be generated. We use language models to select plausible relationships between term words in tweets such as the relationship of "Object-Object" or "Object-relation-Object," which aim to detect the relatedness of an event in tweets. We assume that the data collection of language models contains suitable knowledge on the relationships among term words to discover the elemental relationship among tweets with a statistical analysis to classify events. We explore two types of language models that have obtained high correlation with human judgment such as ConceptNet and LDA-SP. These models are used for calculating the similarity of a pairwise of tweets for detecting events. The relationship between the discriminate term words of the tweets will be discovered by

checking their relatedness under pairs of relations. In addition, the similarity between tweets is computed based on their common term words and the relationship between their discriminate term words. It is intuitive and convenient to apply it in classifier algorithms to classify events in Twitter. The general proposed method consists of four stages as (1) data collection, (2) labeling stage, (3) data modeling, and (4) machine learning shown in Figure 1. Stages 1 and 2 will be discussed in Section 4.1; stage 3 will discussed in Section 3; and state 4 will be discussed in Sections 3.3 and 4.2.

*3.1. ConceptNet Model.* To model the "Object-Object" relationships in tweets, we consider the ConceptNet [16] model. It is a large semantic graph containing concepts and the relations between them. It includes everyday basic, cultural, and scientific knowledge, which has been automatically, extracted from the internet using predefined rules. In this work, we use the most current version ConceptNet 5. As it is mined from free text using rules, the database has uncontrolled vocabulary and contains many false/nonsense statements. ConceptNet contains 24 relations with over 11 million pairs of relation. For example, "Nasa is located in United States" is presented as AtLocation ("Nasa", "United States") in ConceptNet model. Table 2(a) shows list of 24 relations, and Table 2(b) shows samples of four relations as MadedOf, AtLocation, MotivedbyGoad, and RecievesAction. Speer and Havasi [16] provide more details of the model in their paper. We first examine all relations in the ConceptNet 5 database (http://conceptnet5.media.mit.edu/) and define which are relevant

TABLE 2: ConceptNet model. (a) List of relations. (b) Samples of extracted relations.

(a)

MotivatedByGoal; CausesDesire; WordNet/ParticipleOf; MemberOf; HasA; NotDesires; UsedFor; AtLocation; Entails; DefinedAs; InstanceOf; HasPainIntensity; ReceivesAction; SimilarTo; RelatedTo; NotHasProperty; PartOf; HasLastSubevent; TranslationOf; HasProperty; NotHasA; CapableOf; WordNet/adverbPertainsTo; NotCapableOf; LocationOfAction; SimilarSize; HasPainCharater; HasContext; NotMadeOf; HasFirstSubevent; SymbolOf; LocatedNear; NotUsedFor; ObstructedBy; Desires; DerivedFrom; HasSubevent; MadeOf; Antonym; CreatedBy; Attribute; DesireOf; IsA; Causes

(b)

| MadeOf | | AtLocation | | MotivatedByGoal | | ReceivesAction | |
|---|---|---|---|---|---|---|---|
| Atomic bomb | Uranium | Nasa | United states | Fight war | Freedom | Bacteria | Kill |
| Computer | Silicon | Alcoa | Pittsburgh | Get drunk | Forget life | Army tank | Warfare |
| Gas | Oil | Tv channel | Russia | Pen | Write letter | Bread | Cook |
| Song | Music | Aozora bank | Japan | Join army | Defend country | Candle | Burn for light |
| Person | Live cell | Apartheid | Mall | Kill | Hate someone | Tomato | Squash |
| Light | Energy | Golden gate | Bridge | Live life | Pleasure | Tobacco | Chew |
| Carton | Wax paper | Art | Gallery | Sing | Performance | Supply | Store |
| Chocolate | Cocoa bean | Audience | Theatre | Socialize | Be popular | Ruby | Polish |
| Telephone | Electronics | Crab | Coastal area | Study | Concentrate | Money | Loan |
| Window | Glass | Handgun | Army | Visit museum | See history | Life | Save |
| … | … | … | … | … | … | … | … |

to relations in target events by keywords matching (in experiments) to extract relations.

*3.2. LDA-SP Model.* To model the "Object-relation-Object" relationships in tweets, we adapt the LDA-SP model [17], which has been used for the selectional preference task in order to obtain the conditional probabilities of two objects in a relation. In particular, the LDA-SP, using LinkLDA [34], is an extension of latent Dirichlet allocation (LDA) [13] which simultaneously models two sets of distributions for each topic. The generative graphical model of LDA versus LDA-SP is depicted in Figure 2. In LDA-SP, they presented a series of topic models, at which objects belonged to them, for the task of computing selectional preferences. These models vary in terms of independence between $Topic_i$ and $Topic_j$ that is assumed. These two sets represent the two arguments for the relation $R$ $(Topic_i, Topic_j)$. Each topic contains a list of relation words. Each relation, $R$, is generated by picking up over the same distribution, which keeps two different topics, $Topic_i$ and $Topic_j$, sharing the same relation (Figure 2(b)). The LDA-SP is able to capture information about the pairs of topics that commonly cooccur. To model the relations with LDA-SP, we also follow the data preparation in [21], which was automatically extracted by TextRunner [35] from 500 million Web pages. This resulted in a vocabulary of about 32,000 noun phrases, a set of about 2.4 million tuples with 601 topics in our generalization corpus. Some samples of topics extracted through LDA-SP are illustrated in Table 3.

*3.3. Similarity Measures in Tweets.* Classifying events in tweets from Twitter is a very challenging task because a very few words cooccur in tweets. Intuitively, the problem can be solved by exploring the relationships between tweets well; the intrinsic relationship among words may be discovered with a thesaurus. Hence, we present a method to discover the intrinsic relationships between objects based on statistical analysis of language models and then gain the similarity between tweets accordingly. We consider two types of relationships in tweets such as "Object-Object" and "Object-relation-Object."

*"Object-Object".* The event "*Death of Amy Winehouse*" is posted in tweets $T_1$, $T_2$, and $T_3$ shown in Figure 3. Traditional methods can only find one cooccurring term, "*Amy Winehouse*," in the tweets after removing stop words. However, if we analyze and compare the relatedness between the pairs <"*Singer*"-"*Amy Winehouse*">, <"*Amy Winehouse*"-"*passed away*"> and <"*Amy Winehouse*"-"*dead*">, and <"*Amy Winehouse*"-"*R.I.P.*">, closer relationships will be exposed: "*Object-Object*" as "*Topic_1-Topic_2*" where a set of terms {"*Singer*"; "*Amy Winehouse*"} is in $Topic_1$ and a set of terms {"*death*", "*passed away*", "*R.I.P.*"} is in $Topic_2$.

*"Object-Relation-Object".* The event "plane carrying Russian hockey team Lokomotiv crashes" is posted in $T_4$, $T_5$, and $T_6$ shown in Figure 4. We can discover the relationship between "Object-relation-Object" such as <"*Plane*"-"*crash*"-"*KHL team Lokomotiv*">, <"*Plane*"-"*crash*"-"*Russia*">, and <"*Plan*"-"*crash*"-"*KHL team*">. This also exhibits the closer relationships "*Object-relation-Object*" as "*Topic_3-crash-Topic_4*" where the term {"*plane*"} belongs to $Topic_3$ and a set of terms {"*russia*", "*khl team lokomotiv*", "*hockey*", "*khl team*"} belongs to $Topic_4$.

Our method extracts relation tuples from language models such as ConceptNet and LDA-SP. We treat all tweets from Twitter that are contained in the collection equally and then perform to match models of tuples generated from ConceptNet and LDA-SP with them. Hence, if we can discover relation tuples as "third-party" for both tweets and calculate the similarity between the two tweets by comparing the
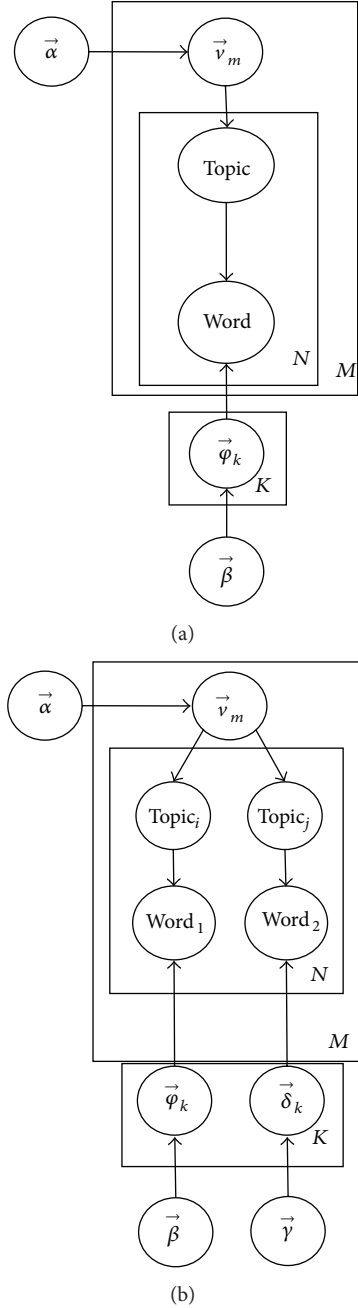
FIGURE 2: Graphical model of LDA model (a) versus LDA-SP (b).

distinguishing term words with these tuples, we may find the real relationship underlying the two tweets. We assume that the data collection language models contain sufficient knowledge about the relationships among term words, from which we can find the elemental relationship among tweets.

For computing similarity between tweets, we derive a set of relations, $R_i = (object_m, object_n)$ matched from language models and tweets combining with Bag-of-Words. Considering two original tweets, $d_1$ and $d_2$, in data collection $D$, we check with $object_m$, $object_n$ existing in each tweet which match with relation tuples $R_i = (object_m, object_n)$ extracted

from ConceptNet model. In using LDA-SP, we exam not only relations but also $object_m$, $object_n$ existing in each tweet and then match them with relation tuples $R_i = (object_m, object_n)$ generated from LDA-SP. We then replace matched objects in tweets by relation tuples from language models. Thus, the relationship between the distinguishing terms of the tweets can be discovered by examining their relatedness under pairs of relations by "third-party." We consider calculating the similarity between two tweets based on their common terms and the relationship between their distinguishing terms. To calculate the similarity between two tweets in an event category, we represent them as vectors:

$$d_1 = (w_1, w_2, \ldots, w_n)$$
$$d_2 = (w_1, w_2, \ldots, w_n), \tag{1}$$

where $w_i$ is the weight of the $i$th feature in the vector of $d_j$ and is defined by the tf-$i$df measure as follows:

$$w_i = \text{tf}_{ij} \times \log_2 \left( \frac{M}{\text{df}_j} \right), \tag{2}$$

where $M$ is the total number of documents in the collection, $\text{df}_j$ is the document frequency, that is, the number of documents in which term $w_i$ occurs, $\text{tf}_{ij}$ is the term frequency of term $w_i$ in document $d_j$, and $\text{tf}_{ij}$ is simply the number of occurrences of term $w_i$ in document $d_j$.

With the relationship between the two distinguishing term words on a diversity of assigned model tuples, we can calculate the similarity of vectors $d_1$ and $d_2$ with the cosine method shown in

$$\text{sim}(d_1, d_2) = \cos(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| |d_2|} = \frac{\sum_{i=n} w_{i1} w_{i2}}{\sqrt{\sum_{i=n} w_{i1}^2} \sqrt{\sum_{i=n} w_{i2}^2}}. \tag{3}$$

For classifying events from tweets, many classifiers first need to calculate the similarity between tweets. $k$NN is one of the best methods of similarity calculation and selection of a proper number of neighbors. Therefore, it is intuitive and convenient to apply similarity calculation between tweets to $k$NN for classifying events. If our proposed method can calculate the similarity among tweets more accuracy, the $k$NN will select more appropriate neighbors for a test case and the classification performance of $k$NN will be higher than original tf-$i$df, since the performance of $k$NN based on the similarity measuring method outperforms other methods with tf-$i$df measure. We conclude that the proposed method is more effective on calculating tweets similarity to classify events. The result will be discussed in more detail in experimentation section.

## 4. Experimentation

*4.1. Experimental Datasets and Evaluation Measures.* We have conducted experiments on the Edinburgh Twitter Corpus [36], a collection of events in Twitter, for event classification. The corpus contains 3034 tweet IDs spread into 27 event

TABLE 3: LDA-SP model. (a) Samples of list topics; (b) sample of list relations.

(a)

**Topic 2:** driver; civilians; soldiers; motorcyclist; teenager; thomas; policeman; us soldier; smith; george; motoris; father; . . .

**Topic 19:** china; business line; japan; government; israel; judge; india; iran; russia; democrats; court; lawmakers; . . .

**Topic 106:** myspacetv videos; the trail; leaves; moses; the curtain; stones; santa; flowers; victim; posters; stars; flames; . . .

**Topic 114:** britney spears; paris hilton; angelina jolie; tom cruise; the actress; lindsay lohan; amy winehouse; singer; . . .

**Topic 116:** fire; violence; the war; the storm; fighting; katrina; explosion; tornado; earthquake; civil war; dead; heaven; . . .

**Topic 171:** john; david; mark; mike; steve; bill; michael; peter; scott; smith; johnson; brown; executive; robert; jeff; brian; . . .

**Topic 251:** police; group; team; company; day; year; case; report; miller; officials; king; wilson; story; news; friday; . . .

**Topic 286:** article; report; author; court; bible; story; letter; paul; reuters; researchers; statement; respondents; . . .

**Topic 390:** car; train; vehicle; bus; fingers; truck; boat; plane; river; route; traffic; driver; aircraft; train; track; bike; . . .

**Topic 428:** airplane; aircraft; pilot; sparks; birds; crew; terrorists; nasa; people; passengers; the captain; bullets; the jet; . . .

**Topic 433:** family; couple; mary; sarah; thomas; elizabeth; margaret; jesus; jane; matt; martin; daniel; frank; anna; nancy; . . .

**Topic 454:** game; operation; experiment; treatment; procedure; scenario; victim; exercise: measurement; error; idea; . . .

**Topic 525:** bush; the president; president bush; jesus; paul; the minister; clinton; smith; george w. bush; obama; . . .

**Topic 561:** the world; christians; mulims; no matter; americans; jews; catholics; normoms; the chinese; hindus; . . .

**Topic 570:** the sun; light; the moon; the beam; earth; mars; venus; laser; darkness; stars; jupiter; a hush; radiation; . . .

(b)

| Relations | Relationship of topics (Topic$_i$-Topic$_j$) |
|---|---|
| Be cite | 525–561; 251–286; 286–251; 251–251; 542–251; 371–286; 542–371; 542–286; 251–162; 134–286; 162–286; 371–251; 286–162; 286–171; 542–454; 286–538; 454–286; 286–10; 134–24; 538–286; 285-286; 575–454; 572–286; 328–286; 19–454; . . . |
| Blame on | 116–428; 329–531; 116–531; 329-329; 329–116; 116–584; 329–584; 584–531; 314–531; 116–329; 480–531; 171–116; 116–160; 239–584; 458–531; 404–531; 584–116; 196–116; 531–458; 584-584; 531–116; 196–531; 176–531; 545–147; 171–2; . . . |
| Crash into | 428–287; 428–571 390–106; 428–139; 428–390; 428–428; 390–139; 390-390; 390–287; 390–428; 428–570; 390–570; 139–106; 139–428; 139-139; 428–328; 287–106; 139–390; 390–328; 139–287; 428–374; 390–374; 287–139; 570–287; 106–428; . . . |
| Spot in | 114–433; 433-433; 116–525; 114–287; 287–433; 114–570; 405–433; 433–405; 251–433; 114-114; 223–433; 570–433; 433–570; 114–132; 287–405; 114–251; 543–433; 230–433; 223–570; 114–424; 433–287; 433–114; 570-570; 433–132; 223–279; . . . |

categories. Currently, some tweets in the dataset are deleted or lost from Twitter. We developed a tool using Twitter API (http://twitter4j.org) to collected documents including tweets, retweets, responses, and quoted tweets; we then filtered documents to guarantee that each event category contains at least 70 tweets. After the removal of noise and stop words, each word is stemmed into its root form. Table 4 shows the rest of nine significant event categories with checked mark for experiments as event 1, event 6, event 7, event 9, event 13, event 14, event 15, event 16, and event 21.

In this study, experiments are evaluated based on the precision, recall, and $F$-measure with our proposed method. The precision, recall and $F$-Measure are the evaluation metrics often used to rate the information retrieval system's performance. Precision is the number of correct results divided by the total number of returned responses; recall is the number of correct results divided by the number of results that should have been returned and $F$-measure is used to balance between the recall and precision as follows:

$$\text{Precision} = \frac{number\_of\_correct\_responses}{number\_of\_responses},$$

$$\text{Recall} = \frac{number\_of\_correct\_responses}{number\_of\_corrects},$$

$$F\text{-measure} = \frac{2 \times recall \times precision}{recall + precision}.$$

(4)

*4.2. Experiments and Comparison.* Checking similarity between tweets before experiments, we select some samples of tweets from experimental datasets as shown in Table 1. We used the tf-*i*df combined with the similarity functions to compare performance before and after using language models. Note that $T_1$ and $T_2$ were discussed in the same event; $T_4$ and $T_5$ were also discussed in the same event. And two pairs of tweets are, respectively, to calculate similarity with stop words removal. The result depicted in Table 5 shows that the tweets using ConceptNet and LDA-SP increase the similarity of questions from the same category. Moreover, if the tweets did not belong to target event like $T_3$ and $T_6$, the method will reduce the similarity measure that helps system performance of classifying efficiently.
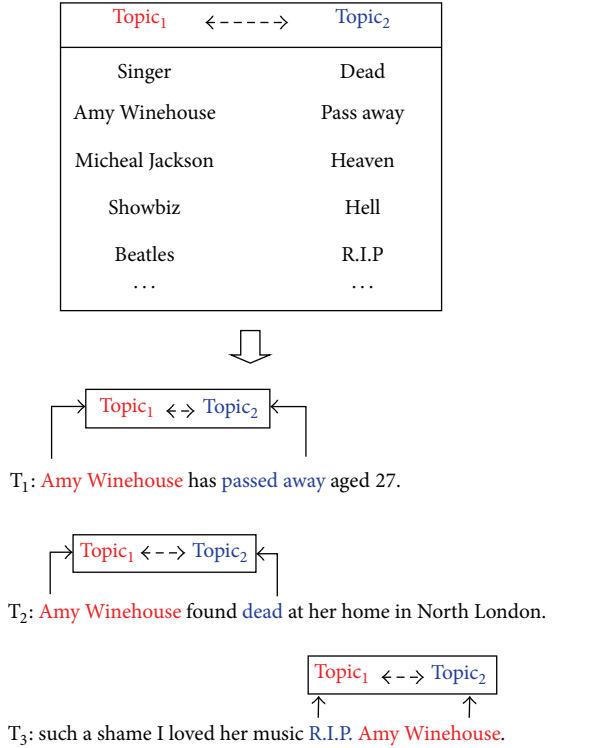
Figure 3: Relationship "*Topic$_1$-Topic$_2$*" in tweets of event "Death of Amy Winehouse."



Figure 4: Relationship "*Topic$_3$-relation-Topic$_4$*" in tweets of event "plane carrying Russian hockey team Lokomotiv crashes."

To classify events, 70% of the tweets for each category are randomly selected for training, and the rest is for testing. In our experiments, we compare the performance of four classifiers implemented as follows: (1) baseline $k$NN (without language model); (2) baseline SVM; and the $k$NN method combining our proposed methods (3) $k$NN-M1 ($k$NN with language model ConceptNet) and (4) $k$NN-M2 ($k$NN with language model LDA-SP). The SVM is also constructed using the tf-$i$df method to weight each vector component of the tweet and is used as second baseline for comparison with our proposed methods. We chose SVM because of a powerful and robust method for text classification [37–39]. The evaluation follows 5-fold cross validation schema. Table 6 shows the performance results applied to 7 categories of events from Twitter. The bold numbers show the best $F$-measure of each event in four methods. For instance, the system obtained the highest $F$-measure of 85.3% in event 1 with method $k$NN-M2. Method $k$NN-M1 yielded better $F$-measure results in most of the event categories: event 6, event 7, event 9, event 14, event 15, and event 16. And, method $k$NN-M2 achieved better $F$-measure result in three categories: event 1, event 13, and event 21.

The overall performance comparison is presented in Figure 5. We can see that the performance of $k$NN-M1 outperforms $k$NN-M2, SVM, and $k$NN. Both of our proposed methods are also higher than the baselines, $k$NN and SVM, in most of the performance metrics. In the overall results, $k$NN-M1, $k$NN-M2, SVM, and $k$NN obtained an $F$-measure of 85%, 84.7%, 78.4%, and 76.8%, respectively.
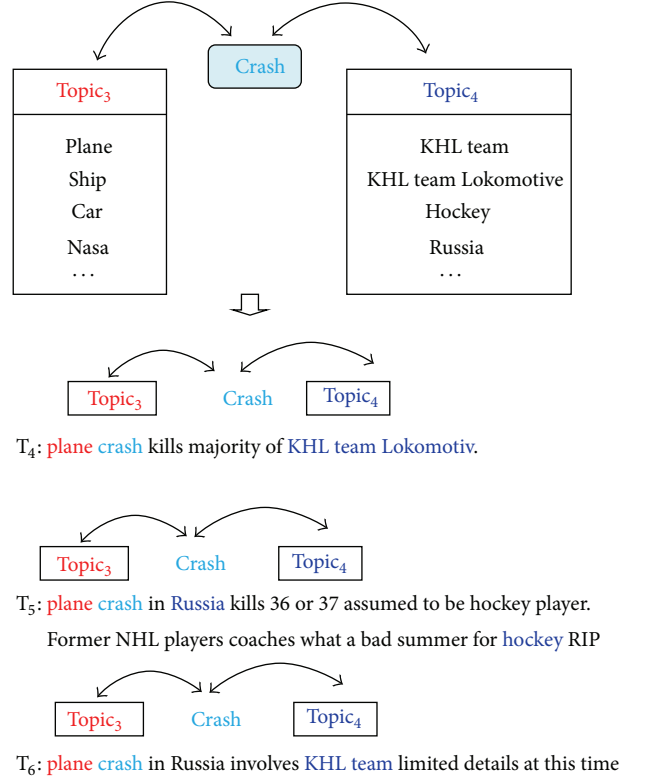
*4.3. Discussions.* We believe that effective performance of our proposed methods is result of the following reasons.

First, noise and exclamative and repeated texts usually occur in the tweets of each event. The following are examples of such tweets. T$_1$: "*Sad day Sky sources now confirming Amy Winehouse is dead A musical legend who died way too young in my opinion,*" T$_2$: "*Amy Winehouse found dead in her London flat according to sky news,*" and T$_3$: "*Hmm. . .omg. . .gruuu Amy Winehouse is dead not totally surprised though ohhh.*" We can observe that {"*Amy Winehouse*"; "*dead*"} is repeated text, {"*gruuu*"; "*ohhh*"} is noise text, and {"*Hmm*"; "*omg*"} is exclamative text. The repeated text will result in a positive value in the similarity measure; however, noise and exclamative texts will result in a negative value in the similarity measure. For preprocessing, stop words had been removed by a defined list of stop words automatically. However, we had checked and revised noise texts manually if they do not belong to list of stop words. For example, a lot of words "deaddddd" will be revised into "dead," or {"RIP", "R I P"} will be revised into "R.I.P."

The second reason we believe our method had effective performance is that quality universal datasets are used to build language models. In this study, more than five billion relation records extracted from Concept are used to build the models. In addition, models from LDA-SP are built by extracting 2.4 million tuples of relations and 601 topics. Furthermore, ConceptNet is a graphical relationship model which uses predefined rules. However, LDA-SP still has some

TABLE 4: Experimental datasets.

| Category | Description | Number of tweets | Checked |
|---|---|---|---|
| Event 1 | Death of Amy Winehouse | 774 | ✓ |
| Event 2 | Space shuttle Atlantis lands safely, ending NASA's space shuttle program | 45 | |
| Event 3 | Betty Ford dies | 8 | |
| Event 4 | Richard Bowes, victim of London riots, dies in hospital | 27 | |
| Event 5 | Flight Noar Linhas Aereas 4896 crashes, all 16 passengers dead | 9 | |
| Event 6 | S&P downgrades US credit rating | 275 | ✓ |
| Event 7 | US increases debt ceiling | 73 | ✓ |
| Event 8 | Terrorist attack in Delhi | 40 | |
| Event 9 | Earthquake in Virginia | 271 | ✓ |
| Event 10 | Trevor Ellis (first victim of London riots) dies | 63 | |
| Event 11 | Goran Hadzic, Yugoslavian war criminal, arrested | 2 | |
| Event 12 | India and Bangladesh sign a peace pact | 3 | |
| Event 13 | Plane carrying Russian hockey team Lokomotiv crashes, 44 dead | 225 | ✓ |
| Event 14 | Explosion in French nuclear power plant Marcoule | 137 | ✓ |
| Event 15 | NASA announces discovery of water on Mars | 110 | ✓ |
| Event 16 | Google announces plans to buy Motorola Mobility | 130 | ✓ |
| Event 17 | Car bomb explodes in Oslo, Norway | 21 | |
| Event 18 | Gunman opens fire in children's camp on Utoya island, Norway | 28 | |
| Event 19 | First artificial organ transplant | 16 | |
| Event 20 | Petrol pipeline explosion in Kenya | 27 | |
| Event 21 | Famine declared in Somalia | 71 | ✓ |
| Event 22 | South Sudan declares independence | 26 | |
| Event 23 | South Sudan becomes a UN member state | 7 | |
| Event 24 | Three men die in riots in Birmingham | 12 | |
| Event 25 | Riots break out in Tottenham | 19 | |
| Event 26 | Rebels capture Tripoli international airport, Libya | 4 | |
| Event 27 | Ferry sinks in Zanzibar, around 200 dead | 21 | |

TABLE 5: Sample of similarities calculated by the proposed methods and the tf-$i$df method.

| Tweets | tf-$i$df | tf-$i$df + ConceptNet | tf-$i$df + LDA-SP |
|---|---|---|---|
| $T_1$: Amy Winehouse has passed away aged 27.<br>$T_2$: Amy Winehouse found death at her home in North London. | 0.16 | 0.365 | 0.4 |
| $T_1$: Amy Winehouse has passed away aged 27.<br>$T_3$: Nelson Mandela, who led the peaceful transition from white-only rule, has died aged 95. | 0.123 | 0.078 | 0.084 |
| $T_2$: Amy Winehouse found death at her home in North London.<br>$T_4$: plane crash kills majority of KHL team Lokomotiv. | 0 | 0 | 0 |
| $T_4$: plane crash kills majority of KHL team Lokomotiv.<br>$T_5$: plane crash in Russia kills 36 or 37 assumed to be hockey player. | 0.433 | 0.452 | 0.468 |
| $T_5$: plane crash in Russia kills 36 or 37 assumed to be hockey player.<br>$T_6$: plane crash, helicopter, was in Moscow with 2 dead. | 0.272 | 0.146 | 0.104 |

errors [17] in computing word statistics. In the experiment results, performance of ConceptNet is better than LDA-SP.

The third reason believed to be behind our method's effective performance is that the models extracted from LDA-SP are intensely analyzed compared to ConceptNet for relationship. However ConceptNet obtained better performance results. Texts from tweets are incomplete sentences that result in failures in grammar parsing for analyzing relation. We did not include grammar parsing for analyzing tweets based on LDA-SP model. Therefore, ConceptNet exhibits a better performance for classifying events from Twitter than LDA-SP.

TABLE 6: Experimental results.

| Category | kNN | | | SVM | | | kNN-M1 (ours) | | | kNN-M2 (ours) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F | P | R | F |
| Event 1 | 76.3 | 71.6 | 73.8 | 75.2 | 75.5 | 75.3 | 86.1 | 77.5 | 81.6 | 88.2 | 82.6 | **85.3** |
| Event 6 | 84.6 | 85.4 | 84.9 | 86.9 | 87.2 | 87.1 | 91.1 | 89.4 | **90.2** | 89.1 | 86.4 | 87.7 |
| Event 7 | 78.9 | 72.3 | 75.5 | 80.4 | 76.2 | 78.2 | 87.5 | 82.3 | **84.8** | 82.4 | 78.9 | 80.6 |
| Event 9 | 83.9 | 78.8 | 81.3 | 85.5 | 80.2 | 82.3 | 93.8 | 92.9 | **93.4** | 87.2 | 83.3 | 85.2 |
| Event 13 | 83.6 | 72.4 | 77.5 | 82.8 | 75.6 | 79.1 | 86.2 | 80.5 | 83.3 | 87.3 | 82.6 | **84.9** |
| Event 14 | 70.1 | 67.8 | 68.9 | 71.6 | 70.0 | 70.8 | 85.2 | 78.7 | **81.8** | 83.8 | 74.3 | 78.8 |
| Event 15 | 79.3 | 71.5 | 75.2 | 81.0 | 70.8 | 75.6 | 90.1 | 87.9 | **88.9** | 88.8 | 85.8 | 87.3 |
| Event 16 | 80.5 | 72.4 | 76.2 | 82.5 | 73.1 | 77.5 | 85.7 | 80.0 | **82.8** | 85.5 | 79.6 | 82.5 |
| Event 21 | 81.6 | 74.1 | 77.7 | 82.4 | 76.8 | 79.5 | 83.9 | 77.8 | 80.7 | 85.4 | 77.1 | **81.0** |
| Overall | 79.5 | 74.4 | 76.8 | 79.9 | 77.0 | 78.4 | 87.9 | 82.4 | **85.0** | 87.4 | 82.3 | 84.7 |
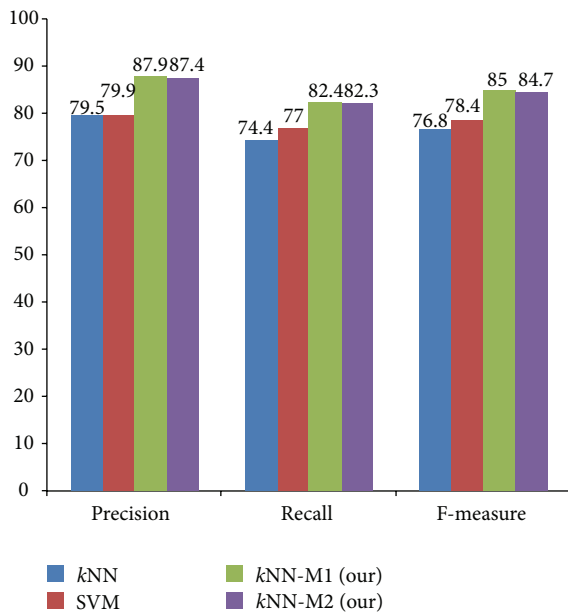


FIGURE 5: Overall performance comparisons.

## 5. Conclusion and Future Work

We have presented methods to classify events from Twitter. We first find the distinguishing terms between tweets in events and calculate their similarity with learning language models: LDA-SP and ConceptNet. Next, we discover the relationship between the distinguishing terms of the tweets by examining them under each model. Then, we calculate the similarity between two tweets based on their common terms and the relationship between their distinguishing terms. The outcomes make it convenient to apply kNN techniques to classify events in Twitter. As a result, our approach obtained better performance results with both ConceptNet and LDA-SP than other methods.

Regarding future work, the research has been suggested with attractive aspects to improve as follows. First, this approach can be considered for future work, including it with a larger corpus and experimenting with other event types.

Second, we will continue to investigate how to apply grammar parsing in tweets so that we can analyze deeply relationships to serve for classifying events. Finally, the research can be applied unsupervised learning with semantic similarity models as pointwise mutual information (PMI) [40, 41] and latent semantic analysis (LSA) [42, 43].

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

## References

[1] N. Dalvi, R. Kumar, and B. Pang, "Object matching in tweets with spatial models," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM '12)*, pp. 43–52, February 2012.

[2] J. Eisenstein, N. A. Smith, and E. P. Xing, "Discovering sociolinguistic associations with structured sparsity," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11)*, 2011.

[3] K. Gimpel, N. Schneider, B. O'Connor et al., "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL '11)*, pp. 42–47, June 2011.

[4] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11)*, pp. 359–367, June 2011.

[5] B. Locke and J. Martin, *Named entity recognition: adapting to microblogging [Senior thesis]*, University of Colorado, 2009.

[6] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami, "Safety information mining—what can NLP do in a disaster," in

*Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP '11)*, pp. 965–973, 2011.

[7] E. Benson, A. Haghighi, and R. Barzilay, "Event discovery in social media feeds," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11)*, pp. 389–398, Portland, Ore, USA, June 2011.

[8] K. Kireyev, L. Palen, and K. Anderson, "Applications of topics models to analysis of disaster-related twitter data," in *Proceedings of the NIPS Workshop on Applications for Topic Models: Text and Beyond*, December 2009.

[9] J. Weng and B. S. Lee, "Event detection in twitter," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, Barcelona, Spain, July 2011.

[10] Q. Diao, J. Jiang, F. Zhu, and P.-E. Lim, "Finding bursty topics from microblogs," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers (ACL'12)*, pp. 536–544, Jeju, South Korea, 2012.

[11] Z. J. Gao, Y. Song, S. Liu et al., "Tracking and connecting topics via incremental hierarchical Dirichlet processes," in *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM '11)*, pp. 1056–1061, Vancouver, Canada, December 2011.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pp. 851–860, New York, NY, USA, April 2010.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[14] X.-H. Phan, C.-T. Nguyen, D.-T. Le, L.-M. Nguyen, S. Horiguchi, and Q.-T. Ha, "A hidden topic-based framework toward building applications with short web documents," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 961–976, 2011.

[15] D. T. Vo and C. Y. Ock, "Extraction of semantic relation based on feature vector from Wikipedia," in *PRICAI 2012: Trends in Artificial Intelligence: Proceedings of the 12th Pacific Rim International Conference on Artificial Intelligence, Kuching, Malaysia, September 3–7, 2012*, vol. 7458 of *Lecture Notes in Computer Science*, pp. 814–819, Springer, Berlin, Germany, 2012.

[16] R. Speer and C. Havasi, "Conceptnet 5: a large semantic network for relational knowledge," in *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pp. 161–176, Springer, Berlin, Germany, 2013.

[17] A. Ritter, Mausam, and O. Etzioni, "A latent dirichlet allocation method for selectional preferences," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pp. 424–434, July 2010.

[18] M. Cheong and V. Lee, "Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base," in *Proceedings of the 2nd ACM Workshop on Social Web Search and Mining (SWSM '09), Co-located with the 18th ACM International Conference on Information and Knowledge Management (CIKM '09)*, pp. 1–8, November 2009.

[19] N. Glance, M. Hurst, and T. Tomokiyo, "Blogpulse: automated trend discovery for weblogs," in *Proceedings of the WWW Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*, 2004.

[20] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th International Conference on the World Wide Web (WWW '04)*, pp. 491–501, 2004.

[21] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pp. 37–45, 1998.

[22] R. Nallapati, A. Feng, F. Peng, and J. Allan, "Event threading within news topics," in *Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM '04)*, pp. 446–453, November 2004.

[23] B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims, "Identifying temporal patterns and key players in document collections," in *Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory, and Applications (TDM '05)*, pp. 165–174, 2005.

[24] W. H. Dai, H. Z. Hu, T. Wu, and Y. H. Dai, "Information spread of emergency events: path searching on social networks," *The Scientific World Journal*, vol. 2014, Article ID 179620, 7 pages, 2014.

[25] W. H. Dai, X. Q. Wan, and X. Y. Liu, "Emergency event: internet spread, psychological impacts and emergency management," *Journal of Computers*, vol. 6, no. 8, pp. 1748–1755, 2011.

[26] H. Z. Hu, D. Wang, W. H. Dai, and L. H. Huang, "Psychology and behavior mechanism of micro-blog information spreading," *African Journal of Business Management*, vol. 6, no. 35, pp. 9797–9807, 2012.

[27] X. H. Hu, T. Mu, W. H. Dai, H. Z. Hu, and G. H. Dai, "Analysis of browsing behaviors with ant colony clustering algorithm," *Journal of Computers*, vol. 7, no. 12, pp. 3096–3102, 2012.

[28] S. Harabagiu and A. Hickl, "Relevance modeling for microblog summarization," in *Proceedings of the 5th International Conference Webblogs and Social Media (ICWSM '11)*, pp. 514–517, Barcelona, Spain, July 2011.

[29] D. Inouye and J. K. Kalita, "Comparing twitter summarization algorithms for multiple post summaries," in *Proceedings of the IEEE 3rd International Conference on Privacy, Security, Risk and Trust (PASSAT) and IEEE 3rd Inernational Conference on Social Computing (SocialCom '11)*, pp. 298–306, Boston, Mass, USA, October 2011.

[30] B. Sharifi, M. Hutton, and J. Kalita, "Automatic summarization of twitter topics," in *Proceedings of the National Workshop Design and Analysis of Algorithm*, pp. 121–128, Assam, India, 2010.

[31] H. Takamura, H. Yokono, and M. Okumura, "Summarizing a document stream," in *Advances in Information Retrieval*, vol. 6611 of *Lecture Notes in Computer Science*, pp. 177–188, Springer, New York, NY, USA, 2011.

[32] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: real-world event identification on twitter," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, 2011.

[33] A. M. Popescu, M. Pennacchiotti, and D. A. Paranjpe, "Extracting events and event descriptions from Twitter," in *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*, pp. 105–106, April 2011.

[34] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5220–5227, 2004.

[35] M. Banko and O. Etzioni, "The tradeoffs between open and traditional relation extraction," in *Proceedings of the ACL '08: HLT*, 2008.

[36] S. Petrović, M. Osborne, and V. Lavrenko, "The Edinburgh twitter corpus," in *Proceedings of the NAACL HLT 2010 Workshop on*

*Computational Linguistics in a World of Social Media (WSA '10)*, June 2010.

[37] T. Joachims, *Learning to classify text using support vector machines [Ph.D. dissertation]*, Kluwer Academic Publishers, 2002.

[38] D. T. Vo and C. Y. Ock, "Learning to classify short text from scientific documents using topic models with various types of knowledge," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1684–1698, 2015.

[39] W. Zhang, T. Yoshida, and X. Tang, "Text classification based on multi-word with support vector machine," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 879–886, 2008.

[40] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[41] B. Gerloff, "Normalized (pointwise) mutual information in collocation extraction," in *Proceedings of the Biennial GSCL Conference*, 2009.

[42] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[43] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2-3, pp. 259–284, 1998.