Event Identification in Social Media using Classification-Clustering Framework

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Nasser Alsaedi

Cardiff University
School of Computer Science & Informatics

Jan 2017

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed ..... (candidate)

Date .....

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed ..... (candidate)

Date .....

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University' s Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed ..... (candidate)

Date .....

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Abstract

In recent years, there has been increased interest in real-world event detection using publicly accessible data made available through Internet technology such as Twitter, Facebook and YouTube. In these highly interactive systems the general public are able to post real-time reactions to “real world" events - thereby acting as social sensors of terrestrial activity. Automatically detecting and categorizing events, particularly small-scale incidents, using streamed data is a non-trivial task, due to the heterogeneity, the scalability and the varied quality of the data as well as the presence of noise and irrelevant information. However, it would be of high value to public safety organisations such as local police, who need to respond accordingly. To address these challenges we present an end-to-end integrated event detection framework which comprises five main components: data collection, pre-processing, classification, online clustering and summarization. The integration between classification and clustering enables events to be detected, especially “disruptive events" - incidents that threaten social safety and security, or that could disrupt social order. We present an evaluation of the effectiveness of detecting events using a variety of features derived from Twitter posts, namely: temporal, spatial and textual content. We evaluate our framework on large-scale, real-world datasets from Twitter and Flickr. Furthermore, we apply our event detection system to a large corpus of tweets posted during the August 2011 riots in England. We show that our system can perform as well as terrestrial sources, such as police reports, traditional surveillance, and emergency calls, even better than local police intelligence in most cases. The framework developed in this thesis provides a scalable,

online solution, to handle the high volume of social media documents in different languages including English, Arabic, Eastern languages such as Chinese, and many Latin languages.

Moreover, event detection is a concept that is crucial to the assurance of public safety surrounding real-world events. Decision makers use information from a range of terrestrial and online sources to help inform decisions that enable them to develop policies and react appropriately to events as they unfold. Due to the heterogeneity and scale of the data and the fact that some messages are more salient than others for the purposes of understanding any risk to human safety and managing any disruption caused by events, automatic summarization of event-related microblogs is a non-trivial and important problem. In this thesis we tackle the task of automatic summarization of Twitter posts, and present three methods that produce summaries by selecting the most representative posts from real-world tweet-event clusters. To evaluate our approaches, we compare them to the state-of-the-art summarization systems and human generated summaries. Our results show that our proposed methods outperform all the other summarization systems for English and non-English corpora.

Acknowledgements

When I write these acknowledgments, I realize that my formal education will end very soon. Reviewing the three years and three months of my graduate study at the Cardiff School of Computer Science & Informatics, not only has Cardiff University given me a very solid education in computer science and informatics, but it has helped me to build my confidence to pursue career success after my graduation. I am greatly indebted to many people for the completion of this thesis, and to still more for making the journey a uniquely rewarding one.

First of all, I would like to thank my country and my sponsors; I would like to thank H.H. Sheikh Mohamed bin Zayed Al Nahyan, Crown Prince of Abu Dhabi and Deputy Supreme Commander of the UAE Armed Forces, for his continuous support and guidance. I would also like to thank H.H. Sheikh Saif bin Zayed Al Nahyan, Deputy Prime Minister and Minister of Interior for his support to education and encouraging students and researchers. I sincerely thank my sponsor Abu Dhabi Police GHQ and H.E. Major General Mohammed Al Romaithi, Commander-in-Chief of Abu Dhabi Police, for his role, support and encouragement.

This thesis and the research it contains could not exist if not for the unfailing support of my supervisor, Pete Burnap. His guidance and input have provided drive, shaped my research and have given me the confidence required for producing and defending ideas both within and outside of research. A big thanks also goes to my second supervisor Omer Rana for his experience, intellectual guidance and belief in me. It has been a

very long journey!

Finally, my time as a research student would not have been the same without the constant presence, support and the positive atmosphere from the School of Computer Science & Informatics. In addition, I am grateful to the fellow researchers with whom I shared an office for the research period. I am also grateful to all my colleagues and friends in Cardiff who have always pushed me and provided encouragement to make me who I am today. They deserve my utmost thanks.

Most importantly, thanks to my family, for supporting me all the way throughout the journey. The thesis journey ends, but ours continues!

Contents

List of Publications

The work introduced in this thesis is based on the following publications:

- Nasser Alsaedi, Pete Burnap, and Omer Rana. Can we predict a riot? disruptive event detection using twitter. *ACM Transactions on Internet Technology (TOIT)*, 17(2):18:1–18:26, March 2017

- Nasser Alsaedi, Pete Burnap, and Omer Rana. A combined classification-clustering framework for identifying disruptive events. In *Proceedings of the 7th ASE International Conference on Social Computing, Stanford University, CA., USA*, SocialCom '14, 2014

- Nasser Alsaedi and Pete Burnap. Arabic event detection in social media. In *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14 _ 20 April*, CICLing '15, pages 384–401. Springer, 2015

- Nasser Alsaedi, Pete Burnap, and Omer Rana. Identifying disruptive events from social media to enhance situational awareness. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 934–941, New York, NY, USA, 2015. ACM

- Nasser Alsaedi and Pete Burnap. Feature extraction and analysis for identifying disruptive events from social media. In *Proceedings of the 2015 IEEE/ACM*

International Conference on Advances in Social Networks Analysis and Mining 2015, ASONAM '15, pages 1495–1502, New York, NY, USA, 2015. ACM

- Nasser Alsaedi, Pete Burnap, and Omer F. Rana. Automatic summarization of real world events using twitter. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 511–514, 2016

- Nasser Alsaedi, Pete Burnap, and Omer F. Rana. Sensing real-world events using arabic twitter posts. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 515–518, 2016

List of Figures

List of Tables

List of Algorithms

List of Acronyms

TDT Topic Detection and Tracking

NLP Natural Language Processing

IR Information Retrieval

TC Text Categorization

API Application Programming Interface

SED Social Event Detection

MPS Metropolitan Police Service

DSS Decision Support System

DSRM Data Science Research Methodology

NB Naive Bayes

SVMs Support Vector Machines

POS Part-of-Speech

NER Named Entity Recognition

MICI Maximal Information Compression Index

TF-IDF Term Frequency - Inverse Document Frequency

KLD Kullback-Leibler Divergence

DFT Discrete Fourier Transformation

LSH Locality-Sensitive Hashing

LDA Latent Dirichlet Allocation

SVD Singular Value Decomposition

SCAN Structural Clustering Algorithm for Networks

FP false positives

FN false negatives

TP true positives

TN true negatives

AP average precision

NDCG Normalized Discounted Cumulative Gain

NMI Normalized Mutual Information

ROC Receiver Operating Characteristics

ROUGE Recall-Oriented Understudy for Gisting Evaluation

PR Phrase Reinforcement

MDS Multi-Document Summarization

PRF Pseudo Relevance Feedback

DTM Decay Topic Model

GDTM Gaussian Decay Topic Model

TCV Tweet Cluster Vector

NMF Non-negative Matrix Factorization

pLSA Probabilistic Latent Semantic Analysis

LCS Longest Common Subsequence

Chapter 1

Introduction

1.1 Background and Motivation

Microblogging, as a form of social media, is a fast emerging tool for expressing opinions, broadcasting news, and facilitating the interaction between people. The ease of publishing content on social media sites and the wide spread of various electronic devices (e.g. cellphones, tablets, etc.) have enabled users to report real-life events as they happen around them. One of the most representative examples of social media is Twitter, which allows users to publish short tweets (messages within a 140-character limit) about any subject. The range of widely known events includes community-specific events, such as local gatherings, or can be wider-reaching national or even international in significance. For example, the Iranian election protests in 2009 were extensively reported by Twitter users [65, 165]. Another good example, where Twitter was employed as a resource for the US government to communicate with citizens, was the outbreak of swine flu when the US Center for Disease Control (CDC) used Twitter to post the latest updates on the pandemic [125].

People tend to comment on real-world events they encounter, both local and global, when a topic suddenly attracts their attention, for example, a sporting event [8, 40], adverse weather update [102], or terror attack [104, 99, 84], etc. For the purposes of this study an event can be defined as an occurrence at a specific time and place that is associated with a topic (Chapter 2). From this definition, we can infer that these events

have several characteristics: i) there are enough users interacting and connecting with the event using tweets (therefore if an event was not reported by users in social media platforms, it will not be identified by our framework), ii) these tweets are discussing the same topic in similar words, and iii) the event takes place at a specific time and within a geographical boundary.

As an example of an event, consider a possible football match "Manchester United vs Liverpool". Users and supporters will use similar vocabulary to describe and report this event. They might use players' names, managers' names, and locations (e.g. the venue or surrounding areas), and also verbs, adjectives and nouns, etc. in reporting the match. Mining these features offers the potential for events to be detected and summarised promptly [127]. User engagement may depend on the event's significance and scale, which could play an important role in identifying events [55]. In this thesis, the terms *detection* and *identification* are used interchangeably.

Recently, increased interest has been shown in real-world event identification from social media sites. It is very challenging to automatically organize user-generated content with respect to events, as well as identifying and characterizing these events according to scale in real time. First, the speed and volume at which data arrive, where tweets arrive continuously in chronological order, and the size of the Twitter network produce a continuously changing dynamic corpus. The significant amount of "noise" presented in the stream is another key challenge; in fact noise constitutes around 40% of all tweets, which have been reported as pointless "babbles" [157] like "let's go to the beach" or "the weather is amazing". In addition, the dynamic nature of events leads to a diverse set of linguistic features. This is compounded by the fact that each social media post (or tag, in the case of photos or videos) is short, which means that only a limited content is available for analysis.

The number of users of social media who actively post content is growing rapidly day by day, so event detection algorithms need to incorporate the minimal number of operations if they are to reduce computational overhead when analyzing real-time

streams. Other challenges are inherent to the microblogging language and nature; they include the frequent use of informal, irregular, and abbreviated words, the high number of spelling and grammatical errors, and the use of improper (informal) sentence structures and mixed language. Additionally, social media characteristics and popularity have attracted spammers to spread advertisements, pornography, viruses, and other malicious activities [13, 57, 73].

Decision makers or researchers might be interested in many kinds of event, making it difficult to anticipate those types which could be important and a priori build detectors for them [124]. Moreover, non-event content is, of course, prominent on Twitter and similar systems; people want to share various types of content, such as personal updates, random thoughts and musings, opinions, and information [16]. The rapid growth of online social media has made it possible for rumors and misinformation to spread very quickly and widely. These online platforms have enabled unreliable and untrusted sources to spread large amounts of unverified information to users, which leads to other challenging problems [116].

Yet social media services such as Facebook and Twitter are providing researchers with new opportunities thanks to the availability of the data they provide. These platforms are by definition free and openly accessible and also pervasive via Smartphone apps, not to mention being part of a widespread subculture of social media sites that encourages users to acquire a large pool of friends [129]. In addition, the social media have substantially reduced communication response time below that needed by the industrial media, in significantly altering the rate at which information is exchanged and consumed [154]. The restriction on the length of a Twitter message invariably means that the tweets do not necessarily contain well-formed ideas, yet full enough for users to make sense of what they read in them [8, 129].

Furthermore, not all events reported in social media are reported in newspapers and other traditional media due to their nature and the freedom of social media platforms. Messages posted on Twitter (tweets) have been reporting everything as they occur from

stories of everyday life to the latest local and global news and events. Thus Twitter can be used as a varied, valuable, and continuous source of information that enables users and organizations to acquire actionable knowledge [9].

Many recent approaches have been proposed for identifying events in the social media which rely on a set of manually selected terms to retrieve event-related documents or those related to particular types of event. Some of these approaches are limited to widely discussed events and are not designed to report small-scale incidents. Moreover, other existing methods have the main drawback of requiring an a priori specification of the total number of topics or are not ideal for social media event detection because they may not be able to capture events in real time given the velocity and scale of the updates in social networks. Moreover, small-scale event detection systems can only detect small or particular type events such as only fire incidents or only car accidents (or road and traffic updates) which limit their overall usability and functionality. There are also disaster identification approaches which are mainly used in filtering, searching, and analyzing tweets during natural disasters such as earthquakes, tornadoes, etc.

In this thesis, we propose a general online classification-clustering framework which is able to handle a constant stream of new documents with a threshold parameter that can be modified experimentally during the training phase. Our proposed framework can detect both large (widely reported) or small scale events (only reported by small number of users), however it does not classify events as large or small events. The high volume of updates from the social media sites is the input of the system, which identifies a number of events in a particular region, the associated sub-events (details), and "disruptive" events - incidents that threaten social safety and security, or that could disrupt social order (Chapter 4). The event detection framework comprises five main elements: data collection, pre-processing, classification, online clustering and summarization. Social media data are very noisy; hence, the first step in this framework after collecting data is pre-processing, which aims to reduce the amount of noise before classification by reducing the number of attributes and putting tweets in the correct

form before they can be analyzed.

The next step is to separate event-related tweets and non-event content using machine learning algorithms, particularly classification learning. Then we compute the features of messages in order to extract similar characteristics and apply incremental online clustering algorithms to assign each message in turn to a suitable event-based cluster after calculating the message's similarity to the existing clusters. The integration between classification and clustering makes it possible to detect events including disruptive events within a particular time-frame (daily or hourly). Finally, we tackle the task of selecting the most salient social media content for an event (cluster) in a process called summarization or representation.

Here, we are interested not only in identifying events and their associated social media documents throughout the day, but also in the identification of a special type of event called "disruptive events". We define a disruptive event in the context of the social media as a special type of event that obstructs the achieving of the objective(s) of another event or interrupts the routine of another event (Chapter 4). Disruptive events have different characteristics and we hypothesize that they can be captured by a set of features: temporal, spatial and textual. Therefore, understanding the features of social media content that single out disruptive events is a key motivation behind this work. One way to optimize the identification of the patterns and signals that indicate an event is to undertake feature selection (optimization), because not all features are expected to lead to better system performance or contribute equally towards improved machine classification and/or clustering accuracy.

Disruptive events, like other events, range from large-scale events, often global in scale such as terrorist attacks or disaster-related events, to small-scale and localized incidents such as fires, car accidents, and events threatening public order. In this thesis, we define a small-scale event as: An occurrence that discusses a particular topic at a specific time and place but only reported by few number of users, whereas large-scale event is an event that is reported by large number of users. Some of the large-

scale events start as small-scale incidents before they escalate and become damaging to the wider society and business. It has been noted that detecting of small-scale events is essential to improving situational awareness of both citizens and decision makers [132, 153, 76] and hence this remains a well motivated research topic for the social computing community. In this thesis, we propose a novel approach to event detection that aims to overcome many challenges and provide a system for detecting both large-scale events and related small-scale events. The approach is based on the integration of supervised machine learning algorithms to detect larger-scale events, and unsupervised approaches to clustering, disambiguating and summarizing smaller sub-events, with the goal of improving situational awareness in emergency situations by automatic methods.

1.2 Hypothesis and Research Questions

The hypothesis of the present research is as follows.

We can automatically identify real-world events including disruptive events as they happen from posts on the social media in a particular place and for a predefined time period to improve public safety and decision support.

The research questions assist in understanding the scope of the work in this thesis. There are four main research questions:

RQ1 Can we detect "events" in real time from the streaming media and introduce a strategy to integrate this knowledge into a Decision Support System (DSS)?

RQ2 Can we identify sub-event details including disruptive events and their context within the streaming media (topic clustering)?

RQ3 Since not all features are expected to improve a system's performance, can we investigate the dynamics of event/topic identification of three kinds of influential

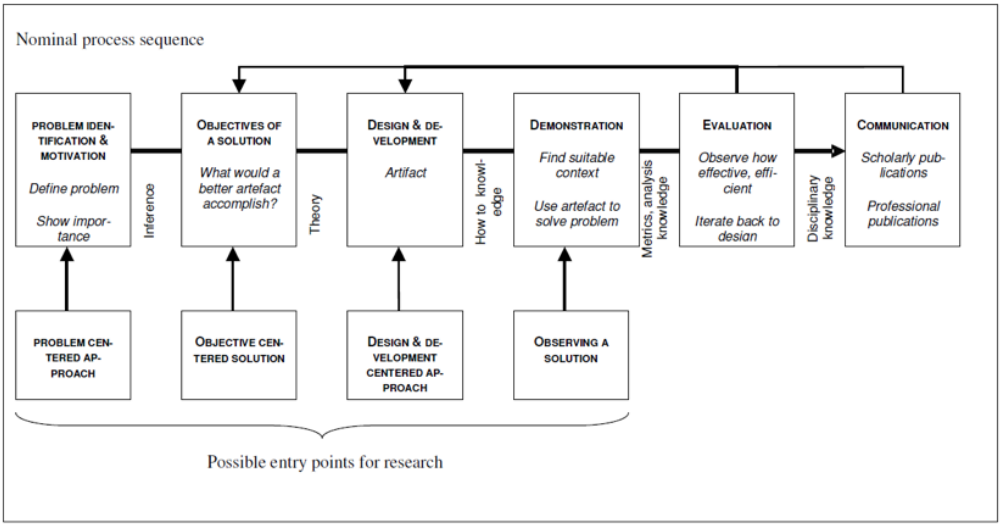


Figure 1.1: Design Science Research Methadology (DSRM) Process Model. Adapted and modified from Peffers et al. [110].

feature: temporal, spatial and textual, in order to optimize feature selection and to improve the effectiveness of topic clustering?

RQ4 Can we summarize events to enable decision makers to read effectively only high quality summaries of most representative posts from Twitter?

1.3 Research Methodology

The main purpose behind our work in this thesis is to develop techniques for identifying real-world events in real-time using a classification-clustering framework to improve the performance of social media event identification. To this end, and to verify the hypothesis, we design a general methodology that we use in the different parts of this thesis. Our methodology uses the Data Science Research Methodology (DSRM) introduced by Peffers el al. [110] as depicted in Figure 1.1. Each step is described and related to the thesis chapters as follows:

1. **Problem Identification and Motivation:**

   This phase involves critical thinking of the research problem and modelling strategies and justifies the value of designing a solution to that problem. The first step of this phase involves the identification of gaps in related literature as presented in Chapter 2. A literature review investigates several existing models and techniques for the task of event detection in social media as well as identifies shortcomings in current approaches and looks at what can be done to improve social media event detection. In the second step, the research hypothesis statement and the research questions are identified as presented in Chapter 1. The third step requires the choice of the data source and the development tools that will be used to test the hypotheses. The last step involves research planing by dividing the main problem into tasks (classification, clustering, feature selection and summarization) and identifying the required milestones.

2. **Objectives of the Solution:**

   This stage requires knowledge of the state of the problem, and current solutions and their efficacy. The problem definition in the previous stage is used in order to propose the objectives of the solution. In this research, our problem is most similar to the event detection and tracking task, whose objective is to identify events in a continuous stream of news documents (e.g., newswire). However, our problem exhibits some fundamental differences from traditional event detection that originate from the focus on social media sources. The objective of this work is to develop a framework to mine, analyze and summarize events from social media sites. The aim of this thesis also to test the performance of a number of methods and models with the intended outcome of selecting the best performance methods for the event detection task.

3. **Design and Development:**

   This step aims to design and develop a solution of the problem. This step is explained in chapters (three, four, and five). The entire design of the proposed

framework is outlined in Chapter 3. The design of different machine learning methods used in text classification for separating event content from non-event is introduced in Chapter 3. In Chapter 4, we design an incremental on-line clustering algorithm which assigns similar messages in turn to similar clusters based on a similarity measure. We also explore three sets of features (temporal, spatial and textual) and combinations of them, in Chapter 4. Finally, the design of three techniques for the task of microblog summarization is presented in Chapter 5.

4. **Demonstration:**

   This step involves using the developed framework in a suitable context. In this thesis, different experiments are carried out in chapters (three, four, and five) using samples of realistic and representative data sets for a representative number of users with different locations, communities, languages, and backgrounds to demonstrate the effectiveness of the proposed framework. We have also demonstrated that the proposed framework can be used to identify real-world events, in Chapter 4, when we compare its performance against other leading approaches using Twitter posts from the UK riots in 2011, and publicly accessible reports received by the Metropolitan Police Service during the UK riots in 2011.

5. **Evaluation:**

   This step observes and measures how well the proposed framework supports a solution to the problem. It involves assessing the effectiveness of the proposed framework compared to other existing methods. In this research, once the methods are developed, researchers start a thorough testing process for each element of the framework. First, the three machine learning methods are extensively evaluated to separate event content from non-event according to standard evaluation metrics (Chapter 3). Next, we extensively test the effectiveness of our online clustering algorithm using several large real-world datasets from different social media platforms. Several experiments are conducted to compare the proposed clustering approach performance against many leading approaches in

the event detection task. Finally, three automatic summarization/representation techniques for summarizing Twitter messages are successfully tested on English, Arabic and Japanese language tweets to test their applicability across multiple languages. We also compare the three proposed summarization techniques with a number of recent and leading summarization systems

6. **Communication:**

   The main contributions of this thesis have been published in peer reviewed scholarly publications. This thesis resulted in seven publications, six conference papers and one journal article paper. The publications are listed in the list of publications section.

1.4 Main Contributions

This research describes effective techniques for event detection, event monitoring (tracking), topic clustering and event summarization. The combination of supervised learning and unsupervised learning enables the identification of main events and sub-events, including disruptive events, and supports our hypothesis that the social media can be used as primary sources of information. In addition, one way to optimize the identification of patterns and signals that would be indicative of an event is to undertake feature selection experiments. We aim to understand the effectiveness of a range of features for identifying events, in particular, features that would distinguish "normal" events from disruptive events. We then propose techniques that reduce noise and focus on summarizing Twitter messages linked with events to improve event reasoning, visualization, and analytics. Our methods are language independent, and satisfy real-time requirements as well as being suitable to the huge volume of data. Therefore, the main contributions of this thesis are the following:

- We propose a novel framework that identifies the relationship between social

media activity and real-world events and detects key events throughout the day;

- Using temporal, spatial and textual features, our framework is able to detect disruptive events in a given place for a certain time;

- Extensive feature analysis and feature selection are performed in order to demonstrate that these features contribute differently in the process of decision-making with regard to the management of real-time disruptive events.

- We develop several approaches for summarizing microblogging posts linked with events without the need for prior knowledge of the entire dataset;

- We evaluate the behaviour and effectiveness of the machine learning algorithms, the on-line clustering and summarization techniques on large real-world datasets.

- We validate the overall model performance on several events (including a Formula 1 car racing event) as well as the MediaEval2012 Social Event Detection (SED) benchmark [108] to show the effectiveness of the framework. We further evaluate it against other leading approaches using Twitter posts from the UK riots in 2011, and a publicly accessible account of *actual reported* intelligence obtained and reports received by the Metropolitan Police Service during this event. Smaller scale events included localized looting, violence and criminal damage. The results show that our system can detect events related to the riots as well as terrestrial sources did - in some cases we detect the event *before* the intelligence reports were recorded.

- We extend and test the applicability of our algorithm to identify events and disruptive events in the Arabic microblogging context (Arabic is the most complex and challenging language regarding data mining, due to its orthography and morphology [37]). To the best of our knowledge, this study is the first attempt to identify real-world events in Arabic from social media posts, which itself can be considered a contribution. This work has been previously published at the following venues [6, 12].

1.5 Organization of the Thesis

In this thesis, we investigate and evaluate solutions to the problem of online real-world event identification for both large-scale and rare events such as car accidents in a given location. We present techniques that are related to three main areas: text classification, online clustering and automated summarization for various microblogging platforms. In particular, we investigate and examine new and exciting techniques in order to propose an event identification framework which can be generalized in the future to develop a social awareness system or a credible source of information. The outline of this thesis is as follows:

Chapter 2 discusses different definitions of an event with examples from real-world scenarios. We then review the related work on event detection in the social media (advantages and limitations of each work) and the main applications that could arise from a scalable event detection approach. Then, we describe the first two steps in our framework: data collection and pre-processing. Details are given of the datasets that we have used for the purposes of evaluating the proposed framework in total and in its various aspects. This chapter also presents various pre-processing techniques and we analyze the impact of preprocessing social media content for the task of event detection.

Chapter 3 investigates different machine learning methods used in text classification for separating event content from non-event, mainly the Naive Bayes classifier [75], Logistic Regression [47] and Support Vector Machines (SVMs) [62]. Throughout our extensive experimentation, we evaluate these classification methods and their effectiveness at distinguishing event and non-event messages. Furthermore, we aim to investigate methods to improve the performance of the classification results; thus we consider several features which capture patterns in the data, such as the n-gram presence or n-gram frequency, the use of unigrams, bigrams and trigrams, linguistic features such as parts-of-speech (POS) tagging and Named Entity Recognition (NER). Parts of this chapter were published at the following venue [8].

Chapter 4 defines the term 'disruptive event' in the context of the social media. Then we outline the incremental on-line clustering algorithm which assigns each message in turn to a suitable event-based cluster after calculating similarity of the post to the existing clusters. We also test the effectiveness of the on-line clustering using different events and various datasets. To explore three kinds of feature (temporal, spatial and textual) and combinations of them, in order to achieve better system performance, we implement an improved model for feature selection that is suitable for microblog data. We perform extensive feature analysis and feature selection in order to demonstrate that these features contribute differently to the process of decision-making regarding the management of real-time disruptive events. Then, we validate the effectiveness of our framework using several large real-world datasets from Twitter and Flickr. We compare the overall performance of our system using the optimized model in terms of Precision, Recall, F-Measure and NMI to the performance of many leading systems, namely, Spatial LDA [106], unsupervised methods [16], [167], topic models [148] and a graph-based approach [131]. We further evaluate it against other leading approaches using Twitter posts from the UK riots in 2011.Together with parts of Chapter 3, the research in these two chapters has been published at the following venues [10, 7, 12].

Chapter 5 presents three techniques for the task of microblog summarization. Summarization methods are language independent, they satisfy the real-time requirement and are suitable for high volumes of data. We evaluate our proposed techniques using two noisy datasets according to well-known metrics (quality, relevance and usefulness). We also validate our system against the state-of-the-art methods including centroid method (Becker et al.) [17], Zubiaga et al. sub-event detection and then the tweet selection method [167], Xu et al.; a graph-based approach [160] and a hybrid TF-IDF (term frequency summarization approach) [59]. We further validate the system using English, Arabic and Japanese corpora. Our results show that our proposed methods outperform all the other summarization systems for English and non-English corpora. The research in this chapter extends work that has been previously published at the following venue [11].

Chapter 6 presents our conclusions from the research and discusses the possible directions that future work might take.

Chapter 2

Event Detection and Characterization

2.1 Introduction

The term ‘event’ can be defined in a number of ways, depending on the domains and the interest of the users or decision makers. The definition of an event varies in granularity as well, depending on the way in which the event detection will be applied. Our focus is to define and connect these definitions to our task of identifying and characterizing events in the social media. The goal of identifying events and their associated documents on social media sites is to monitor real-time social media streams and extract information. We seek in particular information of high value to public safety organizations such as the local police or emergency departments who need to respond accordingly.

The rapid growth of Internet-enabled communication technology in the form of social networking services (often collectively referred to as the social media) and their associated smartphone apps has enabled billions of global citizens to broadcast news and ‘on the ground’ information during ‘real world’ events as they unfold. Twitter, for example, has been studied as an emerging news reporting platform [113, 157, 105] and has been widely used to disseminate information about the Arab Spring uprisings [143, 6] and other disaster-related incidents [57, 136, 25, 158]. The interaction between people, events, and Internet-enabled technology, presents both an opportunity and a challenge to social computing scholars, public sector organizations (e.g. governments and poli-

cing agencies), and the private sector, all of whom aim to understand how events are reported using social media and how millions of online posts can be reduced to accurate but meaningful information for making wise decisions and carrying out productive action.

In this chapter, we survey the various definitions of 'event' in the literature from a variety of academic disciplines, especially Topic Detection and Tracking (TDT), which include tasks such as event identification, tracking, and filtering, as well as topic segmentation and event summarization. Then we overview some recent works that use user generated content sources for global and local event identification in the social media, and highlight both their benefits and shortcomings. In addition, we present our end-to-end integrated event detection framework which contains five main elements: data collection, pre-processing, classification, online clustering and summarization. Section (2.4, 2.5) discusses the first two stages, data collection and pre-processing. In data collection, we describe the datasets that we use in our subsequent experimental chapters, including three large-scale, real-world datasets from Twitter and one dataset from Flickr [109]. Section 2.5 explores traditional text processing steps such as stop-word elimination and stemming.

2.2 Event Definition and Examples

The definition of 'event' varies across academic fields, from social computing to Topic Detection and Tracking (TDT). Even within a specific domain, researchers often disagree on what precisely constitutes an event or the characteristics of an event [93]. An event on a social media platform can be loosely defined as a specific thing that happens at a specific time and place [3]. Wang et al. [154] defined an event as: "An occurrence causing changes in the volume of text data that discusses the associated topic at a specific time". This definition suggests that the occurrence is characterized by topic and time, and often associated with entities such as people and location. Of

course, if an event was not reported by social media users, then it will not be identified by our framework. For example, if a conference or a local gathering or a small car accident were not reported in social media sites, then event detection systems will not be able to detect such events from social media. Furthermore, [135] defined an event as something that happens with a defined beginning and ending within the scope of a document. These definitions touch on an important aspect of an event, namely, its necessary association with the time dimension.

All these definitions share the same aspects of an event; hence, an event can be characterized by one or more of the following attributes: Topic, Time, People and Location. In other words, according to the knowledge from social media sources, events can be represented by four kinds of attribute: what (topic or keywords), when (time), where (location) and who (people) [77]. We define an event as a real world happening that is reflected by change in the volume of text data that discusses the associated topic at a specific time and place, similar to Dong et al. in [42]. For the purposes of this study, we define a small-scale event as: An occurrence that discusses a particular topic at a specific time and place but only reported by few number of users, whereas large-scale event is an event that is reported by a large number of users.

Chen and Roy [29] slightly adapted the definition of an event in the context of a set of photos rather than text. Thus, if a set of photos represents an event, it should at least satisfy the following three constraints: (1) The group of photos concerns the same specific thing. That is, the content of the photos should be semantically consistent. (2) The group of photos should be taken within a certain time segment. (3) The group of photos should be taken within a similar location. We can also generally extend or modify the definition of event to handle a set of videos.

These events range from widely known (global) ones (such as political events or professional sports/games) to smaller-scale, local events (such as a community gathering or a local conference). *Global* events are events that refer to real world happenings whose effects are not restricted to a certain location. *Local* events, by contrast, are

events that refer to real world happenings whose effects are restricted to a certain location [21]. People are interested in both types of event: the former are important world events that one should know about, and the latter affect our everyday life more directly [156].

Real world events are also divided into *known* (predictable or planned) events such as concerts or conferences and *unknown* (unpredictable or unplanned) events such as earthquakes. Obviously, known events should be relatively easy to detect, given the long time of planning and preparation for them that is involved [151, 16]. Events can be further divided into *periodic* and *aperiodic* events. The former are events that occur in sequence at roughly equal intervals, whereas an aperiodic event happens only once within a given period [29, 53]. Other researchers describe periodic events as scheduled or known events where the system is provided with the time that the events will start, scheduled in advance, so the system knows when to start looking for them [167]. Examples of scheduled events are sporting events, ceremonies and product launches. He et al. [53] defined and investigated periodic and aperiodic events and applied spectral analysis using Discrete Fourier transformation (DFT) to categorize the features of both.

Different events have different context-specific features but are generally conveyed on social media using verbs (actions), nouns (names, places, topics ...), adjectives (descriptive) and prepositional phrases (describing proximity and location). Our aim is to represent the data extracted from the social media as a time-line of events (clusters), where each cluster contains sufficient data to discriminate between events and summarize them into actionable information for use by public safety officials and policy makers.

Experimentally, events can be characterized by burst detection, where different events have different terms causing a "volume change" to the use of these terms when discussed on social media platforms [8]. The rationale behind our approach is that events belonging to the same topic often share a set of keywords. Tweets discussing the same

event, for instance an air crash, tend to share similar patterns or words, such as airplane, crash, death, and injuries [163]. An event is therefore conventionally represented by a number of keywords showing a burst in their frequency count [67].

Although these keywords are informative and used to discriminate on-topic and off-topic documents, they cannot differentiate between similar events or events of the same type. Hence event detection methods that are based only on word occurrence/frequency analysis cannot distinguish between two or more events of the same type. Consequently these methods may fail to report multiple airplane crashes, or, more precisely, they may identify the first event and miss the others. Or they might consider all of them as one incident/event which means that one event which could be a major event might be omitted - unless the system has the ability to distinguish the discriminative features (words or phrases) for classifying the topic from those discriminative for event distinction [163].

Therefore, relying on textual content alone is generally insufficient particularly for small-scale incidents such as car accidents. Many road accidents are of a kind that happens every day and has effects that last for a few hours on different roads. To achieve good system performance, small-scale incident detection that analyzes tweets for incidents needs to distinguish crashes from one another [132]. Hence, additional knowledge of the time and the location is needed and probably other entities (ex. Url, link or picture) should be brought in to distinguish between similar events and aggregate further information about the kinds that occur repeatedly.

We assume that the task of event detection may be needed in one of the following three scenarios: (a) different events occurring in the same location within one time period, where we assume that each event can be characterized using different *textual features*; (b) similar multiple events in different locations, where we assume that the most appropriate features will be *temporal* and *spatial features*; and (c) similar events in the same location at almost the same time. In this case we assume that they are the same event and group them together, treating the new documents as updates of earlier

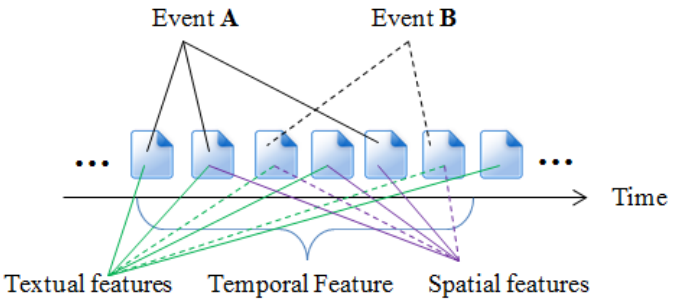


Figure 2.1: Document clustering using different sets of features

ones.

One of the main empirical foci of this work is an exploration of the most effective features in the data from the social media for event detection. Consider a text stream $D = (D_1, D_2, ..., D_n)$ where D_i is a document, and the length of D is $|D|$. A document d_i consists of a set of features,$(F_1, F_2, ..., F_k)$, and is reported at time t_i. The text stream, D, is divided into time windows, W_i of the same length e.g. per day, per 12 hours, per minute. The problem of real-time event detection is to find an optimal set of features to detect events in each unit of time, where all known events are identified and correctly summarized. Figure 2.1 illustrates the relationship between events and the different sets of features.

2.3 Related Work: Event Detection in Social Media

The general topic of detecting real-world events from the social media has drawn considerable research interest. Research efforts have focused on real-time event detection and tracking, social media analysis, micro-blog summarization and information visualization. We describe the related work in three areas: large-scale (global) event detection, small-scale (local) event detection, and systems used to extract crisis relevant information from the social media.

Large-scale event detection: For large-scale (global) events such as the 2016 United

States presidential election or the reaction to the 2014-2015 Ebola Outbreak in West Africa, Petrovic et al. [112] presented an approach to detect breaking stories from a stream of tweets using locality-sensitive hashing (LSH). Becker et al. [16] proposed an online clustering framework to identify different types of real-world event. Then the researchers used different machine learning models to predict whether a pair of documents belonged to the group of real-world events or not. The authors in [148] proposed an efficient methodology for performing event detection from large time-stamped web document streams. Their methodology successfully integrated named entity recognition, dynamic topic map discovery, topic clustering, and peak detection techniques. All these approaches, however, are limited to widely discussed events and are not designed to report rare and potentially disruptive small-scale incidents.

Large-scale event detection has also been explored through the clustering of discrete wavelet signals built from individual words generated by Twitter [157]. Auto-correlation then, by modularity graph partitioning, filters out the trivial words (noise) and cross correlation groups together with words that relate to an event. Similarly, Cordeiro in [35] proposed a continuous wavelet transformation based on hashtag occurrences combined with topic model inference using Latent Dirichlet Allocation (LDA) [20]. In fact, LDA and its variants form a widely used statistical modelling approach implemented in event detection tasks [148, 106, 35, 150]. However, these methods have the main drawback of requiring an a priori specification of the total number of topics, which leads to problems when the total number of event exceeds this number.

Other approaches have focused on structural networks and graph models to discover events in the social media feeds. Benson et al. [18] presented a structured graphical model which simultaneously analyzed individual messages, clustered them according to event, and induced a canonical value for each event property. Using a different graph analytical approach, Sayyadi and Raschid [130] used a KeyGraph algorithm [100] to convert text data into a term graph based on co-occurrence relations between the terms. Then they employed a community detection approach to partition the graph. In this,

each community is regarded as a topic and the terms in the community are eventually considered the topic's features. In addition, Schinas et al. [131] used the Structural Clustering Algorithm for Networks (SCAN) to detect "communities" of documents. These candidate social events were further processed by splitting the events that exceeded a predefined time range into shorter events. Then the researchers used a classification approach based on median geolocations and accumulated tf-idf vectors for each cluster to separate the relevant and irrelevant candidate events. However, these graph partitioning algorithms are not ideal for social media event detection because they may not be able to capture events given the velocity and scale of the updates in social networks in real time as events unfold.

Small-scale event detection: Various methods have been proposed to identify small-scale (local) events such as fire incidents, traffic jams, shooting incidents etc. from social media streams. Walther and Kaisser [153] developed spatiotemporal clustering methods which monitor the specific locations of high tweeting activity and cluster tweets that are geographically and temporally close to each other. A machine-learning module is then used to evaluate whether a cluster of tweets refers to an event based on 41 features, including the tweet content. Another clustering approach is presented in [132] by Schulz et al., with a small-scale incident detection pipeline based on the clustering of incident-related micro-posts. It uses three properties that define an incident: (1) incident type, (2) location and (3) time period. Various techniques are adopted to increase the quality of their clustering approach: (A) the incident type determination, using supervised machine learning (Semantic Abstraction); (B) geotagging of tweets based on their geolocalization; and (C) the extraction of the time period of the incident. Although both methods are very specific to particular incident type (for example [132] is specific to car incidents, fire incidents, and shooting incidents) without giving aspects of the general context (i.e. what are the other events around these specific events?), it is critical that the system can provide insight into the other ongoing events and sub-events arising amid the specific events. This may explain the low recall values of the [132] and [153] approaches when validated using real-world official reports,

32.14% and 4.75%, respectively.

Another event detection system, Twitcident [1] by Abel et al., presents a Web-based application for searching, filtering and aggregating information about known events reported by emergency broadcasting services in the Netherlands. In addition, Watanabe et al. [156] proposed a system called Jasmine, for detecting local events in the real world using geolocation information from microblog documents. It obtains the name list of locations from geotagged tweets and adds positional information to tweets by matching the location name. They use two gazetteers or as they call them (Place Name Database): the first one is the well-known Foursquare service (https://foursquare.com/) and the second is Solr (http://lucene.apache.org/solr/), an open source search platform based on Lucene. A similar system is EventRadar [21] by Boettcher and Lee. EventRadar introduces a statistical method for detecting local events using a temporal and spatial analysis by considering seven- day historic data. The main contribution of EventRadar is that it detects local events without needing a list of locations by finding clusters of Tweets that contain the same subset of words. Another related system is proposed by Li et al. in [76] to detect crime and disaster related Events (CDE) from tweets. They use the spatial and temporal information from tweets to detect new events and extract the meta information by a number of text mining techniques (e.g., geo-location names, temporal phrase, and keywords) for event interpretation. These systems (Twitcident, Jasmine, EventRadar, and TEDAS) can only detect small events and all of them (except EventRadar) require location names in advance, which limit their overall scope and portability.

Event detection for disaster and emergency events: Regarding the use of social media data in disasters, researchers have proposed several visual analytics approaches aiming at real-time microblog analysis that often facilitates interactive means for exploring and identifying anomalies [43]. TwitterMonitor [88] performs trend detection in two steps and analyzes trends in a third step. During the first phase, it identifies bursty keywords which are then grouped on the basis of their co-occurrence. Once

a trend is identified, additional information from the tweets is extracted to analyze and describe the trend. AIDR (Artificial Intelligence for Disaster Response) [58] is a platform for filtering and classifying messages in real time that are posted to social media during humanitarian crises. AIDR uses human-assigned labels (crowdsourcing messages), and pre-existing classification techniques to classify Twitter messages into a set of user-defined situational awareness categories in real time. Vieweg et al. in 2010 [151] analyzes the Twitter logs for pairs of concurrent emergency events, for example, the Oklahoma grassfires (April 2009) and the Red River floods (March and April 2009). Their automated framework is based on the relative frequency of a geo-location and location-referencing information from the users' posts.

In a related work in 2014, Vieweg et al. [150] has enabled filtering, searching, and analyzing of tweets during another natural disaster (the 2013 typhoon Yolanda). The researchers used a supervised classification algorithm to automatically classify the tweets into three categories, labelled Informative, Not informative and Not related to this crisis. Then they employed topic modelling using the LDA [20] model to further classify the informative tweets into 10 clusters according to the Humanitarian Clusters Framework. Similarly, Twitinfo [87] automatically detects and labels unusual bursts in real-time Twitter streams. However, TwitInfo adapts signal processing and streaming techniques to extract peaks and label them meaningfully, using text from the tweets. Additionally, Olteanu et al. [103] created a lexicon of the crisis-related terms (380 single-word terms) that frequently appear in relevant messages posted during six crisis events. Then they demonstrated how the lexicon can be used to automatically identify new terms by employing pseudo-relevance feedback mechanisms to extract crisis-related messages during emergency events.

In research that was mostly analytical, Shamma et al. [136] presented Tweetgeist for identifying structure and semantics in Twitter about media events and sending such information back to the microbloggers to enhance their experience. However, most of the current disaster identification approaches are desinged to detect certain events, such

as earthquakes, tornadoes, etc. and may struggle to detect other non-disaster related events. Another presumption of these approaches is that users have to know the event in advance to represent the keyword quires to be detected. In addition, Thapen et al. [144] built a situational awareness system that uses frequency statistics and cosine similarity based measures to produce terms characterizing localized events (the detection of an outbreak of illness, in their case) and then retrieve relevant news and representative tweets.

It may also be useful to report some studies that have been proposed to identify event phases and the temporal boundaries of mass disruption events. For instance, Chowdhury et al. [32] introduced a system called Tweet4act to automatically determine different phases of an event by extracting content features from each message. They applied the popular k-mean clustering algorithm to classify messages for three crisis events (the Joplin tornado in USA, the Nesat typhoon in the Philippines and the earthquake in Haiti). Similarly but with a broader perspective on events, Iyengar et al. [60] described an approach to automatically determine when an anticipated event started and ended by analyzing the content of tweets using an SVM classifier and hidden Markov model with various textual features, such as bag of words, POS (part-of-speech) tags, etc. Both studies aim to automatically classify tweets according to the three phases of an event: before, during, and after. Additionally, Yin et al. [164] investigated several approaches that have been shown useful for analyzing event to a local level the Twitter messages generated during humanitarian crises. Three key relevant methods for burst detection were evaluated: a) tweet filtering and classification, b) online clustering, and c) geotagging.

In summary and from the above review, although many approaches exist for the task of event detection in social media, they are generally used for large scale events such as large-scale event detection systems and hence are not designed to capture important small-scale events. On the other hand, the small-scale event detection systems are very specific and are limited to detecting some events alone - thus missing the context

of larger events. Moreover, and regarding the use of social media data in disasters, researchers have proposed several event detection systems for disaster and emergency events aiming at filtering, searching, and analyzing tweets during disasters in real-time. Many of these approaches require a priori knowledge of the number of events, or require manually selected terms or phrases to retrieve documents for each event, or are limited to specific type of event (i.e., concerts), or are limited to specific language particularly the English language.

Overall, a common and serious limitation of both, large-scale event detection and small-scale event detection approaches is the real-time elements of these approaches to identify events in social media. Together with the limitations of traditional approaches and disaster event detection approaches, these constitute the motivation behind the research work in this thesis. Hence, we need to propose a general online classification-clustering framework, suitable for large-scale and small-scale social media content, in real-time and for multiple languages. We also need to develop new techniques for selecting top messages that represent an event with high quality, strong relevance and are useful to people looking for information about an event.

In contrast to the above approaches, we propose a novel approach to event detection that aims to overcome many of the above limitations and challenges to provide a system to detect large-scale events and related small-scale events. Our integrated event detection framework consists of five main components: data collection, pre-processing, classification (supervised machine learning algorithms), online clustering (unsupervised learning), and summarization (representation). The proposed system automatically identifies real-world events and disruptive events in a particular time and location. We propose the use of an online clustering algorithm with a sliding window timeframe which can be used to detect large and small-scale events from social media streams - with particular attention to filtering from large to small-scale events.

To this end, we investigate the deployment of a supervised classification model to classify each tweet into 'event class' or 'non-event class' before clustering, as an approach

for reducing computational overhead at the clustering stage by potentially significantly reducing the number of tweets to those containing only event related tweet. Thus clustering, feature selection, and summarization could become much faster and more suitable for real-time analysis. In addition, we validate the overall model performance of our integrated framework on several events and using multiple Twitter datasets to show the effectiveness of the framework. We evaluate it against different leading event detection approaches, described above, including spatial LDA [106], unsupervised methods [16], [167], topic models [148] and a graph-based approach [131]. We also presented a case study of our approach, evaluating it against other leading approaches using Twitter posts from the UK riots in 2011 which is different because this evaluation is based on high quality ground truth data from public Metropolitan Police Service (MPS) reports.

2.4 Data Collection

Since we receive a high volume of widely varied tweets per day, traditional monitoring and analyzing is impractical and moreover it significantly reduces the set of potentially applicable real-time algorithms. We collected user-generated updates directly from the social media using the streaming API (Application Programming Interface) because it allows us to subscribe to a continuous live stream of data. Our goal is to detect events in a given location without any prior knowledge of these events. Generally, Twitter offers two application programming interfaces (APIs) for collecting tweets: one is the search API, which is used to retrieve past tweets matching a user specified criteria; the other is the streaming API, which is used to subscribe to a continuing live stream of new tweets matching a user defined criteria and delivered to the user as soon as they become available. The most open and widely used social medium is Twitter. Other social media sites, such as Facebook and Google+, are largely closed and access to public online discourse is more restricted. This is due to that different social networking sites are used for different purposes, but commonalities do exist. For

instance, the top 3 activities on Twitter are to (1) post about daily activities, (2) upload and share photos, and (3) comment on posts of others; while on Facebook they are to (1) upload and share photos with friends, (2) message with friends on a one-on-one basis, and (3) comment on posts of friends [57].

Terrestrial events by definition occur in space and thus we collected tweets based on a set of keywords describing a country or region (e.g., Iraq, Syria, Egypt, ...) using different languages. We also collected tweets from users who selectively added the required region as their location in their profile metadata or turned on the GPS option on their smartphones. Although Twitter users have the option to enable location services on their account, this feature is off by default and requires users to opt in, but once it is enabled users can geotag their tweets with precise location data in the form of latitude and longitude [139]. Geotagging tweets means that the exact position of where the tweeter was when the tweet was posted is recorded using longitude and latitude measurements. Although the proportion of geotagged tweets with precise location is small [139], these samples of tweets actually contain very valuable information for the purpose of disruptive event detection . Details of the spatial features and geotagging can be found in Chapter 4. Finally, we made use of geographic Hashtags (e.g., #Ramadi, #Aleppo, #Cairo, #Dubai ...) in the data collection process.

The data were stored temporarily using a MongoDB database, which is open-source, easy to use and boasts high access speed and memory. In addition, MongoDB is suitable for storing short texts and supports different indices with a standardized querying interface [70, 8]. We stored all the collected tweets for 24 hours and then we released them. In the following section we describe the various datasets that were used in the course of the thesis.

2.4.1 F1 Twitter Corpus

In this study, our first dataset, which consists of around 1.7 Million tweets (1698517),

was collected from 15 October 2013 to 05 November 2013 using Twitter's Streaming API because it allows us to subscribe to a continuous live stream of new data. Our initial aim was to monitor and analyze the disruptive events in a particular region that were associated with major occasions. In this case, we chose (FORMULA 1 GRAND PRIX 2013) as the occasion. It was hosted in Abu Dhabi between 1st and 4th November 2013, but we extracted data for 15 days before the start to identify also the differences in the Twitter sports messages reported before the event and during it, so as to train the online clustering algorithm and set the thresholds. The number of Arabic tweets is 890,658 where English tweets are 439,191. Around 22% of tweets were published in other Latin script and other languages. Nearly 16,850 unique hashtags appear in the first dataset involving roughly 674,000 unique users.

We collected tweets based on a set of keywords that describe Abu Dhabi and sport in general in different languages, mainly in Arabic and English. We also collected tweets from users who chose to add Abu Dhabi (or the surrounding cities in the UAE) as their location. Figure 2.2 shows the tweets volume in Abu Dhabi, which clearly indicates the rise of sports messages posted during the F1 event. Figure 2.2 also shows an increase in the total frequency of all tweets in Abu Dhabi for the F1 period because of its popularity and due to the various associated events such as financial events, entertainment events, disruptive events, etc. (This dataset is used in [8, 6, 7, 9]).

2.4.2 Abu Dhabi Twitter Corpus (Crime detection)

One of the main objects of interest in the present work is the task of disruptive event detection in the Arabic language; for this reason, we restricted this dataset to Arabic tweets and eliminated all the non-Arabic ones. We collected tweets from the Abu Dhabi area for the interval between 26th November 2014 and 8th December 2014. This dataset consists of 1,161,854 Arabic tweets with approximately 13,400 unique hashtags and roughly 590,000 unique users.

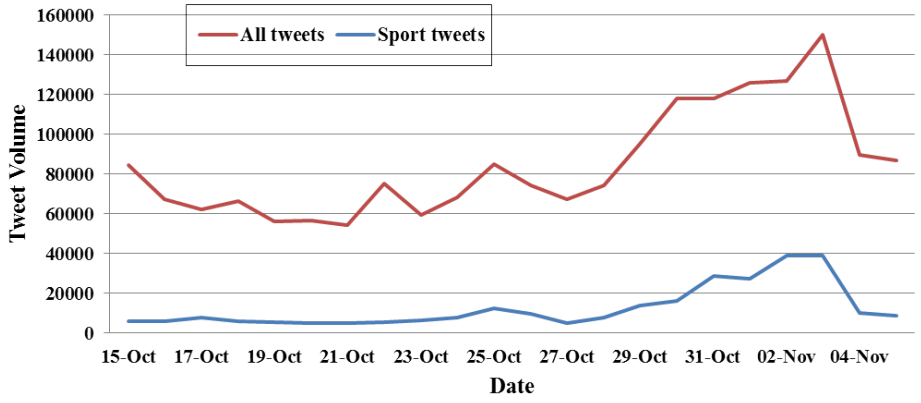


Figure 2.2: The volume of tweets in the first data set from 15th October to 5th November in Abu Dhabi.

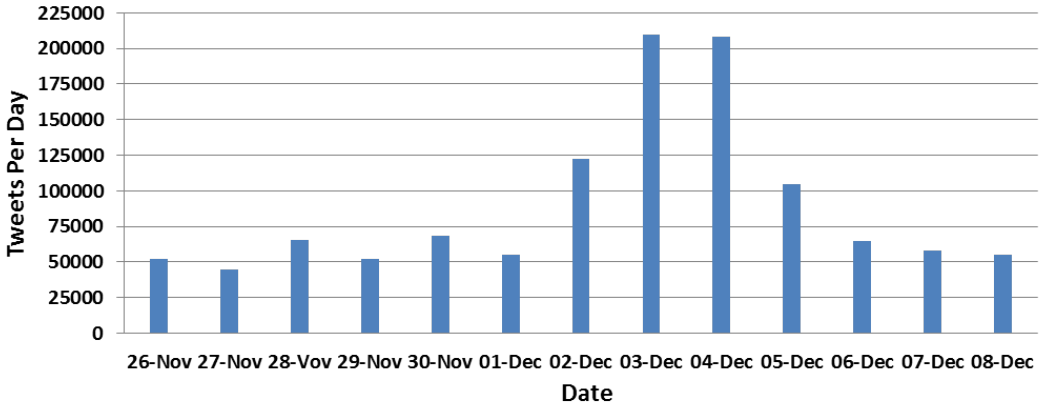


Figure 2.3: The volume of tweets in the second data set from 26th November to 8th December in Abu Dhabi.

A considerable change in the volume of tweets was noted from 2nd to 5th December 2014 due to the double crime (considered to be a terrorist attack) on 2nd December 2014 which was unprecedented in the nation's peaceful history. An American woman was murdered in a shopping mall and the woman suspect in this murder planted a primitive bomb on the doorstep of an American citizen in a different location; at this point she was held. Figure 2.3 shows the volume tweets in Abu Dhabi between 26th November 2014 and 8th December 2014 where we notice the rise in the volume tweets from 2nd to 5th December 2014. (This dataset is used in [9, 6]).

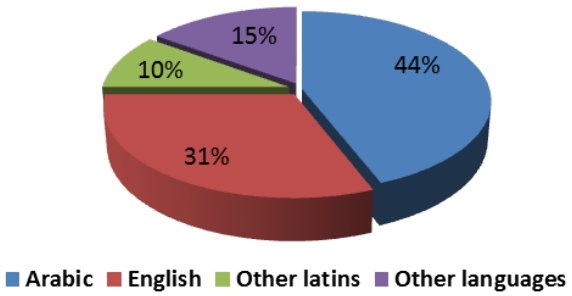


Figure 2.4: The distribution of languages used in the Middle East dataset

2.4.3 Middle East Twitter Corpus

This dataset consists of 40 million tweets and was collected between 1st October 2015 and 30th November 2015 using Twitter's Streaming API. This is a general collection of tweets and is used here to show that our event detection system is useful for extracting information from socially-generated content on a broad range of topics. Our aim was to predict and analyze events and disruptive events, so we extended the geographical search location to the Middle East region, collecting tweets from users who chose one of the Middle Eastern countries as their location. We addressed this query using the center coordinates of the Middle East area of 29.298o latitude and 42.551o longitude. Nearly 425,000 unique hashtags appear in the 40 million tweet corpus involving roughly 18,000,000 distinct user accounts. Figure 2.4 shows the language distribution of the third dataset. (This dataset is used in [12]).

2.4.4 MediaEval2012 Flickr Corpus

In order to further validate our approach and to test the applicability of our framework on different social media sites such as Flickr [29], an Internet image community website, we evaluated it against other approaches in the context of the MediaEval2012 Social Event Detection (SED) international benchmark [109]. The SED competition comprised three challenges to a common test dataset of images with their metadata

(timestamps, tags, geotags). The associated timestamp is the time at which the image was published. The goal of the first challenge in the test collection was to identify public technical events, such as exhibitions and fairs that had taken place in Germany. The goal of the second challenge was to find all the soccer events that had taken place in Hamburg (Germany) and Madrid (Spain) (i.e. soccer events and celebrations). The goal of the third challenge was to find demonstration and protest events involving the Indignados movement occurring in public places in Madrid and all over Spain, which were related to the financial crisis and national politics.

SED provided 167,332 photos collected from Flickr.com that had been captured between 2009 and 2011. All the photos were originally geotagged. However, before supplying the XML photo metadata archive (including any tags, geotags, time-stamps, etc.) to the task participants, the geotags were removed for 80% of the photos in the collection (randomly selected). Although the SED dataset included photos augmented with metadata, we focused on textual metadata, in order to treat all the photos as documents. We consider the use of visual information from our algorithm to be included in future work. (This dataset is used in [12, 10]).

2.4.5 2011 Riots in England

Our final dataset consists of 1.6 million tweets and was generated during the 2011 riots in England, which began as an isolated incident in Tottenham on 6th August but quickly spread to other parts of London and other cities in England and gave rise to levels of looting, destruction of property and violence not seen in England for more than 30 years [92]. This event was selected because of a publicly available record of the intelligence and incidents reported during this period that offers a gold standard evaluation dataset. Data were purchased from the Twitter reseller Gnip from 6th August until 12th August 2011 using the following query #londonriots OR #tottenham OR #enfield OR #birminghamriots OR #UKRiots OR #Croydon OR #hackney OR #tottenhamriots OR #tottenhamshooting OR #Londonriots OR #riotcleanup OR #rioting OR

#manchesterriots OR #liverpoolriots OR #bullring OR #enfieldriots OR #croydonriots OR #Londonsburning OR #prayforlondon. We selected the most popular hashtags that had attracted the attention of most users. They are reflected as peaks in the tweeting rates which use these hashtags. In the process of selecting these hashtags, the system considered only sudden increases in the recent tweeting activity that used these hashtags. (This dataset is used in [10]).

2.5 Data Pre-processing

Due to the conversational and messy nature of tweets, the data needed , cleaning before proceeding with analysis. The goal of this step is to represent data in a form that can be analyzed efficiently and to improve their quality by reducing the amount of trivial noise (i.e. deleting posts that are irrelevant to the events). We employed traditional text processing techniques, such as stop-word elimination and stemming. Moreover, posts that were less than 3 words long were removed, as were messages where over half the total words were the same word, since these posts were less likely to have useful information.

- **Stop-word Elimination**

  Among many words, some words are too frequent to function as useful feature since in addition they do not convey any semantic significance to the texts or phrases that they appear in. For example, the verbs "be" and article "the" can be seen in almost all documents. Such words are called stop-words and are often removed from the feature set [15]. One problem in stop-word elimination is that a word can be a stop-word for a data set, but can also be a useful feature for another data set. Therefore, we used the classical stop-word list as well as deriving our own stop-list. Term frequency (TF), Inverse Document Frequency, (IDF), and TF-IDF are the criteria used for classifying stop words (some of which are Twitter-specific stop words, such as "lmao"). The TF-IDF is the best criterion of

the three approaches inspired by [81]. This was expected, since IDF is more reliable than TF alone and the TF-IDF is an additional refinement of IDF. Similarly, we then added more stop words which were determined under the same criteria (TF, IDF, and TF-IDF) as those of the training corpus, to create an Arabic stop word list in addition to the Khoja stemmer [41] stop word list. The resulting list of the Arabic stopword list consisted of 1,377 words and the English stopword list consisted of 441 stop words.

- **Stemming**

  Stemming is a pre-processing step in text mining applications in general, as well as being a very common requirement of Natural Language Processing (NLP) systems and Information Retrieval (IR) systems. Stemming is the process for reducing a derived word to its stem, base or root form. It is usually done by removing all affixes (suffixes and prefixes) attached to index terms. Stemming and lemmatizing are similar in that stemming involves reducing a word variant to its 'stem' and lemmatizing involves reducing a word to its 'lemma'; however, stemming is done by applying a set of rules to a word in its context but ignoring its context and what part of speech (POS) it is [61]. In contrast, lemmatizing requires obtaining the 'lemma' of a word, which involves first understanding the POS and then reducing the word to its root form. Here we focus on stemming only, but in the next section we investigate the effect of POS tagging on short text classification.

  As shown in [115], stemming is not a process to apply to all languages. It is not applicable to Eastern languages, for example, Chinese. In this work we use the Khoja stemmer [41] for Arabic tweets and the Porter stemmer [114] for English and other Latin posts. For the remaining European languages, we implement the Snowball Stemmer [115]. We do not use stemming for any Eastern languages, such as Japanese.

2.6 Summary

In this chapter, we presented various definitions of events in the literature over a variety of domains and connected these definitions to our task of identifying and characterizing events in social media. Then, we provided an overview of the event detection literature, targeting mainly those works on social media event detection, since detecting and analysing events from social networking sites is the main focus of this thesis. A literature review of the event detection techniques and systems was provided with particular focus on key existing works on large-scale event detection systems, small-scale event detection systems, as well as systems that have been used in disasters and emergency events. Discussion on the main strengths, limitations and gaps in the state-of-the-art approaches that are relevant to the task of event identification in social media was given.

Then, we provided a short description of the first two stages of our proposed framework; data collection and pre-processing. In data collection, we provided a listing of each dataset that we used in this thesis and detailed their statistics, including three large-scale, real-world datasets from Twitter and one dataset from Flickr. These datasets are used in the subsequent chapters to evaluate each component of the proposed framework (classification, clustering, feature selection, and summarization). Furthermore, we explored the traditional preprocessing steps to reduce the amount of noise in the tweets such as stop-word elimination and stemming.

Chapter 3

Classification and Categorization

3.1 Introduction

In the previous chapter, we discussed the first two stages of our proposed framework; data collection and pre-processing where we introduced the various datasets that we used in this thesis and then we briefly mentioned the preprocessing steps that we have used in our framework such as stop-word elimination and stemming. In this chapter, we begin by outlining our classification-clustering event identification framework (Section 3.2) and then we briefly discuss the different elements of the event identification framework.

Section 3.3 focuses on the third step of our proposed framework, event classification. We describe related efforts on event classification using supervised machine learning algorithms. Then we use three supervised machine learning methods, (Naive Bayes, Logistic Regression, and Support Vector Machines) in order to automatically separate the event-related messages in a stream of tweets from all the non-event content. We also evaluate and compare the performance of these methods according to standard evaluation metrics in Section 3.4. Finally, we investigate ways to improve the performance of the classification results using many types of features such as the presence of n-grams or their frequency, the use of unigrams, bigrams and trigrams, linguistic features such as parts-of-speech (POS) tagging and Named Entity Recognition (NER).

The research question we aim to address in this chapter is:

RQ1 Can we detect "events" in real time from the streaming media and what are the best features for the event classification task?

3.2 Proposed Event Identification Framework

The social media generally produce a large number of posts per hour on a wide variety of topics, rendering human monitoring impractical as well as significantly reducing the set of applicable real-time algorithms. In this thesis, we propose a novel approach to detecting the small and large events from social media sites. Figure 3.1 shows the framework, which allows the automatic identification of meaningful events from the social media in real-time for a specific time and place. The five-step framework consists of data collection, pre-processing, classification, on-line clustering and summarization. The clustering step is divided into two sub-steps: on-line clustering and feature selection. The proposed framework aims to collect data (step 1) over certain time windows for a given location which is supported by the automatic detection and summarization of events from social media. Tweets are usually composed of incomplete, noisy and poorly structured sentences due to the frequent presence of abbreviations, irregular expressions, ill-formed words and non-dictionary terms [13, 57]. Therefore, the pre-processing (step 2) applies some pre-processing techniques to reduce the amount of noise in the tweets and consequently reduce its possible negative impact on the event detection task, as was shown in the previous chapter.

Classification, on-line clustering, feature selection, and summarization are the main elements of our proposed framework. These elements together enable us to perform the challenging task of identifying events and their associated documents over social media streams. Information describing events from those who report them in order to gather information about the ongoing events in a given area can be critical in many situations.

We use classification models (step 3) in our framework to identify event tweets based

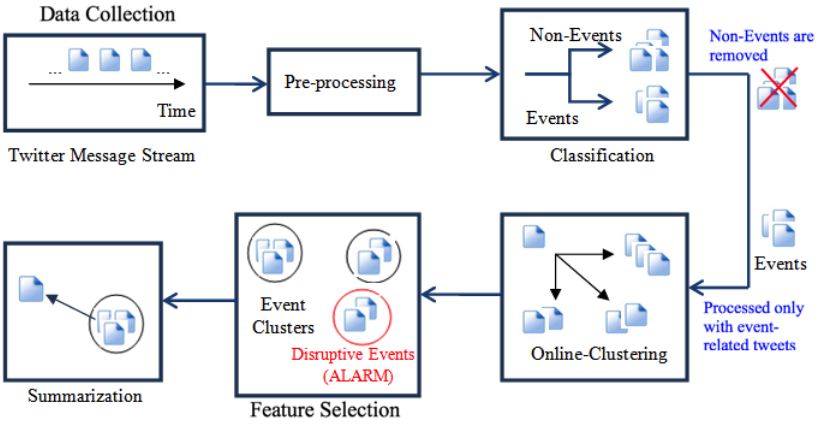


Figure 3.1: Event Detection Framework for Social Media Content

on the textual content of the tweets. Each tweet is represented as a single document and the TF-IDF weights of textual terms are used as features to train the classifier. Three supervised machine learning algorithms (Naive Bayes, Logistic Regression, and Support Vector Machines) are presented and evaluated for the microblog classification of events. Classifying before the step of on-line clustering reduces the computational overhead at the clustering stage because the number of tweets is significantly reduced once the non-event tweets have been removed. Then the online clustering algorithm (step 4) that we discuss in Chapter 4 is used to group together topically similar event tweets and identify the topic that they share.

Our proposed clustering algorithm is suitable for the social media domain and employs a similarity metric technique to exploit a variety of textual and non-textual features. For each identified event cluster, which may contain hundreds of tweets, we summarize each event cluster by selecting the messages that best represent it, using one of the summarization approaches (step 5) that we have developed (see Chapter 5). In this chapter, we focus on using supervised machine learning algorithms (step 3) in order to automatically separate the event-related messages in a stream of tweets from all the non-event content.

3.3 Machine Learning approaches

Our aim in using supervised machine learning is to build a model that makes predictions about future instances based on the evidence of instances supplied, in the presence of uncertainty [68]. Supervised learning is splitted into two broad categories: classification or regression. Text Categorization (TC) or classification is the problem of labelling natural language texts with one or more thematic categories drawn from a predefined set [152]. In regression, the goal is to predict a continuous measurement for an observation. That is, the response variables are real numbers. Statisticians use the word regression for the process of predicting a numeric quantity [159].

Text categorization (or classification) of texts into topical categories has a long history, dating back at least to the early 1960s. Until the late 1980s, the most effective approach to the problem seemed to be that of manually building automatic classifiers by means of knowledge engineering techniques, i.e. manually defining a set of rules encoding expert knowledge of ways to classify documents according to a given set of categories [134]. In the 1990s, with the booming production and availability of on-line documents, automated text categorization has witnessed a renewed and increased interest, which has prompted the machine learning paradigm to construct automatic classifiers to emerge to the point where it has certainly superseded the knowledge-engineering approach.

In the last ten to fifteen years, automated content-based document management tasks have gained prominence in the information systems field, largely due to the widespread and continuously increasing availability of documents in digital form [134, 159]. Varieties of social media, as a form of this digital Internet technology, are now emerging rapidly. People are using social media platforms such as Twitter, Facebook and YouTube to interact with others, share information and report real-life events.

A number of recent studies have used different techniques to address the classification of tweets for a variety of purposes, such as sentiment analysis [39, 107], news detection

[129, 113], event characterization and information extraction [124, 127] and many other applications. But classifying the texts generated by the social media faces many major challenges. Short texts have an adverse effects on results because they contain very few words (features), sometimes too few for the classification task. Moreover, since the number of tweets is high and the nature of streaming data is dynamic, the classification process must be quick enough to cope with the real-time requirement. Another challenge is that tweets are not as well-written nor thoughtfully composed as articles or blogs. A third challenge is, of course, noise (in the form of typos or less situation-specific tweets) is conveyed through the microblogging services which are hard to discover automatically and sometimes requires human judgment.

Many of the existing text classifiers represent a tweet according to a bag of words (BOW) approach, and then use machine learning techniques over vectors whose features are derived from the occurrence frequencies of these words. One classification method discovers rules from certain training sets with known classes, and then uses these rules to predict new classes. For instance, Sriram et al. [141] classified tweets under a predefined set of five generic classes (news, events, opinions, deals, and private messages) in order to improve information filtering. Moreover, Sankaranarayanan et al. [129] built a news processing system that identified the tweets corresponding to late breaking news. They used and trained a Naive Bayes Classifier to classify incoming tweets as either news or junk (not news) to improve the quality of news tweets. Sakaki et al. [127] trained a classifier to recognize the features derived from individual tweets (e.g., the keywords in a tweet and the number of words it contained) and detect a particular type of event, such as an earthquake or a typhoon. They formulated event detection as a classification problem and trained a Support Vector Machine (SVM) on a manually labelled Twitter dataset comprising positive events (earthquakes and typhoons) and negative events (other events or non-events).

Much research has gone into the area of sentiment classification. For example, Go et al. [49] introduced an approach for automatically classifying the sentiment of tweets with

emoticons, using a range of machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) in distant supervised learning. In addition, Pang and Lee [107] researched the performance of various machine learning techniques (Naive Bayes, maximum entropy, and SVMs) in the specific domain of movie reviews, to separate positive from negative reviews. In a similar context and for the same purpose, Saif et al. [126] introduced an approach to add semantics to the training set as an additional feature. They incorporated semantic features into Naive Bayes (NB) model training, using an interpolation approach. Adopting a similar semantic approach but with external knowledge (Wikipedia), Gabrilovich and Markovitch [48] proposed a semantic analysis based on manifest topics grounded in Wikipedia pages. The mapping between each short text and the Wikipedia topics was carried out through a feature generating labelled categories. Support vector machines (SVMs) were implemented as the learning algorithm to build these text categories.

The next step in our framework is classification (step 3) which distinguishes events from noise or irrelevant posts. Generally, classification aims to assign text documents to pre-defined classes. A traditional example would be to automatically label each incoming news story with a topic such as "politics", "sports", or "art". We used supervised machine learning models to separate "event" and "non-event" content on social networking services such as Twitter. In these highly interactive systems members of the general public can post real-time reactions to real world events - thereby acting as social sensors of terrestrial activity. However, non-event content is, of course, prominent on Twitter and similar systems, where people share various types of content, such as personal updates, random thoughts and musings, opinions, and information [16]. Our method provides an efficient way of accurately categorizing event tweets without the need of external data, enabling decision makers to discover event related posts in real time.

The classification step aims to identify event related posts and omit non-event tweets. It subsequently reduces the number of posts to be processed in the following steps

(clustering and summarization) because these steps process only event-related tweets. Our aim in this chapter is to examine whether it suffices to treat event classification in Twitter simply as a special case of topic categorization (with the two "topics" of events and non-events). Then we discuss and experiment with three standard machine learning algorithms for the purpose of identifying tweets that are about events and those that are about non-events. We chose three well-known machine learning algorithms; Naive Bayes classification [75] a statistical classifier based on Bayes' theorem, Logistic Regression [47], a generalized linear model for applying regression to categorical variables, and support vector machines (SVMs) [62], which aims to maximize (maximum margin) the minimum distance between two classes of data using the hyperplane that separates them. We have chosen these classifiers based on the size, nature and the quality of the data. In addition, these classification methods belong to different and distinct supervised classification methods (Generative vs Discriminative). Furthermore, we seek to investigate ways to improve the performance of the classification results; thus we consider those features which capture patterns in the data, such as the presence of n-grams or their frequency, the use of unigrams, bigrams and trigrams, linguistic features such as parts-of-speech (POS), tagging and Named Entity Recognition (NER).

3.3.1 Naive Bayes Classifiers

Naive Bayes Classifiers are a family of simple probabilistic generative classifiers based on applying Bayes' theorem. Naive Bayes was introduced in the 1950s, was studied extensively in the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization [159]. This kind of probabilistic approach makes strong assumptions about the way in which the data are generated, and posits a probabilistic model that embodies these assumptions; then it uses a collection of labelled training examples to estimate the parameters of the generative model. It has been successfully implemented in many information retrieval and natural language processing tasks, such as sentiment analysis and opinion mining [140, 126], event de-

tection [58, 8] and many others [24, 85].

The main reason for using the Naive Bayes model is that despite its simplicity, it has been shown to be a very powerful model. The Naive Bayes model has many advantages, for example, that it is relatively fast to compute, easy to construct and with no need for any complex iterative parameter estimation schemes. Unlike Support Vector Machines (SVMs) or Logistic Regression, the Naive Bayes classifier treats each feature independently. Naive Bayes also tends to do less overfitting compared to Logistic Regression [129]. However, the strong assumption of conditional independence between features reduces the power of Naive Bayes.

In our case, tweets are characterized by the words that appear in them, and a tweet t is represented as a vector $w_1, w_2, ..., w_k$ of (binary or weighted) terms/words. The probability that t is an event is denoted by $P(E|w_1, w_2, ..., w_k)$, which can be rewritten as follows using Bayes' theorem:

$$P(E|w_1, w_2, ..., w_k) = P(E).\frac{P(w_1, w_2, ..., w_k|E)}{P(w_1, w_2, ..., w_k)} \tag{3.1}$$

Similarly, given a tweet t, the probability that it is a non-event tweet is given by $P(N|w_1, w_2, ..., w_k)$, which can also be rewritten using Bayes' theorem:

$$P(N|w_1, w_2, ..., w_k) = P(N).\frac{P(w_1, w_2, ..., w_k|N)}{P(w_1, w_2, ..., w_k)} \tag{3.2}$$

Using the assumption of independence among the words in t as well as our prior calculations of $P(E)$, $P(N)$, $P(w_i|E)$, and $P(w_i|N)$, we introduce the threshold (D) :

$$D = log\frac{P(N|w_1, w_2, ..., w_k)}{P(E|w_1, w_2, ..., w_k)} = log\frac{P(N)}{P(E)} + \sum_{i}^{k} log\frac{P(w_i|N)}{P(w_i|E)} \tag{3.3}$$

If $D < 0$, then the tweet is classified as an event. Otherwise, the tweet is classified as a non-event and discarded. To train and test the classifier we use human annotators to manually label 5000 randomly selected tweets as belonging to one of two classes, "Event" and "Non-Event", which were collected at five different hours in the first and

second weeks of October 2015 (first dataset, Section 2.4.1). These five hours were sampled uniformly at random from five bins partitioned according to the volume of messages per hour over these two weeks. To ease the annotation process, examples were shown to the annotators along with their respective classes. The details of the annotation process are expanded in Section 3.5.1.

The agreement between our three annotators, measured using Cohen's kappa coefficient, was substantial (kappa = 0.807) [34, 26]. Training data (tweets) were transformed into feature vectors and their corresponding category (event or non-event) was provided to the classifier, constituting the training set. From the training data the likelihood of each post's belonging to either class was derived on the basis of feature occurrence in the training data. Whenever a new example is presented, the class likelihood for the unseen data is predicted on the basis of the training instances.

Algorithmic steps:

1. Input posts.
2. Extract features from posts.
3. These features and their corresponding labels are used to train the learning algorithm.
4. New posts are presented to the trained classifier to predict their label according to their extracted features.

3.3.2 Support Vector Machines

Support vector machines (SVMs) are supervised discriminative learning models which are based on the maximization of the margin between the instances and the separation hyper-plane. They can be used for classification and regression analysis and have also been shown effective in linear and non-linear classification (using a kernel function)

[51]. The application to TC of the support vector machine method has recently been proposed by Joachims [62]. SVMs have been shown to be highly effective in traditional text categorization [134, 107]. In addition, they have performed well for short texts [24, 58, 127, 164]. Although the training time of even the fastest SVMs can be extremely slow, they are highly accurate, owing to their ability to model complex nonlinear decision boundaries [51]. SVMs classify unseen data by maximizing the distance between classes of similar data points created using training data and find the best place for new data points; they have been particularly useful for text classification [24].

When two classes are involved [107], the basic idea behind the training procedure is to find a hyperplane for separating the two classes which might be represented by vector $\overrightarrow{w}$, and written as

$$\overrightarrow{w} = w_0 + w_1\partial_1 + w_2\partial_2 \tag{3.4}$$

where ∂_1 and ∂_2 are the attribute values and there are three weights w_i to be learned.

The hyperplane vector $\overrightarrow{w}$ not only separates the document vectors in one class from those in the other, but finds those for which the separation, or margin, is as large as possible. The equation defining the maximum-margin hyperplane can be written in another form [159], in terms of the support vectors (as an optimization problem). The class value y of a training instance can be written as either 1 (for yes, it is in this class) or -1 (for no, it is not). Then the maximum-margin hyperplane can be written as

$$\overrightarrow{w} = b + \sum_i \alpha_i y_i \partial(i).a \tag{3.5}$$

Here, y_i is the class value of the training instance $\partial(i)$, while b and α_i are numeric parameters that have to be determined by the learning algorithm. Note that $\partial(i)$ and a are vectors. Vector a, for instance, represents a test, just as vector $[\partial_1, \partial_2]$, represented a test instance in the earlier formulation. The vectors $\partial(i)$ are the support vectors that

are selected as members of the training set. The term $\partial(i).a$ represents the dot product of the test instance with one of the support vectors: $\partial(i).a = \sum_j \partial(i).a_i$.

Finally, b and α_i are parameters that determine the hyperplane, just as the weights $w_0 + w_1 + w_2$ are the parameters that determined the hyperplane in the earlier formulation. The classification of test instances consists simply of determining which side of the hyperplane $\overrightarrow{w}$ they fall on [159].

3.3.3 Logistic Regression

Logistic Regression is a statistical discriminative model that measures the relationship between the two or more independent variables by estimating probabilities using a logistic function [47]. In a two-class case of logistic regression, the model should perform a regression for each class, setting the output equal to 1 for the training instances that belong to the class (in our case, an Event) and 0 for those that do not (a Non-Event). One way of looking at this linear regression is to imagine that it approximates a numeric membership function for each class. The membership function is 1 for instances that belong to this class and 0 for other instances. Given a new instance, we calculate its membership for each class and select either class 0 or class 1 [159]. Linear regression performs a least-squares fit of a parameter vector [72] β (β is the coefficient vector which we learn from the training set) to a numeric target variable so as to form a model

$$f(x) = \beta^T.x \tag{3.6}$$

where x is the input vector (x is basically our feature set). However, this linear approach is known to suffer from masking problems [72, 159]. A better method for classification is that of linear logistic regression [72, 107], which models the posterior class probabilities

$$h_\beta(x) = \frac{1}{1 + e^{-\beta^T x}} \tag{3.7}$$

The decision boundary for two-class logistic regression lies where the prediction probability is 0.5. Hence, an example is classified as 0 if the value of $h_\beta(x)$ is less than 0.5 and is classified as 1 if the value of $h_\beta(x)$ is greater than 0.5. Note that this model still produces linear boundaries between the regions corresponding to the different classes in the instance space.

Logistic regression, unlike Naive Bayes, holds no strong assumption of conditional independence between features. However, the drawback of this algorithm is that it is not very stable and requires a large number of training data to give good results; producing low variance but potentially high bias [72, 107].

3.4 Summary of Features

One of the factors, which affects the performance of event classifiers, is the choice of the features used for classifier training [126]. Many types of features have been used in the tweet classification task, including (i) feature frequency and presence, (ii) word n-gram features, (iii) Part- Of-Speech tags (POS) features, and (iv) Name Entity Recognition (NER). The use of these features is the main focus of this section.

Some researchers have reported that best performance is achieved using unigrams (single word features) [107, 49], while other works report that bi-grams and trigrams outperform unigrams (two and three word combinations) [24, 39]. However, they are agreed that term-presence gives better results than term frequency. For instance, [39] shows that the presence of words only once in a given corpus is a good indicator of higher precision. Linguistic analysis has also been used on machine learning features to try to improve results. For example, part-of-speech (POS) tagging, a basic form of syntactic analysis, has been used to disambiguate the sense in many applications of natural language processing (NLP), while Named Entity Recognition (NER) is used to extract from a given corpus the proper names or entities, such as those of persons, organizations, and locations.

We believe that the key to achieving success in text classification by machine learning is feature selection. Therefore, we consider different features which capture patterns in the data, such as n-gram presence or n-gram frequency, the use of unigrams, bigrams and trigrams, linguistic features such as parts-of-speech (POS) tagging and Named Entity Recognition (NER) and we investigate which of these features could improve the performance of the classification results. In this thesis, we use the Stanford POS tagger (http://nlp.stanford.edu/software/tagger.shtml) because it has an English tagger model, as well as other models for different languages including Arabic, Chinese, French and German.

For the Named Entity Recognition (NER), we use TwitIE pipeline (An Open-Source Named Entity Extraction Pipeline for Microblog Text), which is a modification of the GATE ANNIE open-source pipeline for traditional (i.e. news) text [22]. GATE uses gazetteer-based lookups and finite state machines to identify and type named entities in newswire text. TwitIE is available to download from https://gate.ac.uk/wiki/twitie.html, usable both via the GATE Developer user interface and via the GATE API [22]. We have used the first dataset (F1 Twitter Corpus, see Section 2.4.1 for details) to evaluate various machine learning approaches. The F1 dataset consists of 38,364 English named entities (23,115 unique entities) and 15,712 Arabic named entities (12,944 unique entities). In one of our baseline models, we build various classifiers trained using a combination of all POS tags and all name entities and use them as baseline model, which is called (POS + NER) model.

From the above discussion, one can notice that although many types of features can be used for training classifiers, one of the questions we aim to answer in this chapter, which set of features or which combination of them is optimal for the task of event classification in Twitter.

3.5 Empirical Evaluation

The aim of these sets of experiments is to select the best classifier out of three machine learning algorithms - Naive Bayes classification, Logistic Regression, and support vector machines (SVMs) - for identifying the events and non-events tweets. Furthermore, we investigate what methods might improve the performance of the classification results; thus, we consider the features which capture patterns in the data such as the presence or frequency of n-grams, the use of unigrams, bigrams and trigrams, linguistic features such as parts-of-speech (POS) tagging and Named Entity Recognition (NER).

We used Weka (http://www.cs.waikato.ac.nz/ml/weka/) for the training and testing classification task. The Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. In addition, it deals with a collection of data mining tasks including testing, analyzing, comparison and the automatic calculation of performance measures. Weka also includes implementations of algorithms for learning association rules, clustering data for which no class value is specified, and selecting relevant attributes in the data [159].

3.5.1 Experimental Setup

Dataset: We use the Twitter Streaming API to collect tweets from 15/10/2013 to 5/11/2013. (See Section 2.4.1 for more details).

Annotations: To train and test the classifier we asked three human annotators to manually label 5000 randomly selected tweets in two classes, "Event" and "Non-Event". Event instances outnumber the non-event ones in the training set, which consisted of 1900 Non-Event tweets and 3100 event-related tweets. In spite of the fact that misclassifying a number of event-related as non-related ones could affect the accuracy of the classifier, it substantially improves the identification of real-world events. A set of instructions and examples was given to the annotators so that they can perform the an-

Table 3.1: The instructions provided to the annotators for the annotation task (Classification), followed by an example tweet.

Instructions: Given a Twitter message, identify whether the message is: (A) Event or (B) Non-Event. Please read the examples and the invalid responses before beginning if this is the first time you are working on this annotation task.
Tweet: **#TrafficUpdate : An #accident has been reported on SMBZ Rd After #GlobalVillage towards Sharjah, please be extra cautious** Overall, the tweet is: ☐ Event ☐ Non-Event

notation task. The instructions that we provided to the annotations are shown in Table 3.1.

In addition, some of the example tweets and annotations (Classes are: Event or Non-Event) that were provided to the annotators are shown in Table 3.2. For additional example tweets and annotations used in this chapter, these can be found in Appendix A.1. Agreement between our three annotators was measured using Cohen's kappa coefficient and was found substantial (kappa = 0.807) [34, 26]. A ten-fold cross validation approach was used to train and test the machine learning methods. For each evaluation, the dataset was split into 10 equal portions and trained 10 times. Every time, the classifier was trained in 9 out of the 10 portions and the tenth was used as test data.

Evaluation Methods: we used standard classification metrics: precision, recall, F-measure and accuracy to measure the effectiveness of the text classification. Generally, the evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, called a 'confusion matrix'. The confusion matrix for a binary classification problem (which has only two classes - positive and negative), is shown in Table 3.3.

Table 3.2: List of example tweets and annotations that were provided to the annotators for the classification task (Classes are: Event or Non-Event).

Tweet	Event or Non-Event
#TrafficUpdate : An #accident has been reported on SMBZ Rd After #GlobalVillage towards Sharjah, please be extra cautious	Event
.@RTA_Dubai holds 4th #Dubai International Project Management Forum in November 2013 http://tinyurl.com/n434h3p	Event
Mohamed bin Zayed holds talks with Bahrain Crown Prince on ways to enhance fraternal ties, GCC coordination efforts, regional developments	Event
Happy Birthday to my brother & one of my favorite collaborators EVER, @Pharrell. Enjoy your day, P!!! https://www.instagram.com/p/BShfSi0hA8O/	Non-Event
The root of all health is in the brain. The trunk of it is in emotion. The branches and leaves are the body. #ZenMoment #HealthyLiving	Non-Event
Wish you were here! It's a glorious Saturday morning in Dubai, time for a stroll before the gallery opens at 10am.	Non-Event

The FP, FN, TP and TN entries have the following meanings:

- The false positives (FP): examples predicted as positive, which are from the negative class.
- The false negatives (FN): examples predicted as negative, whose true class is positive.
- The true positives (TP): examples correctly predicted as pertaining to the positive class.

Table 3.3: The confusion matrix for two-class classification problem

	Predicted Class	
True Class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

- The true negatives (TN): examples correctly predicted as belonging to the negative class.

The precision, recall, F-measure and accuracy measures are widely used in classification. Precision (how often our predictions for a class are correct - a measure of false positives); recall (how often tweets (instances) are classified correctly as the correct class - a measure of false negatives); the F-measure, a harmonic means of precision and recall; and accuracy, the proportion of the correctly classified tweets (both true positive and true negative examples) to the total number of tweets, measuring the overall effectiveness of a classifier. Evaluation measures are given by the following equations:

$$Precision(P) = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{3.9}$$

$$Accuracy(Acc) = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.10}$$

$$F - Measure(F) = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.11}$$

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{3.12}$$

where β is a non-negative real number in the last equation. F-Measure (F) gives equal weight to precision and recall, whereas the F_β measure is a weighted measure of precision and recall. It assigns β times as much weight to recall as to precision. Commonly used F_β measures are F_2 (which weights recall twice as much as precision) and $F_{0.5}$ (which weights precision twice as much as recall) [51].

3.5.2 Experimental Results

The aim of the first experiment was to test the performance of a number of content analysis and machine learning approaches with the intention of identifying the best performing methods for classifying events from real-time microposts. For the second set of experiments, we focused on different features, such as unigrams, bigrams and trigrams, linguistic features such as parts-of-speech (POS) tagging and Named Entity Recognition (NER).

We evaluated our three competing approaches according to standard evaluation metrics. Table 3.4 shows a comparison of the various classifiers with unigram presence, which clearly indicates that the Naive Bayes classifier produces the best results. We believe that the key to achieving success in text classification by machine learning is feature selection. Therefore, we considered different features which capture patterns in the data, such as n-gram presence or n-gram frequency, the use of unigrams, bigrams and trigrams, linguistic features such as parts-of-speech (POS) tagging and Named Entity Recognition (NER) and we investigated which of these features could improve the performance of the classification results. In this work, we build various classifiers trained using different baseline models: (1) word unigrams only is our first baseline model (2) the use of bigrams only as our second baseline (3) the use of trigrams only as our third baseline model (4) the use of a combination of word unigrams and bigrams

Table 3.4: Accuracy, Precision, Recall and F-measure for different classification algorithms.

	Naive Bayes classifier	SVMs Classifier	Logistic Regression classifier
Accuracy	82.13	80.93	76.13
Precision	80.64	79.84	73.91
Recall	86.79	86.54	83.90
F-measure	83.60	83.05	78.30

features as our forth baseline model (5) the use of a combination of NER and POS features (all POS tags and all named entities) as our fifth baseline model (6) Finally, we use a combination of word unigrams, bigrams, NER and POS features (all POS tags and all named entities) as our sixth baseline model.

The classification results from Table 3.5 using bigrams as the feature shows that the performance of Naive Bayes and SVMs classifiers does not improve beyond that of the unigram, but in the case of Logistic Regression a noticeable improvement can be observed. Moreover, the classification accuracies of all three classifiers declined when using trigrams as features, which provides suggestive evidence that the use of n-grams for Twitter classification may not be a good approach due to the limitations on the size of tweets. Hence we eliminated the testing with trigrams and the higher order of n-grams and instead combined unigrams and bigrams in order to improve performance by exploiting their best features. Indeed, the Naive Bayes classifier achieved an accuracy of 83.67%. In addition, we got a boost of approximately 1.3% from using SVMs and an improvement of about 3.3% from using the Logistic Regression classifier. Since our training set size is quite modest, this model is outperformed in the present study by Naive Bayes. This suggests that there is still room to improve the performance of the logistic regression classifier by increasing the number of training examples and annotations.

The use of both part-of-speech (POS) tagging and Named Entity Recognition (NER)

Table 3.5: Comparison of classification accuracies of different classification algorithms over a set of features.

Features	Naive Bayes	SVMs Classifier	Logistic Regression
Unigrams	**82.13**	80.93	76.13
Bigrams	**79.52**	78.18	78.57
Trigrams	72.84	**74.09**	69.97
Unigrams + Bigrams	**83.67**	82.23	79.45
POS + NER	**83.50**	81.92	80.18
Unigrams + Bigrams + POS + NER	**85.43**	83.86	80.22

resulted in better performances because they clarify how words are related to events and they also differentiate between different senses of a word (word-sense disambiguation). The POS and NER approach outperforms the unigrams baseline for all three dataset machine learning algorithms. However, the margin is biggest for the logistic regression model. The final test combines all the successful features (unigrams + bigrams+ POS + NER) which leads to the highest classification accuracy, 85.43%, being achieved by the Naive Bayes classifier.

3.6 Summary

In this chapter, we explored several supervised machine learning algorithms (Naive Bayes, logistic regression, and support vector machines), which are able to detect real-world events using data collected from Twitter. We were able to obtain accurate results for the event classification through the choice of appropriate algorithms and features. We treated the problem as a classification task. First, we trained a binary classifier with positive and negative examples of messages which were event-related or not. To build the classifiers, we investigated a wide spectrum of features to determine which ones could successfully be used as predictors of event posts. The experimental results

show that, of the three machine learning classifiers, Naive Bayes overall performed best, using a combination of all the successful features (unigrams + bigrams+ POS + NER).

Our results show that combining linguistic features (POS+NER) with word unigrams and bigrams outperforms the baseline model trained from unigrams only across all three machine learning classifiers by an average accuracy of 3.44%. The event classification accuracy of using a combination of word unigrams, bigrams, NER and POS features (all POS tags and all named entities) (our sixth baseline model) also outperforms the baseline model trained from using using part-of-speech (POS) and Named Entity Recognition (NER) only, features that are often used in the literature, by an average of 1.30%. Machine learning algorithms can achieve high accuracy for classifying event-related messages. Although Twitter messages have some unique characteristics compared to other corpora, machine learning algorithms are shown to be able to classify tweets for the purpose of identifying real-world events.

Chapter 4

Clustering (On-line Clustering)

4.1 Introduction

After demonstrating in the previous chapter the effectiveness of using supervised machine learning methods, (Naive Bayes, Logistic Regression, and Support Vector Machines) in order to automatically separate the event-related messages in a stream of tweets from all the non-event content, in this chapter we introduce the use of an unsupervised clustering algorithm to group topically similar event-related tweets together in the same group (cluster). Because of the speed and volume at which data arrive, and the challenging nature and the language of the social media content, we propose using a variety of features, namely; temporal, spatial, and textual, to enhance the identification of various events, particularly disruptive events.

We start this chapter by reviewing prior research on unsupervised learning (clustering) for the task of event detection (Section 4.2). Section 4.3 defines the term "disruptive event" in the context of social media, then we give examples of disruptive events in real-world scenarios. In section 4.4, we present our online clustering algorithm to automatically assign each event-related tweet to a cluster according to temporal, spatial and textual similarity features. Section 4.5 explores three sets of features (temporal, spatial and textual) and combinations of them, in order to achieve better system performance, we implement an improved model for feature selection that is suitable for microblog data. In section 4.6, we present extensive experiments to evaluate the effectiveness

of the proposed clustering approach using large real-world datasets. In addition, we present the results for different feature selection experiments in section 4.6.

In summary, the contributions presented in this chapter are as follows:

- We propose a general online clustering algorithm, suitable for identifying small-scale and large-scale events from their social media content;
- We explore and analyze three kinds of clustering feature: temporal, spatial and textual, as well as combinations of them, in order to optimize computational resource use when detecting real-time events; and
- We validate the effectiveness of our framework using several large real-world datasets from Twitter and Flickr. We compare the overall performance of our system using the optimized model in terms of Precision, Recall, F-Measure and NMI to the performance of many leading systems, namely, Spatial LDA [106], unsupervised methods [16], [167], topic models [148] and a graph-based approach [131]. We further evaluate it against other leading approaches using Twitter posts from the UK riots in 2011.

The two research questions we aim to address in this chapter are:

RQ2 Can we identify sub-event details including disruptive events and their context within the streaming media (topic clustering)?

RQ3 Since not all features are expected to improve a system's performance, can we investigate the dynamics of event/topic identification of three kinds of influential feature: temporal, spatial and textual, in order to optimize feature selection and to improve the effectiveness of topic clustering?

4.2 Background of clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than they are to those in other groups (clusters), according to some distance measure [19, 74]. From a machine learning perspective, the search for clusters is the unsupervised learning of hidden patterns in large datasets or databases. In general, clustering is a common technique for statistical data analysis, and is used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics [19]. More specifically, it plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, and many others [146].

Because of the rapid growth in the popularity of microblogging applications such as Twitter, there is a need for micro-messages accurate clustering on a large scale, so as to better organize and manage the massive uploaded content. While document clustering has been well studied and successfully applied in many real-life data mining problems [19], applying traditional clustering methods to micro-messages is less likely to perform well for three main reasons: the inherent sparseness of posts, the massive stream of data posted in microblogging services, and the latency required by the proposed clustering algorithm [146, 63, 133]. Latency is the time time between an event occurring and when the event is detected.

Using online topic model clustering to handle a variable number of topics is one possible solution that would make this alternative approach suitable for our problem. One obvious drawback of topic models for the social media domain is that they associate documents with topics purely based on textual content. As we show in Section 4.5, clustering using learning similarity metrics using a variety of textual and non-textual (temporal and spatial) features is more effective than text-based approaches at determining when social media documents correspond to the same event. Still, topic model clustering is an effective technique for characterizing social media con-

tent [16, 121, 163].

In this chapter, we propose an online clustering algorithm that uses a minimum number of parameters to handle a constant stream of new documents on various events, including small-scale events. We focus in this work on real-world event identification for both large-scale and rare (disruptive) events such as car accidents in a given location.

While many previous works focus on document clustering, topic models, and cluster analysis, the clustering of short text such as tweets has been addressed infrequently in the literature. Kang et al. [63] use the Affinity Propagation (AP) algorithm to cluster similar tweets by choosing the exemplars (cluster centers) that best represent the tweets. The main advantage of affinity propagation is that it is a distributed clustering algorithm which finds the best assignment of all tweets to clusters at the same time rather than greedily assigning each tweet to the best cluster. However, affinity propagation and its variants suffer from a number of serious drawbacks. It is hard to know the value of the parameter preferences which can yield an optimal clustering solution. The hard constraint of having exactly one exemplar per cluster restricts AP to classes of regularly shaped clusters, and leads to suboptimal performance. Additionally, the time complexity of affinity propagation is in the order of $O(N^2T)$, where N is the number of data points and T is the number of iterations. In contrast, the time complexity of k-means (Lloyd's algorithm) is in the order of $O(NKT)$ for K clusters.

Sculley [133] presents a modified k-Means algorithm designed for large scale sparse web page clustering. His idea is to use mini-batch optimization for k-means clustering, which reduces computation cost by orders of magnitude below that of the classic Lloyd's k-means algorithm, while yielding significantly better solutions than online stochastic gradient descent. However, Rangrej et al. [121] and Tsur et al. [146] separately show that this modification is not well suited to microblogs and a graph-based approach using affinity propagation performs better than this modification in clustering short text data with minimal cluster error.

Rangrej et al. [121] conducted a comparative study of three document clustering tech-

niques, namely, k-Means clustering, singular value decomposition (SVD) and affinity propagation. Their results show that affinity propagation performs best according to cluster error, but their comparison does not address scalability because their experiments were conducted on a small set of tweets (only 611 manually handpicked tweets). To cope with the scalability and sparsity of tweets, Tsur et al. [146] constructed a framework that is split into two distinct tasks: the batch clustering of a subset of the data concatenating all the micro-messages with the same hashtag and then applying a k-means algorithm to cluster virtual documents. Using a different clustering approach, Ifrim et al. [56] proposed a topic detection method in Twitter streams based on aggressive term filtering and the hierarchical clustering of Tweets on the tweet-term matrix. Other works have focused on clustering events on Flickr [123] by Reuter et al. Their approach is based on identifying similar documents as a record linkage task and enhancing them by tags (such as geographic locations, titles and the description) as significant descriptors.

As regards detecting global events from the social media, Petrovic et al. [112] presented an approach to detecting breaking news stories from a stream of tweets using locality-sensitive hashing (LSH). This research uses LSH to rapidly retrieve the nearest neighbour of a document and accelerate the clustering task. The LSH reduces the dimensionality of high-dimensional data by hashing input items so that similar items map to the same (buckets) with high probability. Becker et al. [16] designed a two-step approach to first cluster the input Twitter stream using an online clustering approach to identify different types of real-world event. Then they used machine learning models to distinguish event clusters from non-event clusters, looking at features such as temporal, social, topical and Twitter-specific features such as retweets and hasthags. Nevertheless, such approaches are limited to widely discussed events and struggle to report rare occurrences.

Many researchers have proposed models and techniques for the purpose of detecting real-world events using social media data. Some of these approaches are based on

supervised methods while others are based on unsupervised methods. Concentrating only on the supervised classification in this domain requires labeled data that are often hard to acquire. Due to the huge number of tweets at any given moment, labeling all instances is impossible and labeling more than a few can be expensive. The main drawback of clustering techniques is that they require many parameters or prior information about events or are restricted to discovering large-scale events. In this chapter, we propose an online clustering framework that handles a constant stream of documents on various events including small-scale ones while using a minimum number of parameters. To enhance the identification of small-scale events, we investigate in depth a variety of features derived from Twitter posts, namely, their temporal, spatial and textual content.

One way to optimize the identification of the patterns and signals that indicate an event is to undertake feature selection experiments [9]. Because not all features are expected to lead to better system performance or contribute equally to improved clustering accuracy, we seek to evaluate the effectiveness of a range of features for identifying events, especially disruptive events. These features may be presented as follows:

- Temporal features: the time the event is reported on the web, which we assume is very close to the real time of an event or when the event took place. We use the publicly available timestamp of a document which is the only temporal information that we use in our system. The associated timestamp is the time at which the image was published. The temporal features are related to the "speed" and "quality" of information diffusion over time [120]. Temporal features can also improve the average precision of extracting events and discriminating between different events [64, 120];

- Spatial features: to approximate the origin of posted content or to estimate the location of a user or to reveal the location of an event; and

- Textual features: which are representative of the text as published content on Twitter.) including near-Duplicate measure, Favorite ratio, Retweet ratio, Hashtag

ratio, Sentiment ratio (positive, neutral and negative), Url ratio and Dictionary-based feature.

In this chapter we focus specifically on optimizing feature selection to increase the performance results of event clustering, and to reduce the number of features required to lower the computational overheads in calculation and to improve the identification accuracy. In addition, these results are achieved using features selected with consideration for computational resource use, which is important when analyzing a real-time data stream surrounding ongoing events. In this work, we implement an improved version of the unsupervised feature selection proposed by Mitra et al. [91] for the task of feature optimization. Then we use the standard metric of *NDCG* (Normalized Discounted Cumulative Gain) [36] for the same feature selection task.

4.3 Disruptive Event Definition

A “disruptive event" in the context of social media can be defined as:

Definition Disruptive event is a special type of event that obstructs (disrupts) the achieving of the objective(s) of another event or interrupts ordinary event routine. It may occur over the course of one or several days, causing disorder, destabilizing security and may result in a displacement or discontinuity [8].

Consider a disruptive event (a car crash, fire, crime... etc.). Many crashes or fires happen every day in different places and their immediate effects last for a few hours. Small-scale incident detection that analyzes tweets for incidents needs to distinguish different crashes to secure high standards of precision and recall. Therefore, additional knowledge about the time and location is needed to distinguish different crashes and aggregate the information that relates to the same one [132]. The definition of ‘disruptive event’ above touches on important aspects of such events, namely, their necessary

association with a defined time period, a geographical boundary and tweet-related features - i.e. features related to the content of the tweets such as the Retweet ratio, Hashtag ratio, Dictionary-based features, etc. (See Section 4.5.2).

The first and the most important property by which to identify a disruptive event is the time dimension of the incident, when it was reported on the social media. It has been shown that the *relevance of time* for an event can play a considerable role in real-world event identification by studying patterns of word usage over time [64, 118]. The temporal features, together with the contents of retrieved documents, can improve the average precision for extracting events and discriminating between different events [64, 120]. The second aspect of the disruptive event is its spatial features or the *geolocation* (the message sender's geographical proximity to the event. It is known that a person's geographical location (geolocation) significantly affects her/his social connections and activities in the offline world [55, 69]. Recent research has also found evidence to show that offline geography has a significant impact on user interactions on online social media such as Twitter [55]. Moreover, researchers have discovered that users prefer to exchange information with other users from their own country (region, city,...), and that less information is exchanged across national boundaries [69, 7].

The third element that we use to detect disruptive events is the textual features of the texts, including the near-Duplicate measure, Favorite ratio, Retweet ratio, Hashtag ratio, Sentiment ratio (positive, neutral and negative), Url ratio and Dictionary-based feature which can be used to discriminate and characterize disruptive events. We present an in-depth experimentation of the these textual features to indicate how they can be used to identify a disruptive event and which of them are the most discriminative features for the disruptive event detection.

Large-scale events such as the Olympics, the NATO summit, EXPO events, etc. have similar potential to attract disaster due to the high concentration of people; terrorist threats and other disasters gravitate towards them. The identification of disruptive events is extremely important for event planning, crisis management, the organization

of strategic resource and risk assessment. Efficient action in these areas can significantly mitigate the effects of any large-scale disastrous event.

4.4 On-line Clustering Algorithm

The classification step separates event-related documents from non-event posts (such as chats, personal updates, spam, and incomprehensible messages) as shown in Chapter 3. Consequently, non-event posts are filtered. To identify the topic of an event, even when it is a potentially disruptive event, we define a temporal, spatial and textual set of features, which are detailed in the next section. We then apply an online clustering algorithm, which is outlined in **Algorithm 1**.

The objective of online clustering is to automatically assign each document to a cluster according to temporal, spatial and textual similarity measures, with no prior knowledge of the number of clusters or the nature of the real-world events. An event is a vector in which each dimension is the probability of some feature in the event. Each tweet is represented as a TF-IDF weight vector of its textual content, and a cosine similarity metric is used as the centroid similarity function E.

Using a set of features $(F_1, ..., F_k)$ for each document $(D_1, ..., D_n)$, we compute the cosine similarity measure between the document and each cluster $(C_1, ..., C_m)$ where the similarity function is computed against each cluster C_j in turn for j = 1, ..., m and m is the number of clusters (initially $m = 0$). We use **the average** weight of each term across all documents in the cluster to calculate the centroid similarity function $E(D_i, C_j)$ of a cluster. The threshold parameters (τ) are determined empirically in the training phase.

Note that the clustering algorithm considers each tweet in turn, and determines the suitable cluster assignment based on the tweet's similarity to any existing clusters. If there is no cluster whose centroid similarity function $E(D_i, C_j)$ is greater than τ, we increment m by one and create a new cluster C_m for D_i. Otherwise, D_i is assigned

to a cluster C_j with maximum $E(D_i, C_j)$. Therefore, these steps are repeated for all tweets (documents) in a timeframe.

The centroid similarity function $E(D_i, C_j)$ between a document D_i and a cluster C_j is computed by comparing the features (words) of D_i to those of the cluster C_j. We use the average weight of each term across all documents in the cluster to calculate the centroid similarity function $E(D_i, C_j)$ of a cluster. The average is defined as $\frac{1}{|C_j|} \sum_{D_i \in C_j} D_i$.

Algorithm 1: Online Clustering Algorithm

Input: n set of documents $(D_1, ..., D_n)$

Threshold τ

Output: m clusters $(C_1, ..., C_m)$

Step 1: **For** a given τ, compute the centroid similarity function E (D_i, C_j) of each cluster C_j

Step 2: **If** centroid similarity E$(D_i, C_j) \geq \tau$ **do:**

1) A new cluster is formed containing D_i ;

2) The new centroid value = D_i.

Step 3: **If** centroid similarity E$(D_i, C_j) < \tau$ **do:**

1) Assign it to the cluster which gives the maximum value of E(D_i, C_j) ;

2) Add D_i to cluster j and recalculate the new centroid value C_j.

We describe the different set of features $(F_1, ..., F_k)$ in Section 4.5.2. For instance, for the temporal features, we retain the most frequently occurring terms in a cluster in hourly time frames and compare the number of posts published in an hour that contain term t to the total number of posts in this hour. For the spatial features, we make use of three approaches to extract geographic content from clusters. The first one is from Twitter and the latitude and longitude coordinates of its source were provided by the user. The second method depends on shared media (photos and videos) by using the GPS coordination of the capture device (if supported). Third, Open NLP (http://opennlp.sourceforge.net) and Named-Entity Recognition (NER) were imple-

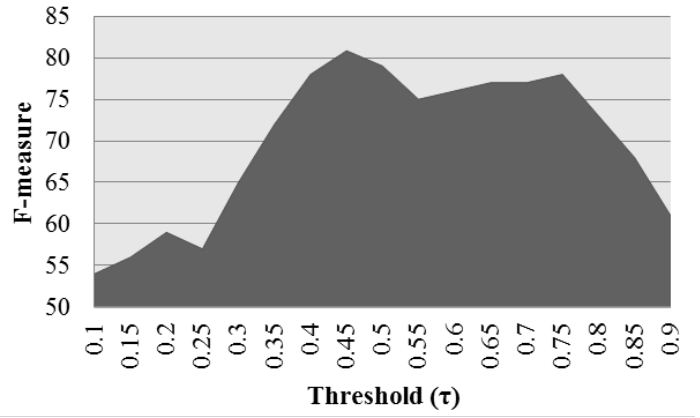


Figure 4.1: F-measure of the online clustering algorithm over different thresholds.

mented for geotagging the tweet content (text) to identify places, organization, street names, landmarks etc. For the other textual features, we calculate each one of them differently. For example, we calculate the Hashtag ratio by computing the ratio of tweets containing hashtag (#) over the total number of tweets in the timeframe. Section 4.5.2 discusses each one of these features as well as it discusses in detail the other textual features.

To tune the clustering threshold τ for a specific dataset, we ran the clustering algorithm on a subset of labelled training data (Information about the labelling and the annotation can be found in Section 4.6.1). We evaluated the algorithm's performance on the training data using a range of thresholds, and identified the threshold setting that yielded the highest-quality solution according to a given clustering quality metric (here we implemented the f-measure). Threshold values for the online clustering algorithm were varied from 0.10 to 0.90 at graded increments of 0.05% with a total of 17 tests in order to find the best cut-off of τ =0.45 (63 character difference). Figure 4.1 illustrates the F-measure scores for different thresholds where the best performing threshold τ =0.45 seems to be reasonable because it allows some similarity between posts but does not allow them to be nearly identical. Details of this experiment and dataset are given in 4.6.2.

It was decided to use an online clustering algorithm for three main reasons: (i) it sup-

ports high dimensional data because it effectively handles the large volume of social media data produced around events; (ii) many clustering algorithms such as K-means require previous knowledge of the number of clusters. Because we do not know the number of events *a priori*, online clustering is suitable, in that it does not require such input; (iii) partitioning algorithms are less effective in this case, because of the high and constant sheer scale of the user contributed messages.

4.5 Feature Selection

Feature selection is a fundamental problem in mining large data sets. The problem is not limited to the total processing time but involves dimensionality reduction to achieve better generalization. Feature selection is an effective way of reducing dimensionality, removing irrelevant data, and increasing learning accuracy. We compute the features of Twitter message clusters in order to reveal characteristics that may help detect the clusters that are associated with events, particularly disruptive events. Not all features are expected to improve the system's performance or lead to more accurate discrimination of the clustering algorithm. Indeed, for many reasons the inclusion of some features could result in worse behavior by the system, such as introducing greater computational cost [9] may lead to overfitting [9, 91] or could result in some scalability issues [154, 146].

Here, we present and analyze in-depth three types of feature: temporal, spatial and textual features. Our experiments led to the following conclusions: first, disruptive events are identifiable regardless of the "influence of the user" discussing them, and can be identified over a variety of topics. Second, temporal features are the best event identifiers and hence should not be disregarded or ignored. Third, a combination of optimum textual features with temporal and spatial features performs best in the event detection task. We believe that these findings provide new insights for gathering information around real-world events as well as being a useful resource for improving situational

awareness and decision support.

4.5.1 Related Work: Feature Selection in Social Media

Traditionally, information retrieval applications do not take full advantage of all the temporal, spatial and textual information embedded in documents to provide alternative search features and user experience [5]. However, in the last few years there has been exciting work on analyzing and exploiting temporal, spatial and textual features for the presentation, organization, and in particular the clustering results. In this section, we summarize the current research on temporal, spatial and textual features and the way in which it has been applied to data mining tasks.

Temporal Features

Time is an important dimension of any information space. It can be very useful for a wide range of information retrieval tasks such as document exploration, similarity search, summarization, and clustering [5]. Temporal information embedded in documents in the form of temporal expressions provides an important means of further enhancing the functionality of current information retrieval applications [4]. Radinsky et al. [119] proposed to incorporate the "temporal behavior" of words in computing their relatedness. The idea originates from the observation that semantically related words do not necessarily co-occur in the same articles; however, they are likely to be employed at roughly the same time. The temporal dynamics of subtopics is studied by Nguyen and Kanhabua in [96] and is used to improve the ranking effectiveness of such queries at particular times.

Social media posts generally come with a creation time-stamp, which can be used for topic detection and tracking (TDT), as demonstrated in [130, 154], which analyzed the evolution of stories and topics over time. Yang et al. [163] study the dynamics of information novelty in some evolving news stories. There has also been much work on

the community structure of the blogosphere. The authors of [30] show that the prediction of information cascades is feasible and the relative growth of a cascade becomes more predictable as more "reshares" are observed over time. Hence, these temporal features are key predictors. Rather than attempting to predict cascades, Elsas and Dumais [45] study the dynamics of document content changes to the rank of documents on the basis of their temporal characteristics. In Facebook, the authors of [14] found that the time it takes for the first reply to arrive is another good indicator of the length of the thread.

The summarization and visualization of dynamic documents has been explored in several studies [136, 97, 155, 43]. In general, tweet summarization has to take into consideration the temporal feature of the arriving tweets, since tweets have not only textual, but also temporal features (a tweet is closely correlated with its posted time) [155]. Shamma et al. [136] studies the relationship between the temporal feature of the arriving tweets and media events. By examining conversation volume and activity over time, they were able to temporally segment a live news event and identify the key people in it. Other work has focused more on visualizing the temporal patterns of messaging behavior on social networks. Dork et al. [43] present a visual timeline-based backchannel for conversations around events. They have introduced topic streams graph and a temporally and topically adjustable stacked graph that visualizes topics extracted from digital backchannel conversations.

Spatial Features (Geospatial, Regional)

Several algorithms have been proposed to estimate the location of Twitter users by means of a content analysis of tweets. Eisenstein et al. [44] built geographic topic models to predict the location of Twitter users in terms of regions (reporting 58% accuracy over 4 regions) and states in the US (predicting 48 US states with 24% accuracy). Hecht et al. [54] built Bayesian probabilistic models from the words in tweets in order to discover the country and state-level location of Twitter users. They were

able to get approximately 89% accuracy with the countries (4 countries), but only 27% accuracy for predicting the states (50 states in the US). Han et al. [50] studied the location estimation problem that is based on the automatic identification of location indicative words; that is, words that implicitly or explicitly encode an association with a particular location.

To tackle the problem of location sparsity, Cheng et al. [31] described a city-level location estimation algorithm, which is based on identifying local words from tweets using statistical predictive models. They achieved approximately 50% accuracy in detecting city-locations. Moreover, Mahmud et al. [86] combine time zone information and content-based classifiers in a hierarchical model at different granularities, reporting accuracy rates of 64% for cities, 66% for states, 78% for time zones and 71% for regions. Our use of spatial features relates to their predictive power when the aim is to identify disruptive events - essentially, whether neighbourhood-, city-, or country-level information is a significant predictor. Note that the above-mentioned studies have been evaluated over different datasets.

Most of the previous studies on geolocation prediction from Twitter data have collected tweets from a specific country, and are limited only to tweets written in English. In addition, most of these studies have focused on the United States, classifying tweets either at a city or state level. The only exception is [50], who focused on a broader geographical area, including 3.7k cities all over the world. The other studies documented in the literature have relied on tweet content, using such techniques as topic modelling [54, 44] to find locally relevant keywords that reveal a user's likely location. Meanwhile, more knowledge can be considered to imply a user's location, such as network features, information from the user's followers and followees.

Textual features

Textual features can be used as individual features (e.g. n-grams), but many studies have combined them to optimize the solution to data mining challenges, such as

information diffusion [83, 104, 30], opinion mining [145, 2, 126], spam and spammer detection [73], and identifying the most knowledgeable posts and influential users [83, 54, 14, 27, 83].

Using the topic model in [27], a set of raw features (number of original tweets, number of retweets, and number of mentions) has been used for identifying the most influential Twitter users. Cheng et al. [30] presented several feature sets and studied their effects on cascade growth prediction. They found that temporal and structural features are key predictors of cascade size. Agarwal et al. [2] investigated two kinds of model: a feature based model and a tree kernel based model for the purpose of sentiment classification. They demonstrate that both models outperformed the unigram baseline model that had previously been shown to work well for Twitter sentiment analysis.

Another key area of research related to the textual features is hashtag popularity, which is considered by Ma et al. [83]. They demonstrated that contextual features (such as the number of users, number of tweets, retweet ratio, etc.) are more effective than content features (such as tweets containing URL, the ratio of neutral, positive, and negative tweets, etc.) in predicting hashtag popularity. Lee et al. [73] created social honeypots to identify spammers on MySpace and Twitter and proposed classification algorithms to distinguish between spammers and legitimate users. In terms of feature space, they extracted and investigated four set of features: (i) user demographics: including age, gender, location, and other descriptive information about the user; (ii) user-contributed content: including "About Me" text, blog posts, comments posted on other users' profiles, tweets, etc.; (iii) user activity features, including posting rate, tweet frequency; (iv) user connections, including number of friends in the social network, followers, following. Overall, all the proposed features have positive discrimination power.

More relevant to the *dictionary-based feature*, Olteanu et al. [103] created a lexicon of crisis-related terms (380 single-word terms) that frequently appear in crisis events. They described an approach toward improving the recall in the sampling of Twitter communications that can lead to greater situational awareness in crisis situ-

ations. However, we create in this work a much richer (1538 terms) and more general lexicon of terms that tend to appear frequently across various disruptive events including; weather, communication, energy, transportation, health, crime, terrorism, politics and others (See Section 4.5.2 Dictionary-based feature). Unlike other researchers, we analyze the role of each feature and group features in the context of detecting events from tweets, extracting the most relevant features and employing them in the attainment of higher system performance. Similarly, irrelevant or redundant features may make optimum results impossible and hence should be discarded. We show experimentally that our framework performs better in practice when selecting the optimum features for the purpose of detecting disruptive events from Twitter.

4.5.2 Clustering Features

Temporal Features

Temporal features are important factors that have been investigated in many studies of event detection via social media. The volume of tweets and the continually updated commentary around an event suggest that informative tweets from several hours ago may not be as important as new tweets [9, 16]. For this reason we identify the most frequent terms in the cluster across a range of time windows. In our experiments we use a range of time windows to improve the efficiency of the event clustering system in terms of accuracy and total running time.

Assumption: Each tweet is associated with a timestamp, which we assume is very close to the real time of an event or when the event took place.

The volume of messages that contains terms related to an event exhibits unique characteristics. By finding these characteristics or the patterns of these terms in the clusters, we can enhance the identification of these events by determining the size of the sliding window. Our goal is to capture the temporal behaviour with a set of temporal features from our clustering algorithm.

Many web documents, especially in microblogging services, are dynamic, with content changing in varying amounts at varying frequencies. Many current systems and search algorithms have a static view of the document content [5]. Therefore, we hypothesize that the temporal dynamics of document content is essential to improve a system's performance. This has been shown to be an important feature in predicting thread length on Facebook [14], a primary mechanism in predicting popularity in Twitter [83], as the most important factor in influencing a cascade through the network [30]. Moreover, it provides better search results, improving the user experience and the functionality of search applications [4].

We retain the most frequently occurring terms in a cluster in hourly time frames and compare the number of posts published in an hour that contain term t to the total number of posts in this hour. The 1-hour time window leads to the best performance, since it requires much less computational time than other settings and produces the second highest degree of accuracy (as is shown in Section 4.6.3).

Spatial Features (Geospatial, Regional)

Events are characterized by a rich set of spatial and demographic features [8]. In this thesis, we make use of four approaches to extract geographic content from clusters. The first one is from Twitter as we extract current users' profiles locations. The second one is from Twitter as well as we obtain the latitude and longitude coordinates of its source were provided by the user. The third method depends on shared media (photos and videos) by using the GPS coordination of the capture device (if supported). Forth, Open NLP (http://opennlp.sourceforge.net) and Named-Entity Recognition (NER) are implemented for geotagging the tweet content (text) to identify places, organization, street names, landmarks etc. These approaches rely purely on Twitter with no need for the user's IP address, private login information, or external knowledge bases, which give them the maximum advantage [31, 86].

Once the geographic content is extracted from each tweet in a cluster, we aggregate

them to determine the cluster's overall geographic focus. The higher the volume of tweets from nearby coordinates, the higher the level of confidence in the location of the event. We make the following assumptions regarding the spatial features:

- A user's location is more likely to appear in his/her tweets than other locations;
- A user's location tends to be closer to the locations of his/her friends in the social network; and
- A user's location is mentioned at least once in his/her tweets or by his/her friends in the social network.

In this thesis, we do not use the network features to identify or extract users' locations, a task that we reserve for future work.

The spatial feature has been shown to be a weak event indicator due to the slow adoption of geospatial features from Twitter users, as shown in [31]. We examined users' locations in our first dataset (Section 2.4.1) which contains around 674,000 unique users and found that 12.7% (85,598 users) of the total user profiles listed their locations as granular, such as a city name, 7.7% (51,898) contained a country name and only 31,004 (4.6%) revealed their locations as a latitude/longitude coordinate. Overall, most users tend to over generalize their location (e.g., East Region), omit it altogether, or write something uninformative (e.g., Middle of the Desert. In addition, Twitter users often rely on shorthand and non-standard vocabulary (non-traditional gazetteer terms) for informal communication; some users simply do not wish to reveal their location, which all makes determining location-terms a non-trivial task [86]. Our results also show some of the differences in user behaviour across regions, languages and backgrounds across the globe as do results from other studies, namely Cheng et al. [31] and Mahmud et al. [86].

We assume that all locations provided by users are correct, although Hecht et al. [54] found that 34% of Twitter users had entered fake locations in their profile. Some users

may intentionally misrepresent their home location to cover their actual location due to privacy concerns. In addition, some users provided locations that differ from their actual location at the time because they are tweeting as they travel. It should also be noted that a Twitter user's location may not be as the location of an event. Similarly, the geotagged location of the tweet may not be the event's location. In addition, it is possible that a tweet about a particular event's location may not be the event's location at the time.

Textual features

There are two main tasks in this thesis as regards textual features: first, we analyze various textual features in order to select the best contributors to the task of event detection. Second, we rank features using performance measures and eliminate irrelevant features that introduce computational cost. First we introduce these features in detail.

- **Near-Duplicate measure**

  The average content similarity over all pairs of tweets posted in a 1-hour time slot cluster (We experimentally evaluate the choice of 1-hour time slot in Section 4.6.3) was calculated using:

$$\sum_{a,b \in \text{set of pairs in tweets}} \frac{similarity(a,b)}{|\text{set of pairs in tweets}|} \tag{4.1}$$

  where the content similarity is computed using the cosine similarity over words from tweet a, b vector representation $\overrightarrow{V}(a)$, $\overrightarrow{V}(b)$ of the tweet content:

$$similarity(a,b) = \frac{\overrightarrow{V}(a).\overrightarrow{V}(b)}{\left|\overrightarrow{V}(a)\right|\left|\overrightarrow{V}(b)\right|} \tag{4.2}$$

  If two tweets have a very high similarity, we assume that one of them is a near-duplicate of the other. The original tweet is considered to be the first tweet

in a particular time frame and/or is the one which is the shorter of the two in length. Even though duplicates are believed to be disadvantaged (newer messages do not add any unique information), several users independently tweeting about an event would effectively increase the confidence level of an event. By detecting near-duplicates, we can tackle several problems such as first story detection [112], rumor identification [84, 116], and news story detection [105] and propagation [113] where newly added content differs slightly from previously-detected versions of the story and this can be considered an update of the story. In our system, we compute the average content similarity over all pairs of messages posted in a (1-hour time slot) cluster. If the two posts have a very high similarity (the cosine similarity is above 0.9), we assume that one of them is a near-duplicate of the other.

- **Retweet ratio**

  Retweeting represents the influence of a tweet beyond the one-to-one interaction domain. Popular tweets can propagate multiple hops away from the source as they are retweeted throughout the network [27]. The retweet feature in a social network can serve as a powerful tool to reinforce a message when not only one user but a whole group of users repeat the same message [83]. Hence, the number of retweets can be used as an indication of popularity [111]. We calculate this attribute by normalizing the number of times a tweet (or a photo or a video) appears in a timeframe to the total number of tweets in the timeframe.

  The retweet ratio can reveal event-related tweets whether users agree with the message or just wish to spread the information (warning, advice, evidence, etc.) to more users. Other applications of retweets concentrate on estimating rumors in social media through analyzing the retweet path of rumor-tweets. While the number of retweets is an indication of popularity, it does not always consider the content of posts where many users retweet/reshare without verifying the content. Most celebrities, such as actors, writers, musicians, and models, have a high

number of retweets, however some ordinary users have a higher retweet rate which indicate that the message's content is vital too. As a consequence, some users suddenly became popular over the course of an event as their retweet ratio increases dramatically. Retweets and mentions, along with indegree (indegree is the number of people who follow a user), are the three main factors that indicate popularity on Twitter, as shown in [27]. We also expect that the RT can play a significant role in identifying disruptive event tweets.

- **Mention ratio**

  A mention is a mechanism used in Twitter to reply to users, engage others or join a conversation in a form of (@username). A user can mention one or more users anywhere in the body of the post. Regarding event reporting, users tend to mention journalists, politicians and official accounts such as news agencies or government official accounts to drive their attention about an event or to add more credibility to their event-related posts. We calculate this attribute by normalizing the number of mentions (@) relative to the number of posts in the cluster.

  Retweets are driven by the content value of a tweet, while mentions are driven by the name value of the user. Such subtle differences lead to dissimilar groups of the top Twitter users; users who have high indegree do not necessarily engender many retweets or mentions [27]. This finding suggests that indegree alone reveals very little about the influence of a user. Another reason is that indegree is dynamically changing and hence we do not use the indegree as a feature in this thesis.

- **Hashtag ratio**

  Hashtags are important features of social networking sites and can be inserted anywhere in a message. Some Hashtags indicate their posted messages (#bbcF1) and others are dedicated originally to events such as (#abudhabigp). In addition, topic related hashtags are used as an information seeking index on Twitter to search Twitter for more tweets on the same topic [32]. The use of hashtags

became a coordinating mechanism for disruptive activity on Twitter [142, 143]. In [142], Starbird and Palen showed that the hashtag is employed as a mechanism for identifying useful social connections in times of crisis. They also noted that a high percentage of "on the ground" users adopts popular hashtags related to protests when reporting street protests [143]. The Hashtag ratio is computed as the ratio of tweets containing hashtag (#) over the total number of tweets in the timeframe.

- **Link or Url ratio**

  Since Twitter is limited to 140 characters per message, it is common in the Twitter community to include links when tweeting to share additional information that makes tweets more informative or for referencing. The co-occurrence of URLs in a cluster or sharing links to popular websites (news agencies or government sites) may confirm that these tweets refer to the same event and improve the level of confidence in an event. This attribute is calculated by the fraction of tweets that contain URL to the total number of tweets in a timeframe.

- **Tweet sentiment**

  Users post real-time messages in microblogging websites giving their opinions on a variety of topics (e.g. news events) which embody positive or negative sentiment [2]. Sentiment analysis over Twitter and other similar microblogs offer organizations a fast and effective way to monitor the public's feelings towards their brand, business, directors, etc. [126]. Here, we first study whether sentiment polarity posts (0 indicates neutral, 1 indicates positive or negative sentiment) are significant features when reporting events. Subsequently, we investigate the influence of positive, negative and neutral sentiment on identifying disruptive events.

  To calculate sentiment we use a semantic classifier based on the use of SentiStrength algorithm [145] which is suitable because it is designed for short informal text with abbreviations and slang. For each tweet, the SentiStrength

algorithm computes a positive, neutral or negative sentiment score. Then we compute the average cluster-level sentiment (set of tweets) in order to study the effect of average positive or negative sentiment with respect to events.

- **Dictionary-based feature**

  This bag of words model uses a dictionary of trigger words to detect and characterize events; these words are manually labelled by experts and decision makers. We use a subset of verbs, nouns and adjectives in the (events and actions) category from WordNet (http://globalwordnet.org) to create our lexicon feature. We created 9 lexicons regarding disruptive events from the clustering scheme, one for each of the following popular topics: weather, communication, energy, transportation, health, crime, terrorism, politics and others. The total number of terms is 1538, which tend to frequently appear across various disruptive events. Each word or term was manually annotated by three independent experts (annotators). The inter-annotator agreement between annotators was calculated using kappa coefficient (kappa = 0.831), which indicates a substantial level of agreement. Table 4.1 shows our lexicons with topics and examples in each category.

4.5.3 Feature Selection Algorithm

Most existing feature selection algorithms were designed for traditional (high-quality) documents containing uniform entities (crisp, clear and easy to extract) rather than low-quality documents containing short, barely comprehensible text, with many spelling and grammatical errors and typos. We chose to implement an improved version of the unsupervised feature selection presented in [91] by Mitra et al. for several reasons: first, it resolves the issue of the high-computational complexity involved in searching large data sets; second, the computation time is reasonable even for large data sets where other algorithms perform well only with medium sized data sets. Third, the unsupervised feature selection results are among the best clustering performances for real-world data sets.

Table 4.1: Topics and sub-topics with examples taken from the corresponding lexicons.

Topics	Sub-Topics	Examples	Total
Weather	Heavy rain, Wind, Fog, Storm, High waves, Flooding, Heat waves, Cold.	Verb: rain, suffer, Noun: fog, visibility, Adjective: heavy, cold, hot,	155
Energy	Blackout, Power lost, Fire, Electricity cut, Water supply, Gas leak.	Verb: lose, leak, continue, Noun: power, signal, authority, Adjective: long, delay,	82
Communication	Signal, Communication lost, Breakdown.	Verb: communicate, restore, Noun: signal, company, Adj: Technical, temporary,	33
Transportation	Public transport, Traffic jam, Accidents, Crashes, Long delay, Services, Hazardous, Roads, Cancellation.	Verb: see, take, Noun: car, crash, plane, train, Adjective: fast, dangerous,	258
Health	Flu, Fever, Virus, Disease, Illness.	Verb: spread, circulate, Noun: influenza, rate, season, Adjective: medical, serious,	45
Crime	Shooting, Theft, Damage, Kidnapping, Homicide, Murder, Manslaughter, Drugs, Threat, Fight, Money laundering, Sexual assault, Illegal, Fraud, Alcohol, Corruption, Internet Crimes.	Verb: witness, report, arrest, Noun: victim, blood, abuse, Adjective: vulnerable, brutal,	342
Terrorism	Terrorist Activities, Explosion, Explosives, Weapons, Hostage, Armed robbery, Bomb, Attacks, Violence, Stabbing, Suicide, Hacking.	Verb: release, support, Noun: email, Syria, knife, Adjective: suspicious, explosive,	230
Politics	Riots, Protests, Political insults, Celebrities, Occasions, News.	Verb: organize, group, Noun: chaos, looting, arson, Adjective: corrupt, violent,	257
Others	Religious, Financial, Social incidents, Death, Rumour.	Verb: spread, die, claim, confirm Noun: truth, correction, rumour, Adjective: false, incorrect,	136

Basically, Mitra et al. [91] proposed an unsupervised algorithm which uses feature dependency/similarity for redundancy reduction, but requires no search. Their method involves partitioning the original feature set into some distinct subsets or clusters so that the features within a cluster are highly similar while those in different clusters are dissimilar. Hence we have two main forces: the attractive forces between similar features and the repulsive forces between dissimilar features. The Maximal Information Compression Index (MICI) is used as a clustering feature similarity measure which can be computed in much less time than many indices used in other methods of supervised and unsupervised feature selection. Finally, a single feature from each cluster is selected to constitute the resulting reduced subset.

Let the original number of features be D, and the original feature set be O where $O = \{f_i, i = 1, ..., D\}$. We represent the dissimilarity between features f_i and f_j by $S(f_i, f_j)$. The higher the value of S, the more dissimilar the features. Let r_i^k represent the dissimilarity between feature f_i and its kth nearest-neighbour feature in R, where R is the reduced feature subset. The unsupervised algorithm of feature selection is outlined in **Algorithm 2**.

The dissimilarity between the two features $S(f_i, f_j)$ is calculated by the Maximal Information Compression Index (MICI). The MICI is a well-known index for measuring dissimilarity between features and has been applied in many pattern recognition and data mining tasks. The Maximal Information Compression Index is defined as:

$$\lambda(x, y) = \left[a - \sqrt{a^2 - 4b(1 - \rho(x, y)^2)}\right] /2 \tag{4.3}$$

where $a = var(x) + var(y)$ and $b = var(x).var(y)$ The *correlation coefficient* (ρ) between two random variables is defined as $\rho(x, y) = \frac{cov(x,y)}{\sqrt{b}}$, $var()$ denotes the variance of a variable, and $cov()$ the covariance between two variables.

Algorithm 2: Feature Selection Algorithm

Step 1: Choose an initial value of $k \leq D-1$. Initialize the reduced feature subset R to the original feature set O.

i.e., $R \leftarrow O$.

Step 2: For each feature $f_i \in R$, compute r_i^k.

Step 3: Find the feature $f_{i'}$ for which $r_{i'}^k$ is minimum.

Retain this feature in R and *discard* k nearest features of $f_{i'}$.

(Note: $f_{i'}$ denotes the feature for which removing k nearest-neighbors will cause minimum error among all the features in R. **Let** $\epsilon = r_{i'}^k$.

Step 4: **If** $k >$ cardinality(R)-1: k= cardinality(R)-1.

Step 5: **If** k=1: **Go to Step 8**.

Step 6: **While** $r_{i'}^k > \epsilon$ **do:**

(a) $k = k - 1$.

$r_{i'}^k$=inf$_{f_i \in R} r_i^k$.

(k is decremented by 1, until the "kth nearest-neighbour" of at least one of the features in R is less than ϵ-dissimilar to the feature)

(b) **If** k=1: **Go to Step 8**.

(if no feature in R is less than the ϵ-dissimilar "nearest-neighbor", select all the remaining features in R)

End While

Step 7: **Go to Step 2**.

Step 8: Return feature set R as the reduced feature set.

4.6 Empirical Evaluation

The goal of these sets of experiments is to evaluate our on-line clustering algorithm for the purpose of identifying events, particularly disruptive events, on large datasets of real-world data from two popular social media sites, namely Twitter and Flickr. We describe the annotation process and the evaluation measures and report the experimental settings (Section 4.6.1). We explore three sets of features in turn, the temporal, spatial

and textual, as well as combinations of them, in order to improve system performance.

We validated the effectiveness of our framework using several datasets of over 40 million Twitter messages (2.4.3). We further evaluated it against other approaches, using the international Flickr MediaEval2012 challenge [109]. To test that our framework was generalizable and language independent, we evaluated it using two real-world datasets from popular but distinct social media sites, Twitter and Flickr. These two datasets contain many languages notably Arabic and English, together with others in the Middle East Twitter Corpus (2.4.3) and German, Spanish and English in the MediaEval2012 Flickr Corpus (2.4.4). We compared the overall performance of our system using the optimized model in terms of Precision, Recall, F-Measure and NMI to the performance of many leading systems, namely, Spatial LDA [106], unsupervised methods [16], [167], topic models [148] and a graph-based approach [131].

Finally, we made a case study of our approach by evaluating it against other leading approaches using Twitter posts from the UK riots in 2011, and a publicly accessible account of *actual reported* intelligence obtained and reports received by the Metropolitan Police Service at the time. Smaller scale events include localized looting, violence and criminal damage. The results show that our system can detect events related to the riots as well as the terrestrial sources did - in some cases we detected the event *before* the intelligence reports were recorded.

4.6.1 Experimental Setup

Datasets: We used three large-scale real-world datasets from Twitter and 1 large dataset from Flickr in this evaluation of the proposed online clustering algorithm (see Section 2.4 for more details).

Annotations: Following the classification output we employed three human annotators to manually label 1600 clusters (from our first dataset 2.4.1), selected from the top-20 fastest-growing clusters according to hourly message volume at the end of each

hour in October (800 clusters) and November (800 clusters) 2015. The October data were used for training and refining the clustering algorithm and the November data were used to test and evaluate the clustering output. The task of the annotators was to choose one category from the following eight categories: politics, finance, sport, entertainment, technology, culture, disruptive event and other-event. The other-event category represents all other events which are not related to the above categories. So the annotators manually labelled clusters (representative tweets from clusters). We compared the clustering output with the annotation output where the annotators agreed on clusters. Then we calculated the confusion matrix between the annotators' clusters and the clustering framework clusters. Then we determined the True Positive (TP), the False Positive (FP), and the other measures.

We consider the event thread selection approach presented by Petrovic et al. [112], which selects the fastest-growing threads in a stream of Twitter messages and then re-ranks them based on thread entropy and number of retweet. Experientially and on a large dataset, Becker et al. [16] indicated that selecting clusters based on such re-ranking strategies yields similar results as selecting the fastest-growing clusters. In fact, they showed that the number of events identified by Fastest was similar to the number of events identified by Random, implying that the growth rate of clusters is not an effective indication of event content.

To ease the annotation process, a set of instructions and examples was given to the annotators so that they can perform the annotation task. The instructions that we provided to the annotations are shown in Table 4.2. In addition, some of the example tweets along with their allotted categories (Categories are: Politics, Finance, Sport, Entertainment, Technology, Culture, Disruptive Event and Other-Event) that were provided to the annotators are shown in Table 4.3. For additional example tweets and annotations used in this chapter, these can be found in Appendix A.2.. The agreement between annotators was calculated using Cohen's kappa (kappa = 0.782), which indicates an acceptable level of agreement. For testing we used only the clusters which all the

Table 4.2: The instructions provided to the annotators for the annotation task (Clustering), followed by an example tweet.

Instructions: Given a Twitter message and eight categories (Categories are: Politics, Finance, Entertainment, Sport, Technology, Culture, Disruptive Event and Other-Event), Please note that the Other-Event category represents all other events which are not related to the above categories. Can you choose one category that best describe the Twitter message from the eight given categories? Please read the examples and the invalid responses before beginning if this is the first time you are working on this annotation task.
Tweet: **#TrafficUpdate : An #accident has been reported on SMBZ Rd After #GlobalVillage towards Sharjah, please be extra cautious** Overall, the tweet belongs to which category? ☐ Politics ☐ Finance ☐ Entertainment ☐ Sport ☐ Technology ☐ Culture ☐ Disruptive Event ☐ Other-Event

annotators agreed over (602).

Evaluation Methods: We used the standard classification metrics of precision, recall and the F-measure to measure the effectiveness of our framework. Precision is a measure of false positives. Recall is a measure of false negatives. The F-measure is a harmonized mean of precision and recall. Accuracy is the proportion of correctly classified tweets to the total number of tweets. We also implemented two well-known information retrieval metrics, namely, average precision (AP), $Precision@K$ and *NDCG* [36] to evaluate the overall performance of the event detection task. Averaged precision measures how many of the identified clusters are correct, averaged over hours per

Table 4.3: List of example tweets and annotations that were provided to the annotators for the clustering task (Categories are: Politics, Finance, Sport, Entertainment, Technology, Culture, Disruptive Event and Other-Event).

Tweet	Categories
Mohamed bin Zayed holds talks with Bahrain Crown Prince on ways to enhance fraternal ties, GCC coordination efforts, regional developments #AD #UAE #Bahrain #GCC	Politics
.@RTA_Dubai holds 4th #Dubai International Project Management Forum in November 2013 http://tinyurl.com/n434h3p	Finance
Adios Luis Suarez, I wish you well. You were a great addition to a great club. #LFC #YNWA #Liverpool	Sport
World Health Day, celebrated on 7 April every year to mark the anniversary of the founding of WHO #World_Health_Day Please come along and visit us at Twam Hospital #AlAin	Entertainment
Smart #Dubai launches Dubai Careers future generation of digital recruitment platforms	Technology
. @DubaiCulture's Reading Box initiative attracted high number of school students from the city http://tinyurl.com/l4fp347 #Dubai	Culture
#TrafficUpdate : An #accident has been reported on SMBZ Rd After #GlobalVillage towards Sharjah please be extra cautious	Disruptive Event
Researchers discussed how to improve immunotherapy at #AACR could the common cold virus help? http://po.st/D6IstF #AbuDhabi	Other-Event

day and calculated on the basis of the precision of each cluster per day. Average precision is a common evaluation metric in tasks such as ad-hoc retrieval, where only the set of returned documents and their relevance judgments are available. $Precision@K$ reports the fraction of correctly identified events out of the top-K selected clusters, averaged over all hours, whereas the *NDCG* Normalized Discounted Cumulative Gain metric ranks the top events relative to their ideal ranking and *NDCG* supports documents related to graded judgments and rewards in the top ranked list.

The discrimination power between different proposed features can be measured by generating a Receiver Operating Characteristics (ROC) curve [159, 51]. ROC curves plot false positive rates on the horizontal axis and true positive rates on the vertical axis for varying thresholds. The closer the ROC curve is to the upper left corner, the higher its overall accuracy. The coordinate (0, 1) represents 100% sensitivity (no false negatives) and 100% specificity (no false positives).

4.6.2 On-line Clustering Evaluation

In order to evaluate the clustering performance, we employed three human annotators to manually label 1600 clusters. The task of the annotators was to choose one category from the following eight categories: politics, finance, sport, entertainment, technology, culture, disruptive event and other-event. The other-event category represents all other events which are not related to the above categories. We divided the test set into six datasets according to each day of the annotation task. The annotators' task was to manually label clusters (not tweets) to obtain the total number of events per category per day. To ease the annotation process, examples were given to the annotators along with their allotted categories. We report below the results from our analysis. Table 4.4 shows the average precision percentages achieved by using our on-line clustering algorithm of the cluster on the test set.

Table 4.4: Average precision of the online clustering algorithm, in percent

Date	Politics	Finance	Sport	Entertai -nment	Techno -logy	Culture	Disruptive Events	Average per Day
30-Oct	82.50	81.11	85.71	76.00	78.80	74.29	87.50	80.84
31-Oct	78.71	85.67	80.62	76.87	74.21	83.36	82.04	80.21
1-Nov	84.15	82.52	80.90	74.45	75.75	81.61	84.67	80.58
2-Nov	77.01	79.40	77.29	72.51	72.19	67.50	90.00	76.56
3-Nov	79.91	83.49	90.21	68.96	82.35	83.36	78.17	80.92
4-Nov	84.34	81.33	82.04	74.01	83.99	79.03	82.76	81.07
Average per Topic	81.10	82.25	82.79	73.80	77.88	78.19	84.18	**80.03**

In general, the online clustering algorithm was able to achieve a good performance of 80.03%, although it was inconsistent with respect to topics. For example, the average accuracy of identifying sports events was greater than the average accuracy of identifying entertainment events by about 9%. It is easier to extract and categorize events such as politics, finance, sport and disruptive events than events such as entertainment, technology or cultural events even for humans - but this caused the main disagreement between annotators in the annotation task. Interestingly, the best performance achieved by the online clustering algorithm concerned the disruptive event identification of 84.18%.

There are many ranking techniques and models for ranking the documents/tweets in event clusters. We can rank them by the cluster size (which is a metric based only on tweets) or we can rank them by their distance to the closest Wikipedia page. In this thesis, we rank the tweets by their distance to the closest centroid using the centroid similarity function $E(D_i, C_j)$ that was calculated in Algorithm 1.

In order to further evaluate the performance of our clustering algorithm, we repeated the same experiment with the same setting on a different dataset (Middle East Twitter Corpus) 2.4.3 using *NDCG* and $Precision@K$ evaluation measures. We gen-

erated a set of ground truth events that took place during the period we examined (between 1st October 2015 and 30th November 2015). Trying to keep the ground truth set as objective as possible, we used the list of October 2015 and November 2015 events reported by Wikipedia (http://en.wikipedia.org/wiki/2015#october) and (http://en.wikipedia.org/wiki/2015#november), respectively. Examples of the events described in October 2015 and November 2015 Wikipedia articles are presented in Table 4.5.

Figures 4.2 and 4.3 show the Precision and NDCG scores for for varying K Twitter documents. According to Figures 4.2 and 4.3, our proposed framework was found effective and performed well both in the *NDCG* and $Precision@K$ measures. In fact, our framework has discovered many real-world events such as the refugee crisis and its implications, disasters (e.g. Hindu Kush earthquake) and terrorist attacks (e.g. the Ankara suicide bombings, the Paris attacks, the Beirut bombings, etc.), the war against ISIS, as well as many other events and stories compared to Table 4.5.

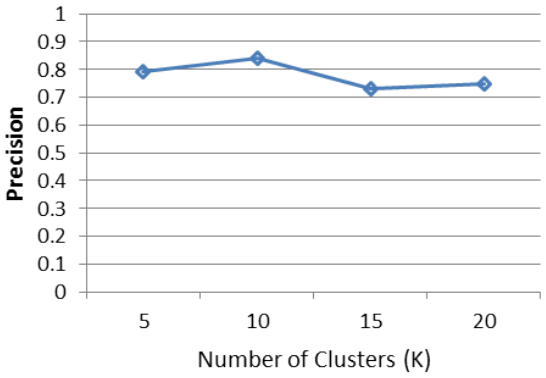


Figure 4.2: $Precision@K$ **of our classification-clustering framework**

Table 4.5: Examples of the events described in October 2015 and November 2015 articles of Wikipedia, which were used as ground truth, for evaluation of the proposed framework..

Date	Events	Description
10 Oct 2015	Ankara bombings	Aseries of suicide bombings kills at least 100 people at a peace rally in Ankara, Turkey and injures more than 400 others.
22 Oct 2015	The refugee crisis	The UN's human rights chief claims the Czech Republic is holding migrants in "degrading" and jail like conditions.
26 Oct 2015	The Hindu Kush earthquake	A magnitude 7.5 earthquake strikes the Hindu Kush region and causes 398 deaths, with 279 in Pakistan, 115 in Afghanistan and 4 in India.
30 Oct 2015	Russian plane crash (terrorist attack)	Kogalymavia Flight 9268, an Airbus A321 airliner en route to Saint Petersburg from Sharm el-Sheikh crashes killing all 217 passengers and 7 crew members on board.
12 Nov 2015	Beirut bombings	Several suicide bombings occur in Beirut, Lebanon, killing 43 and injuring 239. The Islamic State in Iraq and the Levant claim responsibility.
14 Nov 2015	Paris attacks	Multiple terrorist attacks claimed by Islamic State of Iraq and the Levant (ISIL) in Paris, France, result in 130 fatalities.
24 Nov 2015	Russian fighter jet shoot down	Syrian Civil War: Turkey shoots down a Russian fighter jet in the first case of a NATO member destroying a Russian aircraft since the 1950s.
30 Nov 2015	The 2015 COP -21 Conference	The 2015 United Nations Climate Change Conference (COP 21) is held in Paris, attended by leaders from 147 nations.

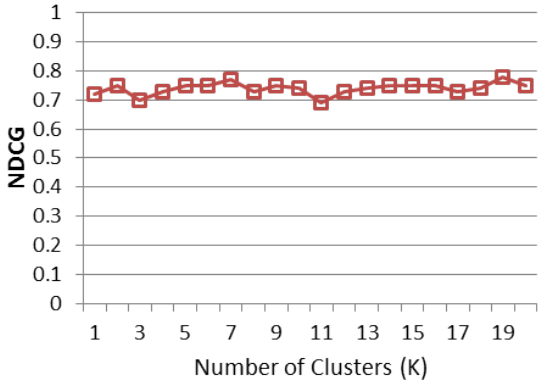


Figure 4.3: *NDCG* at K of our classification-clustering framework

4.6.3 Feature Selection Evaluation

In this section, we present the results for different feature clustering experiments. First, we looked at the performance of the clustering algorithm using a single feature. Then we present the results for the clustering algorithm by combining multiple features. To conclude, we examine the results in more depth by looking at the performance of the framework. In particular, we used accuracy and running time to select the best temporal and spatial setting. Additionally, we used the feature selection method outlined in Algorithm 2 to optimize textual features.

Temporal features

We analyze the efficiency of the proposed temporal features in terms of the event clustering accuracy (A) and the total clustering calculation time (T). We calculated A (Figure 4.4) and T (Figure 4.5) for a range of time windows; 1 minute, 30 minutes, 1 hour, 3 hours, 6 hours, 12 hours and 24 hours. To obtain the best value for temporal features, we had to look at the following optimization problem:
$\{$*Clustering Accuracy* - k . *Clustering Calculation Time*$\}$

where k is the threshold which maximizes the criterion.

The results presented in Figures 4.4 and 4.5 show that the 1-hour time window re-

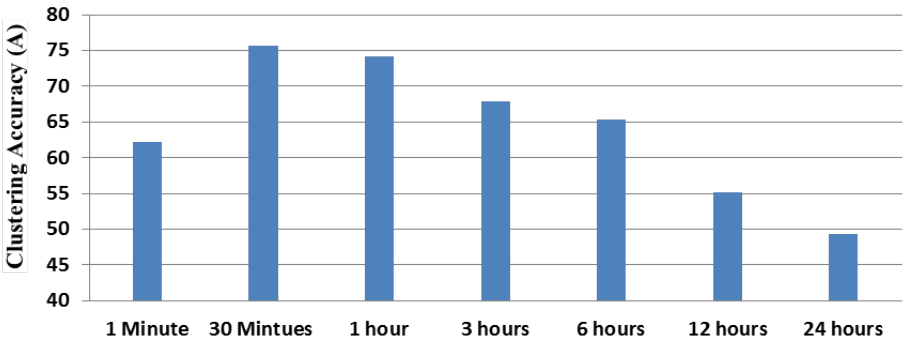


Figure 4.4: Accuracy (A) obtained using various temporal settings

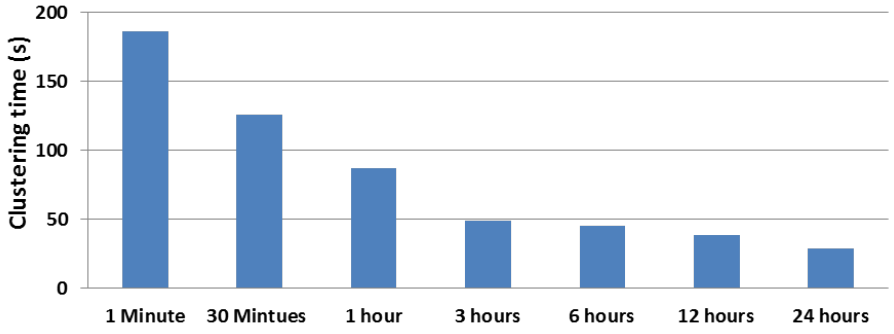


Figure 4.5: Efficiency comparison with various temporal granularities (s)

quires much less computational clustering time than is required for the 1 minute and 30 minute windows, while producing the second best level of accuracy after the 30 minute window. This suggests that tweets published recently are better identifiers of events than older tweets, but also that a lead-in time is required (since 1 minute is too short to provide the same level of accuracy). Clustering the tweets every hour provides a small reduction in the clustering accuracy but significantly reduces the computational processing requirements; therefore for the remaining experiments we set the time window for clustering to 1 hour. Moreover, we find that reporting disruptive events is more likely than other events are to be successful in 1 hour slots.

Spatial features

Here we present the results for different geospatial feature experiments on our clustering algorithm and evaluate them according to the three levels of the Twitter users' location: country, city and neighbourhood. We use the same evaluation approach presented by Mahmud et al. [85] and we use our 3rd dataset (Middle East Twitter Corpus) (2.4.3). We use tweets from the top 100 cities in the middle east region by population (https://www.citypopulation.de/mapindex.html). First, we obtained a bounding box in terms of latitude and longitude for each city using Google's geocoding API (http://code.google.com/apis/maps/documentation/geocoding/). We recorded tweets using the geo-tag filter option of Twitter's streaming API for each of those bounding boxes until we received tweets from 100 unique users in each location. The city corresponding to the bounding box where the user was discovered was assumed to be the ground truth home location for that user. We then used our dataset to extract and collect each user's 100 most recent tweets.

Our final data set contains around 1 million tweets generated by roughly 10,000 users. 220,873 tweets (22%) contained references to countries, cities and neighbourhoods (e.g. street names, neighbourhoods, towns, villages) mentioned in the GATE ANNIE gazetteer (see Section 3.4 for details about the Named Entity Recognition (NER) TwitIE pipeline and the GATE ANNIE gazetteer). From which we had 74,843 tweets (7.5%) contained references to countries, 87,971 (9%) contained cities information, and 54,668 tweets (5.5%) contained references to neighbourhoods. We divided the entire dataset into training (90%) and testing (10%) for 10-fold cross-validation.

We tokenized all tweets in the training dataset, which removed punctuation and other whitespace. All URLs and most tokens containing special characters were then removed, except for tokens that represent hashtags and start with # (e.g., the token #Cairo). Once the tokens have been extracted, we added the Hashtags identifiers which used all tokens that start with the # symbol as terms.

To determine the performance of our clustering algorithm using different location gran-

Table 4.6: Recall comparison using different location granularities.

Location Level	Neighborhood (Local)	City (Intermediate)	Country (General)
Recall	49.04	**53.22**	17.64

ularities, we use the standard evaluation metric Recall (R). Let the total number of users in our test set be n. When this is given to our clustering algorithm, only $n1$ location detections are correct. Hence, we define recall (R) as $\frac{n1}{n}$. Table 4.6 shows the results of the three levels of the Twitter users' location: country, city and neighbourhood our clustering algorithm:

As can be seen from Table 4.6, the city level provides the best overall results, capturing events with around 53.2% accuracy in the city where they occurred. Comparing neighbourhood-level to city-level, we attain similar but slightly better results for the city approach, suggesting that geo-location at the level of neighbourhoods (which is more difficult to obtain via Twitter) is not necessarily required to detect events. An alternative interpretation of this result is that the location detection tools that we used could not handle the misspellings and colloquial terms used for neighbourhood level locations.

Table 4.6 also shows that the performance of the country-level classifier is much worse than that of other classifiers as a result of the users' behaviour. Many users attempt to be more general in their tweets than mere neighborhood-level in order to get more attention. Yet they are trying to be more specific than a country or a region because of the possibility of multiple events in the same time interval. Hence users typically insert hashtags of the city (#abudhabi or #Cardiff) in their tweets rather than country (#UAE or #UK). These results provide some evidence to suggest that events are inherently difficult to identify on the basis of the spatial features on their own.

When it comes to disruptive events, people tend to use city names rather than country or neighbourhood names. For instance, if a user comments on a crime or a terrorist attack or other disruptive event, such as severe weather, s/he tends to include the city

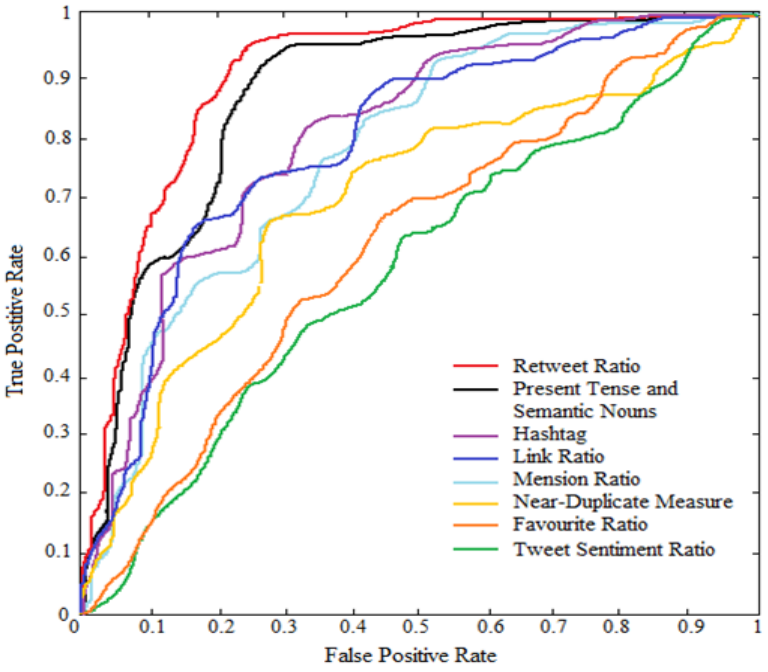


Figure 4.6: ROC curves of the various proposed features

name or s/he might add the geographical hashtag of the city. Consequently, a city-based representation can capture these particular types of event more intuitively.

Textual features

In this section, we investigate the discriminative power of individual textual features in clustering disruptive events in order to show the robustness of each feature individually so the least discriminative features can be removed and thus reduce the computational workload of calculating the results. We use our first dataset 2.4.1 for the evaluation. The results are shown in Figure 4.6 and Table 4.7. Figure 4.6 shows the ROC curve for each feature and Table 4.7 presents the performance results according to the F-measure and the difference between the F-measure of each single feature model to the baseline (Temporal feature is selected as the baseline).

Near-Duplicate measure, Favorite ratio and Sentiment ratio are the least discriminative features, which would suggest that they appear in all types of event, not only in disruptive ones. The bag of words features "Dictionary-based feature", Retweet ratio and

Table 4.7: Comparison of the performance using various textual feature models.

Model	F-measure	F-measure Diff
Baseline (Temporal)	74.14	-
Near-Duplicate measure	74.69	0.55
Retweet ratio	77.57	3.43
Mention ratio	75.73	1.59
Hashtag ratio	77.13	2.99
Link or Url ratio	76.81	2.67
Favorite ratio	74.16	0.02
Tweet sentiment ratio	73.63	-0.51
Dictionary-based feature	77.43	3.29

Hashtag ratio are the most discriminative.

In the middle, there are the averagely influential features, Mention ration and Url ratio, where such tweets require being noticed or promoted by the influential users (celebrities, official accounts or news agents) before their wide spread in Twitter. The time needed to attract potential users may be greater than 1 hour used in our temporal setting. According to our results, the Url or link ratio feature is not as effective as expected. One possible reason is the slow adoption of such tweets especially by influential users and/or the risk of virus or spam links. In a deeper examination of the differences between the features for identifying disruptive events, all the proposed features have positive discrimination power except tweet sentiment, which is investigated in more detail in the next experiment. Combining textual features may lead to better results, for example, if we integrate the Hashtag ratio and the Url ratio, or even combine three features. Such improvement, however, must be reserved for future work.

Tweet sentiment

The first part of this experiment assesses if a tweet has sentiment polarity (0 indicates neutral, 1 indicates either positive or negative sentiment) so as to study the effect of

Table 4.8: F-measures for positive, neural and negative sentiment models, which clearly shows that the negative model outperforms others by at least 1.43%.

Model	F-measure	F-measure Diff.
Positive sentiment ratio	74.27	0.13
Neutral sentiment	74.40	0.26
Negative sentiment ratio	75.83	1.69

sentiments regarding reporting events (reported in Figure 4.6 and Table 4.7). We observe that sentiments in posts are not significant features when reporting events because users are looking for facts and evidence such as photos or videos. Another reason for this is the possibility of negative and positive sentiments of conflict in a cluster, which may cancel each other's influence. Hence, further investigations are carried out in this test to study the influence of positive, negative, and neutral categories in addition to sentiment polarity. The goal of this analysis is to observe whether tweets with positive sentiment undergo a different diffusion process than tweets with negative sentiment.

The goal of the second part of this experiment is to examine whether positive, neutral or negative sentiment tweets have an effect on reporting disruptive events. The main observation made from Table 4.8 is that tweets with negative sentiment lead to a better F-measure than the baseline (temporal) and other sentiment measures. Therefore, negative tweet sentiment has a high adoption rate regarding disruptive tweets, since reporting disruptive events usually involves negative terms and sentiment, whereas events in general can be positive, negative or neutral. Another possible reason is that tweets with negative sentiment are more likely to be retweeted, as shown in [83, 145].

Generally, the supplementary investigation of a tweet's sentiment did not add significant results, suggesting that a fairly small degree of sentiment is typically associated with disruptive events and events in general or that sentiment polarity is hidden within the high volume of tweets. Therefore, the overall level of sentiment in disruptive event tweets was found to be to be quite low and one of their least important features. The

Table 4.9: The most effective textual features (above 1.50 differences)

Rank	Feature	F-measure Diff
1	Retweet,ratio	3.43
2	Dictionary-based feature	3.29
3	Hashtag ratio	2.99
4	Link or Url ratio	2.67
5	Negative sentiment ratio	1.69

limitations of this feature derive from two factors; first, the length constraint of a tweet makes characters very expensive, hence not much sentiment can be afforded. The second reason is related to the technique used in capturing and detecting a sentence's polarity which still is open to challenge and should be further improved.

Ranking top textual features

After investigating each participation feature individually and the further investigation of the sentiment analysis, it was found that features with less than 1.60 differences are not useful; therefore we discard them for the optimum use of textual feature. Only features with high difference (which are most influential) are used to identify disruptive event tweets. The ranking of the most influential textual features is presented in Table 4.9.

Obtaining the optimum model

We use a unigram model as our baseline for this experiment, which is a bag-of-words textual features model (the dictionary-based feature). Figure 4.7 compares the performance of various models: first, we use individual feature models: temporal, spatial and textual. The temporal model uses the 1-hour setting, the spatial model implements the city-level setting and the textual model uses all the features from Table 4.7. Second, a combination of features model: (Temporal + Spatial), (Temporal + Textual) and (Temporal + Spatial +Textual). Third, to build our optimized model we make use of the 1-hour temporal feature, the city-level spatial feature and the most effective textual

features from Table 4.9.

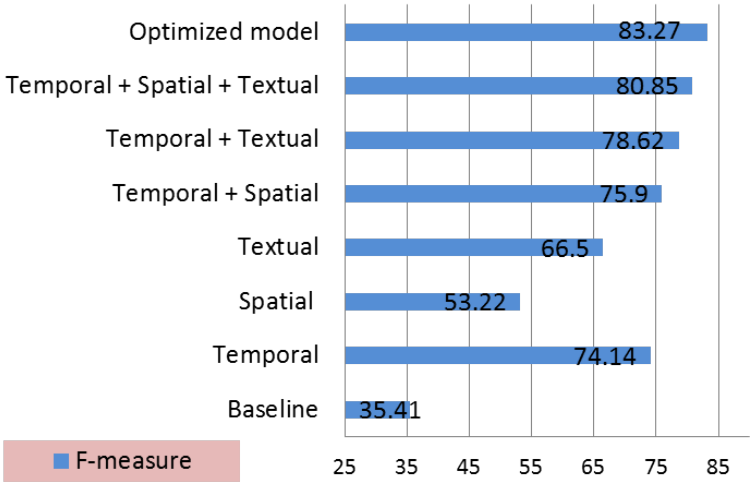


Figure 4.7: Comparison of different models on the event identification task according to the F-measure. Higher is better..

Overall, while each feature set is individually significantly better than the baseline, it is the temporal feature that substantially outperforms all the others, obtaining a performance score of 7.64% over textual features and 20.92% over the spatial features. As shown in Experiment 1, using a 1-hour time window is the most effective in detecting disruptive events. This effect is less when it comes to spatial and textual features. Using the textual feature set without temporal features, we are still able to obtain reasonable performance (66.5%), but this is not as distinctive for disruptive events as when the temporal feature is applied. That is emphatically not the case when using only spatial features, because it is a weak indicator to implement on its own.

From these results it is clear that combining temporal and spatial features gives the best of both with much better performance of F-measure (75.9). More interestingly, integrating the temporal and textual features results in better system performance than using each feature independently. It also outperforms the combination of temporal + spatial by 2.72.

A combination of all three features results in the best performance, but a further invest-

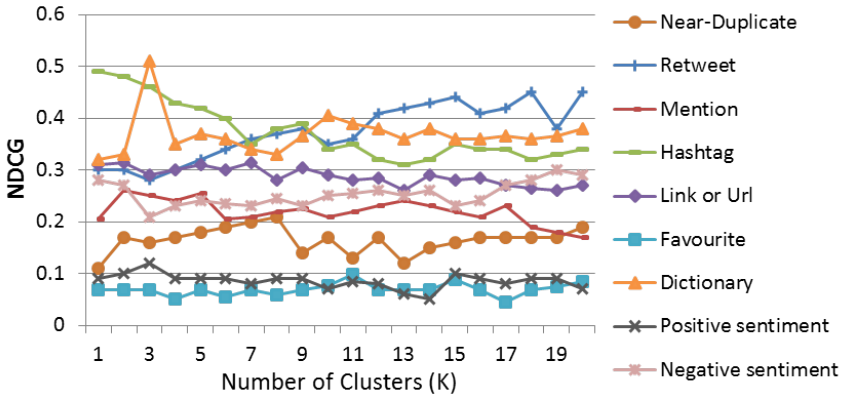


Figure 4.8: Performance of various proposed features

igation which removed unnecessary textual features gave the best model performance overall. The optimized model achieved an F-score of 83.27, higher than the combination of all the features. These results support our claim that not all features should be expected to improve a system's performance; instead they all contribute differently to detecting disruptive events.

4.6.4 Feature Selection using *NDCG* scores

In the previous section, we investigated the discriminative power of the temporal, spatial, and textual features in identifying disruptive events. This section aims to address the same issue by investigating the utility of the these features using *NDCG* scores, while employing a different dataset (the Middle East dataset 2.4.3). We expect to obtain similar results to those obtained using the other ranking approach. The results are shown in Figure 4.8 which illustrates the *NDCG* scores for each feature.

As can be seen in Figure 4.8, we obtain the same results as those we discussed in the last section. The near-duplicate measure, the favourite ratio and the positive sentiment ratio are the least discriminative attributes, whereas the dictionary-based model, the retweet ratio and Hashtag ratio are the best discriminators. The retweet ratio suggests that other users pick up on event commentaries and propagate them further through the

network. Linking content features such as Hashtags and URLs is also highly predictive of events. Given that a hashtag is potentially a topic indicator and that Twitter URLs enrich tweet content with more information, the existence of Twitter URLs can be used as a feature for disruptive event identification.

Hashtags and URLs were already shown previously to have a high correlation with retweetability [111]. Ma et al. [83] found that tweets with URLs and hashtags were more likely to be retweeted. Another important observation in Figure 4.8 is that the negative sentiment model outperforms the positive sentiment model in *NDCG* scores, due to the fact that negative sentiment posts have a high adoption rate as regards reporting disruptive events and that negative sentiment is dominant in tweets about emerging events, as discovered by [55].

Once again, we use feature selection to build our optimized model, this time using *NDCG* scores. Figure 4.9 compares the performance of various models: first, we use individual feature models: temporal, spatial and textual (the textual model uses all the features from Figure 4.8). Second, we use a combination of all features. Finally, we use the temporal feature, the spatial feature and only the most effective textual features from Figure 4.8 (above 0.25 in the *NDCG* evaluation measure) to build the optimized model.

The temporal feature model substantially outperforms spatial and textual models, obtaining a performance score of about 13.2% over textual features and about 38.7% on average over spatial features. Hence, the temporal feature can be judged the most effective in detecting events. Using the textual feature model, we are still able to obtain a reasonable performance of on average, 40% content about an event, provided there is situational awareness information about that event. However, it is emphatically not the case when using the spatial feature in isolation, leading to the conclusion that spatial features are weak indicators to implement on their own.

A combination of all three features results in the best performance, because it gives the best of all features with a much better performance, but further investigation removing

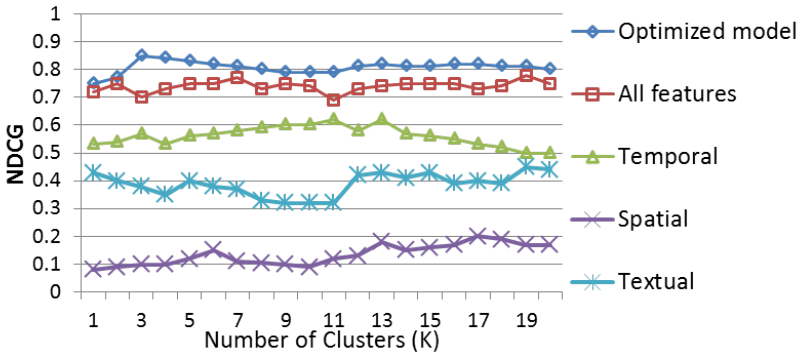


Figure 4.9: Comparison of different models for the event identification task according to *NDCG* scores.

unnecessary textual features (such as the near-duplicate measure, favourite ratio, mention ratio and the positive sentiment ratio) yields the best model performance (average 0.802 *NDCG* score), which confirms the results that were previously presented.

4.6.5 Comparison with Leading Event Detection Approaches

In order to validate our approach, we evaluated it against other approaches in the context of the MediaEval2012 Social Event Detection (SED) international benchmark [109]. The SED competition comprised three challenges on a common test dataset of images with their metadata (timestamps, tags, geotags). The goal of the first challenge in the test collection was to identify public technical events, such as exhibitions in Germany. The goal of the second challenge was to find all the soccer events that had taken place in Spain. The goal of the third challenge was to find demonstration and protest events in Spain (see Section 2.4.4).

Evaluation of the submissions to the SED task was performed by the organizers using ground truth that came in part from the EventMedia dataset [66] (for Challenge 1), and in part from the results of a semi-automatic annotation process carried out with the CrEve tool [166] (for all three challenges). Two evaluation measures were used: a) the F-score for the retrieved images, and b) the Normalized Mutual Information

(NMI) [128] that compared two sets of photo clusters (where each cluster comprised the images of a single event), jointly considering the quality of the retrieved photos and their assignment to different events. Both evaluation measures received values in the range [0-1] with higher values indicating better agreement with the ground truth. Evaluation measures were calculated both per challenge and on aggregate.

The evaluation criteria for each submission took into account the number of detected events (out of all the relevant events in the test set) and the number of correct/incorrect media detected for these events. What we were looking for was a set of photo clusters, each cluster comprising only photos associated with a single event (thus, with each cluster defining a retrieved event). We compared the overall performance of our system using the optimized model in terms of Precision, Recall, F-Measure and NMI to the performance of three leading systems:

- Becker et al. [16]: this is the re-implemented CLASS-SVM and online-clustering method.
- Vavliakis et al. [148]: We used the results of this method using only topics automatically created by the LDA process for topic discovery.
- Schinas et al. [131]: this is based on the Structural Clustering Algorithm for Networks (SCAN) algorithm.

According to Table 4.10, our proposed methodology is effective and outperforms almost every other approach that participated in SED. For Challenge 1, our framework seems to outperform all other algorithms both in the F-measure scores and the NMI evaluation scores. This was the case even though the topics in the challenge are about technical events (mainly conferences) described by a diverse vocabulary and often including relatively few photos; in consequence they raise topics that contain concepts from irrelevant photos which explain the poor performance of the other methods.

The topics in Challenges 2 and 3 are relatively easy to identify automatically, in particular by the LDA method. While our method is comparable to that of Vavliakis et

Table 4.10: Results of the proposed approach against other event detection approaches using MediaEval2012 Detection Task..

	Precision	Recall	F-score	NMI
(a) 1st challenge: Technical events in Germany				
Ours	66.71	**64.93**	**65.81**	**0.5528**
Vavliakis et al. [148]	**80.98**	19.56	31.10	0.2112
Schinas et al. [131]	59.12	11.91	18.66	0.1877
Becker et al. [16]	43.81	46.83	45.22	0.3614
(b) 2nd challenge: Soccer events in Madrid and Hamburg				
Ours	87.19	**86.95**	**86.98**	0.6845
Vavliakis et al. [148]	**91.21**	79.71	84.00	**0.7684**
Schinas et al. [131]	87.05	66.56	74.64	0.6745
Becker et al. [16]	76.74	79.18	78.13	0.7558
(c) 3rd challenge: Protest events in Madrid				
Ours	82.46	**95.67**	**88.94**	0.5326
Vavliakis et al. [148]	**90.76**	84.20	86.11	0.3302
Schinas et al. [131]	88.43	54.61	66.87	0.4654
Becker et al. [16]	79.72	82.66	81.15	**0.5447**

al. in its precision, ours has a much higher retrieval component in terms of the recall (86.95 vs. 79.71), which is an advantage in our settings because we assume that a decision maker is interested in seeing pure clusters with only a few spurious examples. Our system also much outperforms that of Schinas et al., in all challenges in terms of the F-measures and NMI. This difference in performance has a reasonable explanation. The splitting and merging of events into smaller clusters (in Schinas et al.) is limited to short ranges of duration in the Structural Clustering Algorithm.

In addition, classification after clustering has a crucial impact on performance in terms of quality, above all when the time windows are short. Many of the small clusters are filtered out since they do not exceed the predefined thresholds and are considered

non-relevant events (noise). This eliminates many of them when noise is eliminated, which confuses the scoring and ranking of event detection. This explains why our system outperforms Becker et al. approach. Furthermore, it should be noted that the high performance of our framework in identifying disruptive events that is reflected in Challenge 3 shows the effectiveness of our online clustering methodology as well as the feature optimization approach.

4.6.6 Arabic Event Detection in Social Media

We repeated the same set of experiments with the same settings using pure Arabic language tweets in [6] using dataset 2.4.2. We obtained acceptable results which demonstrate the effectiveness of the proposed algorithm, but the results were slightly lower than the results reported for tweets in English and other Latin script languages. In fact, Arabic poses many challenges for data mining tasks [37]. Most of these challenges are due to the orthography and morphology of the language. It is true that some of these challenges are shared with other languages, but Arabic exhibits considerable complexity in the move from theoretical to computational linguistics. Furthermore, processing the language becomes even more challenging given the language used in social networking and microblogging sites, where dialects are heavily used. These dialects may differ from standard Arabic in vocabulary, morphology, and spelling and most do not have standard spellings. Other Arabic accounts use a mixture of Latin and Arabic characters (Arabizi) [6].

4.6.7 Case Study: 2011 Riots in England

In order to further validate our approach, we evaluated it against other leading approaches using the 2011 riots dataset. We used the model based on the training set (from our first dataset 2.4.1), which was evaluated in the previous section. We do not train specifically on the riots data - thus we are testing the generalizability of our model

for a real-world dataset. Our evaluation is based on high quality ground truth data from public Metropolitan Police Service (MPS) reports. On August 4 Mark Duggan was shot in Tottenham by police officers. On the evening of 6th August, following a peaceful protest march to a Tottenham police station, organised by the victim's friends and family, the first outbreaks of public disorder occurred. Then they quickly spread across London and to other cities in England and the levels of crimes and offending increased dramatically, to include looting, violence, burglary, arson and other disorder-related offences, which make this case study and our collected dataset ideal for large scale event detection (riots), and smaller disruptive event detection - from small-scale looting incidents in local shops to one of the largest cases of arson in Europe [92]. In terms of social media, the MPS was clear that at that time its capability for using social media networks as engagement was still in its infancy [92].

We compare the output of our framework to similar existing methods namely, Spatial LDA [106] by Pan et al., unsupervised methods by Becker et al. [16] and Zubiaga et al. [167]. Spatial LDA [106], proposed by Pan et al., combines an LDA model [20] with temporal segmentation and spatial clustering. Becker et al. [16] uses an unsupervised clustering technique to group topically similar tweets together, and computed features (temporal, social, topical, and Twitter-specific) that can be used to train a classifier to distinguish between event and non-event clusters. Zubiaga et al. [167] explores the real-time summarization of scheduled events using a two-step system: (i) sub-event detection and (ii) tweet selection. The first step is based on detecting peaks (reflected as peaks in the histogram of tweeting rates) with an enhancement of two ideas; the sudden increase in the tweeting rate and outlier detection. The tweet selection step selects a representative tweet after ranking all the tweets that were sent in the sub-event. They use the Kullback-Leibler divergence (KLD) weighting scheme for the tweet ranking.

In general, we are not interested in identifying the event type as our framework is general and can be used for various events' types, but for the evaluation purposes we

Table 4.11: Comparison of approaches for disruptive event detection.

	Numberof Real-world events identified				
Incident Type	Police Intelligence	Ours	Becker et al.	Spatial LDA	Zubiaga et al.
Car Accident	-	285	108	74	92
Fire Incident	311	214	121	186	127
Shooting	4	3	1	3	0
Stabbing	5	4	0	3	1
Protest	187	143	106	163	32

	Ours			Becker et al.			Spatial LDA			Zubiaga et al.		
Incident Type	P	R	F	P	R	F	P	R	F	P	R	F
Fire Incident	74.64%	68.81%	71.61%	39.77%	38.91%	39.34%	60.09%	59.81%	59.95%	42.26%	40.84%	41.54%
Shooting	57.41%	75.00%	65.04%	30.22%	25.00%	27.36%	52.75%	75.00%	61.94%	8.43%	0	0
Stabbing	63.64%	80.00%	70.89%	3.55%	0	0	45.18%	60.00%	51.55%	18.29%	20%	19.07%
Protest	77.82%	76.47%	77.14%	53.85%	56.69%	55.23%	38.78%	33.67%	36.04%	32.67%	17.11%	22.46%

used the Dictionary-based feature (See Section 4.5.2) to identify the event-type. A multi-class classification model is implemented, where tweets are classified to one of the following six classes: Car Accident, Fire Incident, Shooting, Stabbing, Protest, and other (any other disruptive events), as defined in Table 4.11.

All three rival methods have been successfully applied to event detection and thus we aim to outperform them using our proposed online clustering algorithms and the temporal TF-IDF summarization (see Chapter 5). Table 4.11 presents the performance of the comparative experiments in terms of Number of real-world events (as reported to MPS) detected, system Precision, system Recall and the F-measure. Precision is defined as the fraction of the retrieved documents that are relevant. Recall is defined as the fraction of the relevant documents retrieved to the total number of relevant documents that should have been returned and the F-measure is defined as a harmonized mean of precision and recall [157, 132, 106, 16].

According to Table 4.11, our proposed methodology is effective and outperforms other approaches. This is the case even though the topics in the *Riots 2011* dataset are disruptive events described by a diverse vocabulary and often comprising relatively few

posts per incident. The MPS did not include car accidents and vehicle damage related to the riots; hence we could not compute the recall measure. However, the number of events detected indicate that our framework can detect four times more real-world incidents than Spatial LDA can and at least twice as well as Becker et al.

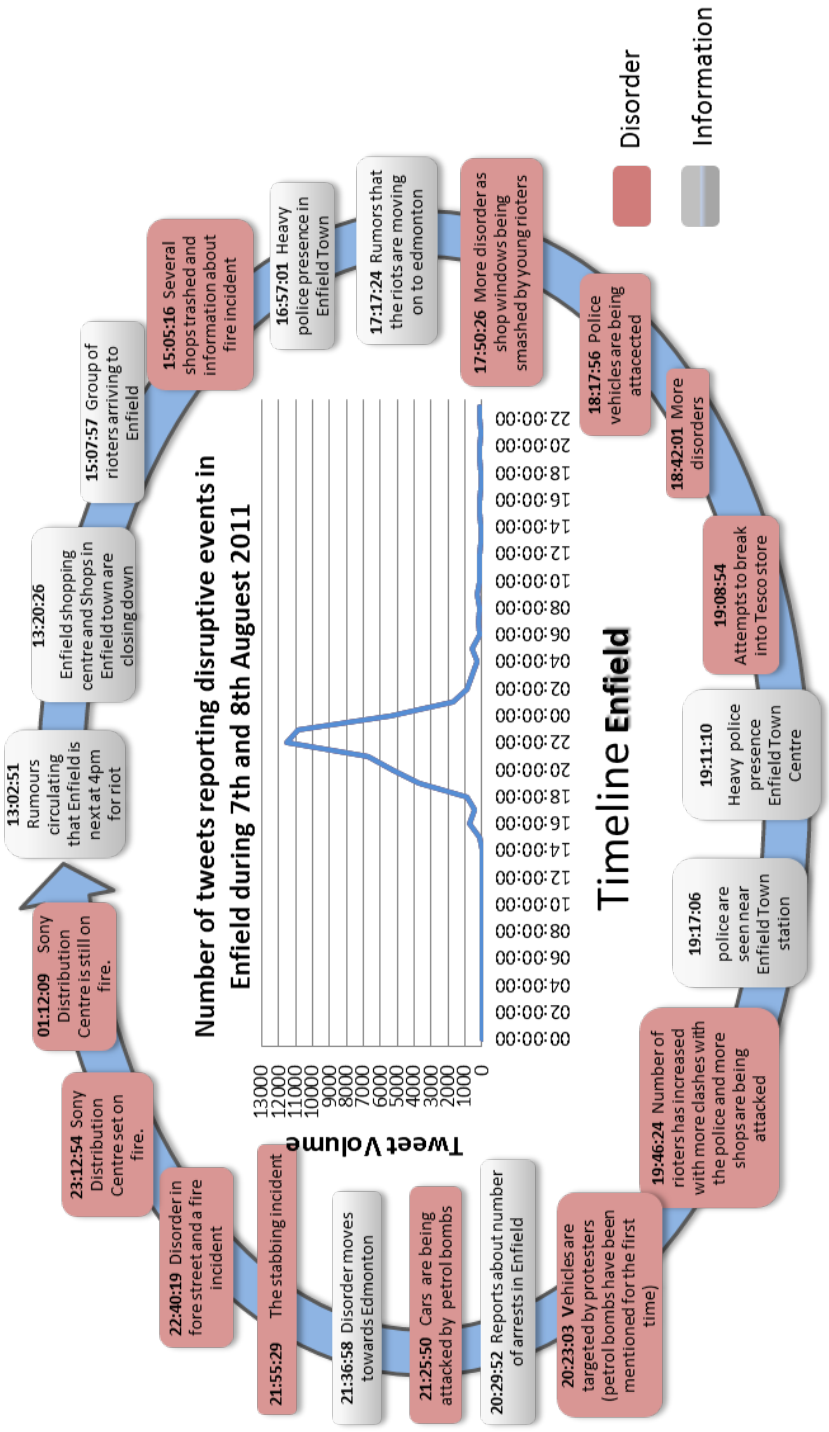
We offer the following explanation as to how the systems we tested could have been impaired; first, not all the events mentioned in the MPS report using traditional intelligence are reported in the social media and vice versa. Second, the presence of rumours and false information in the 2011 England riots and generally in emergencies and disasters is another issue which affects the reported results negatively. The detection of rumours in the social media is beyond the scope of this paper and must be reserved for future work. By studying the life cycle of several rumours and also by investigating the propagation, we may be able to effectively identify social media rumours.

In addition, classification after clustering has a crucial impact on performance in terms of quality, especially with moving time windows. Many of the small clusters are filtered out since they do not exceed the predefined thresholds and are considered non-relevant events (noise). This eliminates many of them together with noise, which confuses the scoring and ranking of event detection and serves to explain why the performance of the approach by Becker et al. is less than ours. The results in Table 4.11 also show that the Spatial LDA approach outperforms Becker et al.'s system only in major events such as fires. However, it fails to achieve such results in other cases because tweets are short and any collection of tweets per hour may contain many more topics referring to multiple small-scale cases such as car accidents or small group protests. Zubiaga et al. [167] approach and similar systems such as [138, 161] are limited to scheduled events such as soccer games; moreover, they require a fixed starting time in order for the system to start looking for new sub-events. This explains why our system performs better than Zubiaga et al. event identification approach.

Visualizations are arguably well suited to displaying real-time disruptive events sensed from social streams. We visualize the real-time output from our system alongside the

post-event visualization provided by MPS in their public report [92] in Figure 4.10. In view of space limitations, we present only the results for the Enfield borough, although the MPS report [92] presents the results in three case studies (Enfield, Croydon and Wandsworth). As can be seen from Figure 4.10 most of the disruptive events including looting, arson, violence, etc. were successfully identified and monitored in real-time and in some cases our system provided information ahead of traditional intelligence. Furthermore, Table 4.12 presents the time difference between the identification of a disruptive incident by our framework and the corresponding police information. The columns show the time of discovery of the events by the summarization of our system (see Chapter 5), the time of the report of intelligence by officials and by how much Twitter leads police intelligence. Entries marked in bold occur first.

From Table 4.12, we observe that the Twitter information was extracted by our system most often ahead of police sources. The latter leads only twice, in both cases by 10 minutes. The delay can result from the time taken by a user to post a tweet, the time to index the post in the Twitter servers, and the time taken to make queries by our system. In fact, our system detected all of the disruptive events which were reported by officials far faster than they did, on average 23 minutes faster. The benefits of identifying accurate intelligence on the disorder are much greater if it is received in real time because this enables decision makers to move ahead of the crisis in such events. These results support the hypothesis that information extracted from the social media can be used as a valuable additional source of intelligence as well as bridging the gap between the use of “big data" and modern policing in the interests of situational awareness and enhanced public safety and decision making.


Timeline Enfield
Disorder
Information
Number of tweets reporting disruptive events in Enfield during 7th and 8th Auguest 2011
Tweet Volume
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
00:00:00
02:00:00
04:00:00
06:00:00
08:00:00
10:00:00
12:00:00
14:00:00
16:00:00
18:00:00
20:00:00
22:00:00
00:00:00
02:00:00
04:00:00
06:00:00
08:00:00
10:00:00
12:00:00
14:00:00
16:00:00
18:00:00
20:00:00
22:00:00
01:12:09 Sony Distribution Centre is still on fire.
23:12:54 Sony Distribution Centre set on fire.
22:40:19 Disorder in fore street and a fire incident
21:55:29 The stabbing incident
21:36:58 Disorder moves towards Edmonton
21:25:50 Cars are being attacked by petrol bombs
20:29:52 Reports about number of arrests in Enfield
20:23:03 Vehicles are targeted by protesters (petrol bombs have been mentioned for the first time)
19:46:24 Number of rioters has increased with more clashes with the police and more shops are being attacked
19:17:06 police are seen near Enfield Town station
19:11:10 Heavy police presence Enfield Town Centre
19:08:54 Attempts to break into Tesco store
18:42:01 More disorders
18:17:56 Police vehicles are being attacected
17:50:26 More disorder as shop windows being smashed by young rioters
17:17:24 Rumors that the riots are moving on to edmonton
16:57:01 Heavy police presence in Enfield Town
15:05:16 Several shops trashed and information about fire incident
15:07:57 Group of rioters arriving to Enfield
13:20:26 Enfield shopping centre and Shops in Enfield town are closing down
13:02:51 Rumours circulating that Enfield is next at 4pm for riot

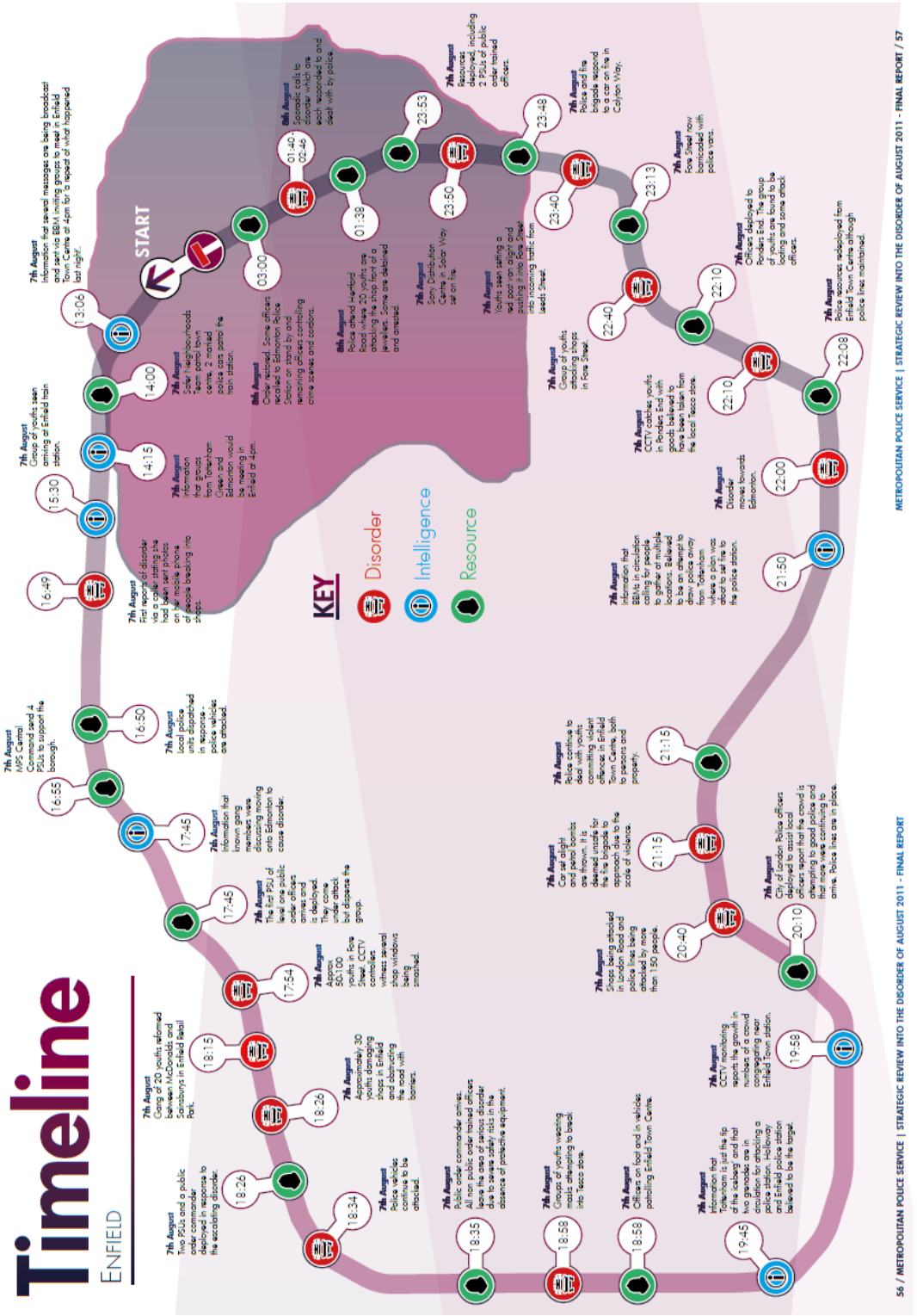


Figure 4.10: Comparison of disruptive events obtained by our framework (top) and MPS (bottom) for Enfield borough. The source of the bottom image is from the Metropolitan Police Service (MPS) report [92].

Table 4.12: Disruptive event exploration using police intelligence and by our framework for Enfield borough on August 7th 2011 (+ when Twitter leads).

Police Intelligence	Summarization by our system	Time/ Police	Time/ Our system	Lead
Information that several messages are being broadcast to meet in Enfield Town Center at 4pm for a repeat of what happened last night.	Rumours circulating Enfield is TONIGHT. #Tottenham #Riots	13:06	**13:02**	+0:04
Information that groups from Tottenham Green and,Edmonton would be meeting in Enfield at 4pm.	#rumour has it #enfield riot k.o's at 4!	14:15	**13:19**	+1:04
Group of youths seen arriving at Enfield train station.	The rioters are now in Enfield and Edmonton. #londonriots	15:30	**15:07**	+0:23
First reports of disorder via a caller stating she had been sent photos on her mobile phone of people breaking into shops.	Ok its officially kicking off in #Enfield Town, one fire and hmv has been smashed in, people coming from all over london to #loot.	16:49	**15:37**	+1:12
Information that known gang members were discussing moving onto Edmonton to cause disorder.	not feeling the rumors that the rioters are looking to move to edmonton and #enfield town. DON'T YOU PEOPLE THINK YOU'VE DONE ENOUGH!!!!	17:45	**17:17**	+0:28
The first PSU of level one public order officers arrives and is deployed. They come under attack.	ok so 9 police vans just drove past my house! ok make that 10! #enfield	17:45	**17:39**	+0:06
Approximately 30 youths damaging shops in Enfield and obstructing the road with barriers.	RT Police car wrecked in Enfield - most rioters looked under 16, lots of young girls throwing concrete slabs through shop windows. #enfield	18:26	**17:50**	+0:36
Police vehicles continue to be attacked.	police car trashed RT @XXXXX: BREAKING: This just happened at #EnfieldTown; Police outnumbered once again; http://yfrog.com/kf4rlauj	18:34	**18:17**	+0:17
Groups of youths wearing masks attempting to break,into Tesco store.	Police horse vans in #enfield tesco car park http://yfrog.com/h7eyhirj	**18:58**	19:08	-0:10
CCTV monitoring reports the growth in numbers of a,crowd congregating near Enfield Town station.	#Enfield Police attacking riotmob with batons and,dogs in the town. Over 230+ riot mobs in #Enfield town	19:58	**19:49**	+0:09
Car set alight and petrol bombs are thrown. It is deemed unsafe for the fire brigade,to approach due to the scale of violence.	*ALERT* Protestors are throwing petrol bombs on passing cars on the A10 from #Tottenham to #Enfield. Avoid the road.	**21:15**	21:25	-0:10

Not reported in the official report as it might not be relevant to the Riots	Teenager stabbed outside #Edmonton WorkingMen's Conservative Club. Medics on scene. #Enfield		21:58	
Disorder moves towards Edmonton.	I hear edmonton is next #enfield	22:00	**21:46**	+0:14
CCTV catches youths in Ponders End with goods believed to have been taken from the local Tesco store.	#Enfield disturbances now spreading to Ponders End #PondersEnd	22:10	**21:36**	+0:34
Group of youths attacking shops in Fore Street.	Carphone warehouse getting smashed up in #edmonton, ridiculous!!	22:40	**21:54**	+0:46
Youths seen setting a red post van alight and,pushing it into Fore Street into incoming traffic from Leeds Street.	Car near fore street about to explode, about 50 man standing off with police. #Edmonton	23:40	**23:12**	+0:28
Sony Distribution Center in Solar Way set on fire.	40 firefighters at a fire in a warehouse on Solar Way in Enfield. #LondonRiots #Enfield	23:50	**23:19**	+0:31

4.7 Summary

In this chapter, we have developed a general on-line clustering approach for the purpose of identifying events on microblogging services which aims to confront many of the challenges and provide a system to detect large-scale events as well as small-scale ones. This chapter surveyed several features that have been cited in the recent literature in order to enrich and optimize the detection performance. Specifically, the temporal features in social media sites such as Twitter are a significant and revealing source of information for, and about, event detection. Extensive experiments were conducted to evaluate the effectiveness of the proposed clustering approach using large real-world datasets. Our findings suggest that our framework yields better performance than many leading approaches in the detection of real-time events.

In order to improve the retrieval performance, we have presented an extensive analysis of various features related directly to Twitter data and shown how they can be used to discriminatively distinguish between disruptive events and other events. The results

make it clear that it is inadequate to consider temporal, spatial, or content-based aspects in isolation. Rather, a combination of features covering all these aspects leads to a robust system which makes possible the best event detection results. Our optimized approach improved the identification accuracy from 80.85% to 83.27%, which is a significant result for event identification tasks. We believe that such spatio-temporal textual knowledge is a crucial asset for many applications, including many computer decision support systems and uses of artificial intelligence.

Regarding the textual features, we show that the Dictionary-based model, Retweet ratio and Hashtag ratio are the most discriminative features, suggesting that references to present time and references to descriptive terms (e.g. live, breaking, etc.) are good discriminators. The retweet ratio suggests that other Twitter users pick up on event commentaries and propagate them more often through the network than non-event tweets. Linking content features, such as Hashtags and URLs, are also very predictive of disruptive events and made more discoverable via a self-defined topic discriminator in the form of a Hashtag.

Chapter 5

Representation and Summarization

5.1 Introduction

In previous chapters, we explored the first 4 steps of our proposed framework; data collection (Chapter 2), preprocessing (Chapter 2), classification (Chapter 3), and clustering (Chapter 4). In this chapter, we focus on the last step of our framework; summarization. Therefore, this chapter focuses on the problem of selecting Twitter content from event clusters. We address this problem by automatically selecting most representative messages that best represent the event, which was identified using an online clustering technique that groups together topically similar Twitter messages (Chapter 4).

Due to the vast number of posts published by hundreds of millions of users, digging through the noise and redundancy to extract and summarize the informative aspects of the content is a very challenging task. Moreover, the Twitter API allows users to see only the most recent posts on a topic, in chronological order; it does not present posts in order on the basis of relevance. This motivates the need for new automatic summarization systems that will give decision makers informative summaries of user-generated content that support intelligence gathering and augment traditional sources of situational information. Such posts are likely to be multilingual so the system should also be capable of handling this. In addition these systems should be able to handle information flows in real time - as events unfold.

The approaches to automatic summarization vary with the definition of "summary". Generally, text summarization is the process of automatically creating a compressed version of a given text so as to provide useful information for the user [38]. The two main approaches to text summarization are extractive and abstractive. *Extractive* methods select a subset of words, phrases, or sentences from the original document to form a summary [11]. In contrast, *abstractive* summarization rephrases the text and drops much of its original phrasing in [46].

In summarizing Twitter streams, summarization (or tweet representation, which is an extractive method) can be viewed as a problem of automatically selecting the most important tweets from one or more event clusters [59]. In the context of the social media, therefore, a summary can be considered a single post if we assume that a post is a summary of what a user observes in the real world about a topic or how she contributes to an ongoing event by giving more details about it. Having many tweets from many users which contribute to the same topic, our online clustering technique groups together topically similar Twitter messages in the same cluster. Then we select the most relevant and useful tweets from this cluster using our proposed summarization techniques. Furthermore, the problem can be defined as a ranking task, in which all the tweets about a certain topic/event are ranked according to a weighting measure. Hence, summarization can be divided into two steps: (i) event detection, followed by (ii) tweet selection [17, 138, 167].

In this chapter, we describe related efforts on event summarization using Twitter and various techniques for selecting most representative Twitter messages for real-world events (Section 5.2). In Section 5.3, we propose three techniques that focus on summarizing Twitter messages; they correspond to events to improve event visualization and analytics. Our methods are language independent, satisfy the real-time requirement and are suitable for the huge quantity of data. We use a frequency-based method, voting approach, and centrality-based approach to select messages that represent an event with high quality, strong relevance and are useful to people looking for inform-

ation about an event. We evaluate our proposed techniques, in Section 5.4, using a real-world dataset of Twitter messages according to ROUGE-1 (see Section 5.4.1) as well as metrics (quality, relevance and usefulness) [17]. Our methods are tested on English, Arabic and Japanese language tweets to test their applicability across multiple languages. We also compare their performance with a number of recent and leading summarization systems, including Becker et al. (centroid method) [17], Zubiaga et al. (sub-event detection then tweet selection) [167], Xu et al. (graph-based approach) [160] and Hybrid TF-IDF (term frequency summarization approach) [59].

The research question we aim to address in this chapter is:

RQ4 Can we summarize events to enable decision makers to read effectively only high quality summaries of most representative posts from Twitter?

5.2 Related Work: Summarization Approaches

The automatic summarization and detection of topics from the social media have often been addressed. Many of these approaches are inspired by previous work on automatic text summarization. Nenkova and McKeown [94] have made an extensive survey of text summarization techniques. In this paper we focus only on systems for summarizing microblog events.

The centroid-based method is one of the most popular extractive summarization methods. MEAD [117] uses an implementation of the centroid-based method that scores sentences according to their sentence-level and inter-sentence features, including cluster centroids, position, Term Frequency - Inverse Document Frequency (TF-IDF), etc. Moreover, MEAD is a flexible platform for multi-document multi-lingual summarization which is publicly available. Similarly, Becker et al. [17] presented three centrality-based approaches (LexRank, Degree and Centroid) to select high quality messages from clusters. These authors found that the centroid approach, which computes the

cosine similarity of the TF-IDF representation of each message to its associated event cluster centroid, outperformed other approaches in three metrics: *quality*, *relevance*, and *usefulness*.

Another approach is the graph-based LexRank, which was introduced by Erkan and Radev [46]. The LexRank algorithm computes the relative importance of sentences in a document (or a set of documents). Then it creates an adjacency matrix among the textual units and finally computes the stationary distribution, treating it as a Markov chain. In their study, they showed that the similarity graph of sentences provides a better view of important sentences than the centroid approach does. The TextRank algorithm [90] is another graph-based approach, that implements two unsupervised approaches for keyword and sentence extraction in order to find the most highly ranked sentences in a document using the PageRank algorithm [23]. Recently, Xu et al. [160] extended the Pagerank ranking algorithm and investigated a graph-based approach which leverages named entities, event phrases and their connections across tweets to create summaries of variable length for different topics. Moreover, Olariu [101] proposed a graph-based abstractive summarization scheme where bigrams extracted from the tweets are viewed as the graph-nodes.

SumBasic [147] is a simple, yet high-performing summarization system based on term frequency. The authors empirically showed that words that occur more frequently across documents are more likely to appear in human generated multi-document summaries. Most recently, Inouye and Kalita [59] developed a new method called "Hybrid TF-IDF", which ranks tweet sentences using the TF-IDF scheme and produces better results for microblogs summarization than all the above-mentioned summarization approaches.

In feature-based approaches, a variety of statistical and linguistic features have been extensively investigated. For example, Sharifi et al. [137] proposed a phrase reinforcement (PR) algorithm to summarize the Twitter topic in one sentence. They extracted keyphrases by exploiting textual redundancy and selecting common sequences of

words. The summary sentence is selected as one of the highest weighted paths in the graph. Nichols et al. [97] extended this idea and generated a journalistic summary method for events in world cup games by employing a phrase graph algorithm on the longest sentence in each tweet. More finely-grained summarization was proposed by considering the detection of sub-events and combining the summaries extracted from each sub-topic (tweet selection, tweet ranking) [97, 138, 167, 161, 28].

For example, Nichols et al. [97] and Arkaitz et al. [167] focused on real-time event summarization, which detects the sub-events by identifying those moments where the tweet volume has increased sharply, and then uses various weighting schemes to perform tweet selection and finally generates an event summary. Shen et al. [138] present a participant-based approach to event summarization. First, the participants of the event are detected and then a mixed model is applied to detect sub-events at the participant level. Finally, the tf-idf centroid approach is used to select a tweet for each detected sub-event. Similarly, Chakrabarti and Punera [28] propose the use of a Hidden Markov Model to obtain a time-based segmentation of the stream that captures the underlying sub-events. A key limitation is that these algorithms can be applied to periodic events only, such as sports events, but not to longer term events or aperiodic events.

Several previous works have leveraged the importance of monitoring the evolution of an event. For example, Ng et al. [95] derived three features from timelines and used them in supervised learning to enhance multi-document summarization (MDS). Lin et al. [78] proposed a language model with dynamic Pseudo Relevance Feedback (PRF) to obtain relevant tweets, and then generated storylines via graph optimization. In [33], the authors (Chua and Asur) proposed two topic models (Decay Topic Model (DTM) and the Gaussian Decay Topic Model (GDTM)) that take advantage of temporal correlations in the data to extract relevant tweets for summarization. Other work has characterized the temporal patterns of tweets. Wang et al. [155] proposed a prototype called Sumblr which supports the continuous summarization of a tweet stream. Sumblr

employs a tweet stream clustering algorithm to compress tweets into Tweet Cluster Vectors (TCVs) and maintains them online. Then it uses a TCV-Rank summarization algorithm to generate online summaries and historical summaries of arbitrary time durations.

Other researchers have proposed models for the purpose of summarizing micro-blog events in Twitter, including the use of Non-negative Matrix Factorization (NMF) by Xintian et al. [162], a structured retrieval approach propsed by Metzler et al. [89], Structured Probabilistic Latent Semantic Analysis (PLSA) proposed by Lu et al. [82], and many more [52, 98, 33]. However, some of these algorithms can be applied only to periodic events such as sports events and not to longer term events or aperiodic events, while others do not perform particularly well on large real-world multilingual corpora. Therefore, we limit our comparison to the most recent leading summarizers, namely, Becker et al. [17], Zubiaga et al. [167], Xu et al. [160] and the Hybrid TF-IDF Summarizer [59].

5.3 Proposed Summarization Techniques

We propose three methods for summarizing a set of Twitter posts: Temporal TF-IDF, the Retweet Voting Approach and Temporal Centroid Representation. For all the proposed methods, we use a one-hour time window based on the best temporal settings, as was shown in the previous chapter particularly Section 4.6.3 as well as in [9]. The temporal TF-IDF is based on extracting the highest-weighted terms, as determined by the TF-IDF weighting in two successive time frames. The voting method considers the highest number of retweets that a post has received in a given time window as the criterion for the most representative post in this window. This method reflects users' choices, since they are the ones who determine which message is the most 'valuable' by propagating it. The temporal centroid method selects posts that correspond to each cluster centroid as the summary of the cluster with respect to the time dimension. Next,

we describe these methods and provide an analysis of the results.

5.3.1 TEMPORAL TF-IDF

The algorithm is inspired by the fact that users tend to use similar words when describing a particular event as well as observations obtained from [122]:

1. High frequency words, like stop-words, occur in approximately the same percentage of documents, no matter whether the document set is small or large and similarly, low frequency words such as "murder" occur very rarely across small and large datasets.
2. The document frequency distribution of one corpus can be used to approximate another.

We propose a novel temporal Term Frequency - Inverse Document Frequency (TF-IDF) that generates a summary of top terms without the need for prior knowledge of the entire dataset, unlike the existing TF-IDF approach [128] and its variants. Temporal TF-IDF is based on the assumption that words which occur more frequently across documents over a particular interval (timeframe) have a higher probability of being selected for human created multi-document summaries than words that occur less frequently [147].

Typically, the TF-IDF approach requires a knowledge of the frequency of a term in a document (TF) as well as the number of documents in which a term has occurred at least once (DF). The need for *a priori* knowledge of the entire data set introduces the significant challenge of using this approach where continuous data streams must be summarized in real time as an event unfolds. In addition, the adopted scheme must be flexible enough to update frequently (every minute, every 10 minutes, hourly, every 3 hours - depending on the time-frame size). Hence, the iterative calculation of term weights should be taken into account.

To overcome these limitations, we introduce the temporal TF-IDF where we consider a set of posts in a cluster to represent a document. The total number of clusters equals the total number of documents which is a subset of the entire dataset or corpus. This reduces the overall computational complexity and overcomes the limitations of the TF-IDF based approaches, in which the document set to be clustered must be known in advance. After the first cluster timeframe, we use the clusters from the previous timeframe with the documents in the recent one, to add more relevance and usefulness to our results, such as emerging keywords. Consequently, we use the document frequency distribution of two timeframes instead of one, taking into account the changing dynamic and narrative of the event. Therefore, a collection C consists of all Twitter posts from two timeframes. We define the TF-IDF weighting scheme of a new document d for a collection C (All Twitter posts from two timeframes) as:

$$w_{ji} = \frac{1}{norm(d_i)} f_{ji} \times log(1 + \frac{N}{N_j}) \tag{5.1}$$

where f_{ji} is the frequency of a word in document d_i, N_j is the document frequency of a word in a collection, w is weight of tweet, and N is the total number of documents in a collection. In order to avoid the bias caused by documents of different lengths, the length of each document vector is normalized so that it is of unit length $norm(d_i)$. This summarizer selects the most weighted post as the summary, as determined by the Temporal TF-IDF weighting.

5.3.2 Retweet Voting Approach

Many studies have illustrated the power of retweeting for many tasks such as predicting the most influential users [27], identifying the most knowledgeable posts [111], ranking and measuring information propagation [30] and analyzing network structure [71, 30]. Voting algorithms have been successfully implemented in many data mining applications [8]. Here we implement the highest number of retweets as a measure of the representation task through a voting algorithm. Voting algorithms have been used

in many applications where they may be considered in the context of the social media, taking into account the following features:

- The average length of a post.
- The total frequency of features in a post.
- The retweet count (the number of times a tweet in a cluster has been retweeted), or the favorite count, or the mention count.
- The inclusion of multimedia files, such as photos, videos.

Using the retweet count as the ranking method in a cluster has several advantages; first, it represents the influence of a tweet beyond one-to-one interaction [27]; second, retweeting serves as a powerful tool to reinforce a message when not only one but a group of users repeat the same message [111]; third, the number of retweets is an indication of popularity [27], so we are in a way summarizing the cluster by the highest degree of agreement from the users themselves. In addition we can generalize this method and apply it to other social networking sites, such as Facebook (number of Shares), Instagram (number of likes), Pinterest (number of Repins), etc. in one time-frame. We can also extend this approach to rank events/clusters by calculating the total number of retweets per cluster. However, using this method of representation also suffers from many drawbacks:

1. The content of a tweet is not always taken into consideration, many users retweet without even reading. For instance, most celebrities have a high number of retweets on account of their popularity.
2. A tweet with a high number of retweets may be repeated over time because it receives most attention and the Retweet Count generally increases with time. Thus, the Retweet Score is not a comprehensive measure.

Many techniques including classification and clustering have been used successfully to distinguish between messages about real-world events and non-event messages; hence, most messages from celebrities were removed in this thesis, unless they were pertinent. To overcome the second drawback, we introduced a normalization factor to calculate a Change of Retweet Score over time, replacing a Retweet Score.

A Retweet Score (rt) is defined as the ratio of the number of "retweets" that a tweet gets (u_i) to the total number of retweets (u_{all}) from all the posts in the target cluster. It is defined as

$$rt = \frac{|retweet(u_i)|}{|retweet(u_{all})|} \tag{5.2}$$

A Retweet Score Change is defined as the number of times that a tweet has been retweeted in the current time-frame (rt_{cur}) and is calculated by subtracting the number of retweet counts from the previous time-frame (rt_{pr}) of the same post.

$$rt\ change = rt_{cur} - rt_{pr} \tag{5.3}$$

5.3.3 Centroid Representation Method

The centroid approach takes account of the centrality measure of a tweet with respect to the overall topic of the cluster [17, 11]. It computes the cosine similarity of the TF-IDF representation of each message to its associated event cluster centroid, where each cluster term is associated with its average weight across all cluster messages. Then it selects the messages with the highest similarity value, because they represent the average weight of all terms in clusters. The main idea behind this method (since it is based on frequency across all messages) is to identify posts of high quality that are most relevant to an entire cluster. The difference between our proposed centroid method and other centroid methods is that we include the time dimension. We select the post which has been a centroid for the longest time on average over a time-window, rather than taking the final centroid at the end of this time-window. We believe that

studying the temporal aspects of posts reveal additional information about their quality, relevance, and usefulness.

5.4 Empirical Evaluation

The aim of these sets of experiments is to investigate and select the best summarization techniques. We perform three different sets of experiments: In the first set of experiments, we compare our proposed approaches to other recent leading summarizers, including Becker et al. [17], Zubiaga et al. [167], Xu et al. [160] and Hybrid TF-IDF Summarizer [59]. Furthermore, we compare the competing approaches according to users' perceptions of quality, relevance, and usefulness. In the third set of experiments, we evaluate and compare the performance of the summarization systems using different languages, namely; English, Arabic, and Japanese. Finally, we discuss the implications of our findings and experimental results.

5.4.1 Datasets and Setup

Dataset: We use the Twitter Streaming API to collect around 1.7 million tweets (1698517) posted from 15 October 2013 to 05 November 2013. This dataset was collected as part of our work on Twitter event identification (F1 Twitter Corpus). (See Section 2.4.1).

Annotations: We selected the top 10 event clusters per day, with an average of 320 posts per cluster, using the online clustering algorithm outlined in Chapter 4. For each event cluster we selected the top 5 posts according to our proposed approaches (whether by a Temporal TF-IDF, Retweet voting, or Temporal centroid method). In total 3000 tweets were manually annotated for the annotation task (for the summarization step), from a total of 200 event clusters. (Note that the selection of the top 10 event clusters per day based on the online clustering algorithm can introduce a bias toward

large-scale social media events, therefore evaluating our summarization approaches using a random selection of event clusters is reserved for future work).

We used three human annotators to label each post according to three desired goals, as reported by Becker et al. [17]:

1. Quality: refers to the textual quality of the messages, which reflects how well they can be understood by a human. High-quality messages contain crisp, clear, and effective text that is easy to understand.

2. Relevance: how well a Twitter message reflects the information related to its associated event. Highly relevant messages clearly describe their associated event.

3. Usefulness: the potential value of a post for someone who is interested in learning details about an event. Useful messages should provide some insight into the event, beyond simply stating that the event has occurred.

The annotators labeled each message on a scale of 1-4 for each attribute, letting a score of 4 signify high quality, strong relevance, and clear usefulness, and a score of 1 signify low quality, no relevance, and no usefulness. A set of instructions and examples was given to each annotator so that they could perform their task as well as the assessments had been done without reference to any model summaries. We used the CrowdFlower crowdsourcing system (http://www.crowdflower.com) to annotate the tweets. The level of agreement between annotators was substantial to high, with kappa coefficient values = 0:92; 0:89; 0:61 for quality, relevance, and usefulness, respectively. After the annotators became familiar with the topics and the summarization task, each annotator was asked to summarize each cluster in order to generate **gold standard** summaries. The annotators where only provided with a subset of the posts from event clusters (the top 5 posts which were selected by each summarization approach, a total of 15 posts per event cluster). The instructions that were provided to the annotations along with an example are shown in Appendix A.3.

Evaluation Methods: The similarity metric we used for evaluation and comparison between system summaries was the ROUGE metric proposed by Lin and Hovy [80]. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is an intrinsic metric for automatically evaluating summaries which is based on BLEU (Bi-Lingual Evaluation Understudy), an algorithm for evaluating machine translation by Lin [79]. The ROUGE metric counts the total number of matching n-grams (excluding stop-words) between the true summary and the summary generated from the model. Given a document D (tweets from the same cluster so they are describing the same topic), S donates a set of reference summaries, X donates an automatic summary (generated from one of the summarization systems), let i donate an N-gram, let $count(i, X)$ donate the frequency of i within X, and $count(i, S)$ donate the frequency of i within S, The ROUGE can be expressed as follows:

1. First, ask N humans to produce a set of reference summaries (S) of D document.
2. Then run the system or the model to generate an automatic summary X.
3. Calculate the percentage of the unigrams from the reference summaries (S) appearing in the automatic summary X. ROUGE-1 is computed as follows [79]:

$$ROUGE-1 = \frac{\sum_{S \in ReferemceSummaries} \sum_{unigrams \in S} min(count(i, X), count(i, S)}{\sum_{S \in ReferemceSummaries} \sum_{unigrams \in S} count(i, S)} \tag{5.4}$$

It has been shown that ROUGE scores correlate well with human judgments [80]. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE-3 (tigrams), and ROUGE-4 (quadrigrams). ROUGE-L is the Longest Common Subsequence (LCS) based statistics, ROUGE-SU is the skip-bigram plus unigram-based co-occurrence statistics, and ROUGE-W is the weighted LCS-based statistics that favors consecutive LCSs [79]. In this work, we use **ROUGE-1** scores as a fitness function for measuring summarization quality, because it showed the widest variation of all the methods

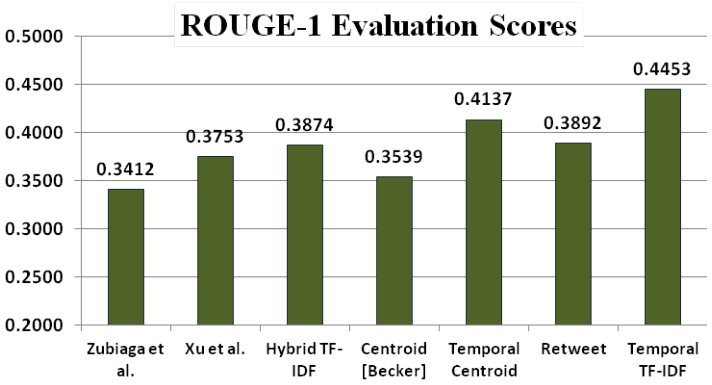


Figure 5.1: Results of our proposed approaches against other summarization techniques (The Y-axis shows the ROUGE-1 scores).

ROUGE-2, 3, 4, L, SU and W metrics [161, 28]. In evaluation, the ROUGE metric is easy and more convenient than human evaluation, but it is not so reliable. Hence, we have also evaluated the summarization techniques, using metrics, for quality, relevance and usefulness [17, 160].

5.4.2 Experimental Results

We conduct several experiments to evaluate different aspects of our summarization techniques. In the first experiment, we compare our proposed approaches to those of other recent leading summarizers, including Becker et al. [17], Zubiaga et al. [167], Xu et al. [160] and Hybrid TF-IDF Summarizer [59]. We selected these baselines because they have been shown to be effective for summarizing tweets and also represent different methods as described in section 5.2. We evaluate the other summarizers using the automatic ROUGE-1 evaluation. The values of the ROUGE-1 scores are presented in Figure 5.1.

Our approaches performed well when compared to other summarization methods. The Temporal TF-IDF adds more knowledge when determining both TF and IDF in two timeframes. Our Centroid algorithm performed better than the other approaches due to its inherent assumption that each cluster revolves around one central topic. In addition,

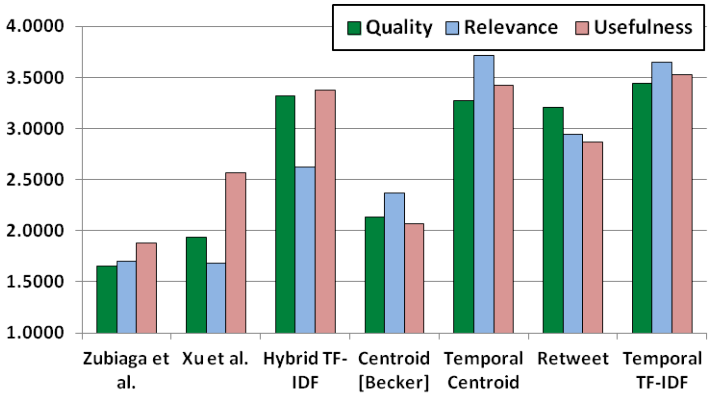


Figure 5.2: Comparison of content selection techniques.

the Retweet approach produces more satisfactory results than those in the summaries that used a baseline approach. Note that the ROUGE scores are based solely on the n-gram overlap between the system and reference summaries, which may not be the most appropriate measure for evaluating event summaries. Hence, further experiments are needed to investigate the proposed methods using more sophisticated evaluation measures.

The second experiment compares the competing approaches according to users' perceptions of quality, relevance, and usefulness. Figure 5.2 summarizes the average performance of these approaches across all 50 test events.

All three of our proposed approaches received high scores for quality the Temporal TF-IDF producing the highest score. In other words, our approaches are able to select clear, informative summary according to human judgements. The Temporal TF-IDF technique also receives a high score for usefulness, indicating that its selected messages are useful with respect to the associated events. The Temporal TF-IDF takes two timeframes into consideration, providing more details about an event than other methods can. The Temporal centroid and the Temporal TF-IDF, on average, select messages that are either somewhat relevant or highly relevant, which indicates that the Retweet voting approach is affected negatively toward the most influential users, for instance celebrities.

One noted problem with the method of Xu et al. is that this graph-based summarization is sensitive to document length, because its similarity estimation, especially in short tweets, mainly depends on commonly used words and not words with high IDF scores' unfortunately this increases the chances of selecting pointless information [161]. We also find that graph models tend to assign high salience scores to long tweets containing several #hashtags. Some of these tweets are pointless (low quality) or unrelated (of minimum relevance) to the topic.

Overall, it seems from the first two experiments that the simple frequency-based summarizers, namely Temporal TF-IDF and Hybrid TF-IDF, performed better than summarizers that incorporate more information or more complexity, using for example graph-based methods or centroid-based approaches. This possibly has much to do with the special nature of Twitter documents; they often have very little structure and are composed of so few words that trying to find relationships between pairs of documents is not particularly helpful. Therefore, more complex relational models will probably not capture more topical information than frequency models do and the added complexity of interrelationships does not help in summarizing Twitter posts. Moreover, Temporal TF-IDF outperforms Hybrid TF-IDF because our Temporal TF-IDF is more sensitive to changes over time, as can be clearly seen in the measures of both quality and relevance.

In comparison to other methods, Zubiaga et al. and similar systems such as [138, 161] are limited to scheduled events such as soccer games. They require the starting time before the system can start looking for new sub-events. The sub-event detection step (based on detecting peaks) fails to detect important events/topics and this reduces the chances of selecting valuable information (high quality, more relevant and very useful). This may explain why the Zubiaga et al. approach performance is lower than the results reported in this paper. Similarly, the centroid-based summarizers such as Becker et al. attempt to reduce redundancy but do so by clustering the documents first and then summarizing on the basis of these clusters. However, their clustering approach does

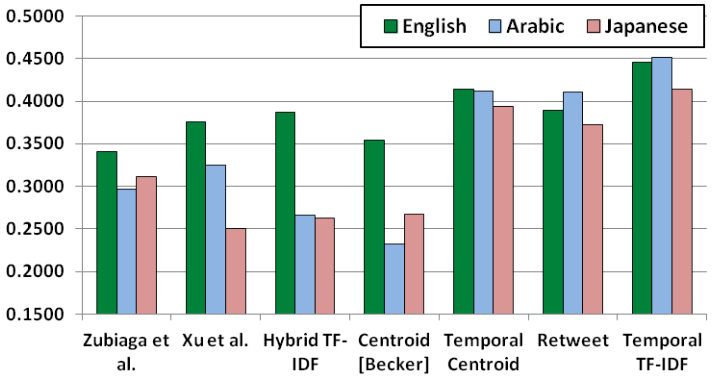


Figure 5.3: ROUGE-1 results of various summarization techniques for different languages. The Y-axis shows the ROUGE-1 scores.

not seem to improve performance particularly on small-sized clusters, which contain very few tweets in each cluster. This can be clearly seen by the low quality and the less significant usefulness measures in Figure 5.2.

For the third experiment, we generated subsets of our dataset to evaluate and compare the performance of the summarization systems using different languages. We randomly created three smaller subsets of English, Arabic and Japanese posts (number of posts: 200, 500 and 40, respectively). We deliberately chose English, Arabic and Japanese because they belong to distinct language families (the Indo-European, Semitic and Altaic languages, respectively). The results of the average ROUGE-1 values obtained for the English, Arabic and Japanese corpora are shown in Figure 5.3.

The results in Figure 5.3 confirm the findings from the first two experiments and in fact are consistent across all the languages considered. The results in Figure 5.3 show that our proposed approaches outperform other summarizers for morphologically-rich languages such as Arabic and Japanese. For Japanese and Arabic, the performance of our Temporal TF-IDF and Temporal centroid supports the claim that these methods handle well the variety of morphological phenomena present in these languages. The simple Retweet voting method achieves good results across all languages, strengthening our claim that users' choices are reliable, because they decide which is the message that

represents an event (cluster) best. The method can be used with no additional knowledge of the language of a message for the task of summarizing the micro-blog. Finally we combined the three summarization techniques in a visualization to generate meaningful real-time updates in order to facilitate the exploration of ongoing events for the end-users and decision makers and improve their understanding. We present the visualization tool in the Appendix B.

5.5 Summary

The rate of information growth due to the social media content and the real-time requirement of many tools have shown the need to develop efficient summarization techniques. Here we implemented three summarization techniques; the Temporal TF-IDF, the Retweet voting approach and the Temporal centroid method. Based on the results reported in this chapter, the temporal frequency based method achieved the best results both in ROUGE scores and in human evaluation scores. The centroid representation also reflects the topic/event; hence, the centroid representation can claim to have performed well. Not far behind them, the user's choice (the retweet voting algorithm) achieved good results too, which puts it among the best techniques for summarizing Twitter topics. Our evaluation also shows that our proposed methods perform well across a variety of language families, and we present here results that improve on the state-of-the-art at present for several noisy real-world datasets, including a multilingual corpus.

Chapter 6

Conclusions and Future Work

Social networking sites such as Twitter, Facebook and YouTube publish high volumes of user generated content as events occur, making them a potential data source, valuable for event analysis that could be transformed into actionable knowledge. The increased use of these fast-growing social media services has created vast amounts of content generated by people from different countries and diverse backgrounds. Messages posted on these platforms have been reporting everything from daily life stories to the latest local and global news and events. Moreover, social media platforms, particularly Twitter, offer a rich source of real-time information during mass convergence and emergency events such as disasters.

The main aim of this research was to set out a general classification-clustering framework for the purposes of detecting real-world events, both large and small, in real-time. The event detection task was performed in several stages: data collection (Chapter 2), preprocessing (Chapter 2), classification (Chapter 3), clustering (Chapter 4) and summarization (Chapter 5). Event detection aims at finding real-world occurrences that unfold over space and time using the textual content of user generated posts. Furthermore, we have implemented other features as well such as temporal and spatial features to add extra knowledge of the time and the location of an event. Using all of these features (temporal, spatial and textual features) and through several experiments, we were able to discriminatively distinguish between events, particularly disruptive events.

The results of our experiments demonstrated that it is not adequate to consider tem-

poral, spatial, or content-based aspects in isolation. Rather, a combination of features which covers all three aspects leads to a robust system that encourages the best event detection results. Extensive experiments were conducted to evaluate the effectiveness of the proposed framework using large real-world datasets. Our experiments suggested that our framework yields better performance than many leading approaches in real-time event detection.

This thesis presented original work to integrating several data mining techniques for detecting events for a specific time period and place, such as data pre-processing, supervised machine learning, topic clustering, and event summarization. This thesis makes four main contributions. First, we proposed a framework that integrates the classification-based approach with unsupervised clustering algorithms to discover events of different types, specifically small-scale disruptive events. Second, we presented an extensive analysis of various features related directly to social media data and show how they can be used to distinguish between disruptive events and other events.

Then we validated our model on several large-scale data sets from Twitter and Flickr to show the effectiveness of the framework. Our evaluation included a comparison with some leading event detection approaches which revealed that our approach outperformed existing state of the art systems. Finally, we presented and analyzed three methods that produced summaries automatically by selecting the most representative posts from real-world event clusters. This final chapter first summarizes the work in this thesis (in section 6.1) and concludes with recommendations to extend the research and suggestions for future work (in section 6.2).

6.1 Conclusions

In this thesis, we presented event identification, characterization, feature selection techniques, and event summarization, each of which serves as an integral part of applications that interact with real-time events and their associated documents, on social

networking sites. Specifically, we outlined several challenges raised by both the characteristics of the social media and those of events when identifying event content in real-time scenarios. We also outlined some of the promising opportunities for exploring and analyzing events from different social media sites (Chapter 1). We also presented our hypothesis, contribution, and motivation for this research in Chapter 1.

In Chapter 2 we discussed various efforts to define events in the context of the social media and defined different types of event, such as global and local, known and unknown, and periodic as opposed to aperiodic. Our proposed framework can detect all these types of events, however it does not differentiate between these events' types. Our framework, in fact, is not developed to identify the type of an event. In the same chapter, we reviewed existing solutions, models, systems, approaches, and tools in the literature and addressed their limitations. The first two steps of the framework (data collection and text processing techniques) were also described in Chapter 2, along with various datasets that were used in the thesis.

We began Chapter 3 by presenting our developed five-step framework which includes data collection, preprocessing, classification, on-line clustering and summarization. Then we detailed three machine learning algorithms, namely; the Naive Bayes classification, Logistic Regression, and support vector machines (SVMs) which we implemented to automatically classify messages related to real-world events and non-event posts. We found that the naive Bayes classifier outperformed other machine learning classification techniques. Furthermore, we investigated methods to improve the performance of the classification result; thus we considered different features which capture patterns in the data such as n-gram presence or n-gram frequency, the use of unigrams, bigrams and trigrams, linguistic features such as parts-of-speech (POS) tagging and Named Entity Recognition (NER). The experimental results showed that a combination of all the successful features (Unigrams + Bigrams+ POS + NER) gives the best classification results of 85.43% of all event-related messages.

In Chapter 4, we first proposed an online clustering framework to identify large and

related small scale events from the social media content. Then we presented an in-depth comparison of the three types of feature (temporal, spatial and textual) that could be useful for enhancing the event detection task. As we showed in Chapter 4, temporal features are the best event identifiers and hence they should not be disregarded. Moreover, textual features can be used to improve the overall performance of event detection. Importantly, the experimental results showed that a combination of optimum textual features with temporal and spatial features leads to the best event identification performance. We evaluated the framework by testing the results of its implementation and compared them with some leading event detection approaches using multiple large datasets of millions of social media messages written in many languages. The results shown in Chapter 4 confirm that our system can perform as well as terrestrial sources at detecting disruptive events from social media sites.

Chapter 5, finally, has detailed the task of automatically summarizing real-world events from microblogging sites, such as Twitter, and has presented three automatic summarization methods that work by selecting the most representative posts from event clusters. Our methods use term frequency, voting, and post centrality to select messages that represent an event with high quality, strong relevance and that are useful to people seeking information about the event. To evaluate our approaches, we compared them to the state-of-the-art summarization systems and human generated summaries. Our results showed that our proposed techniques for selecting the top documents for each event outperform other summarization systems for English and non-English corpora.

6.2 Future Work

This thesis proposed an end-to-end approach for identifying real-world event content on social networking sites such as Twitter. Inspired by previous theoretical work that bridges the social sciences, linguistics, and computer science, we used factors in three

broad categories that could affect a user's Twitter engagement with real-world events. These are: (i) temporal information, (ii) the content of tweets (including expert dictionary lexicons), and (iii) geolocation (the tweeter's geographical proximity to the event). During the course of this thesis, we have shown how our classification-clustering approach is able to distinguish between daily events and disruptive events for a certain location in a particular time window.

This framework can be generalized as a situational awareness system for the purpose of enriching decision making which can be implemented in many fields such as crisis management, information intelligence, or even daily police work. Our results support the claim that the use of social media for the purposes of information gathering could be used as a complement to traditional intelligence and is not to be used independently. There are various ways in which the work in this thesis could be taken further in extended work and projects.

Chapter 1 & 2: Future work could take many directions. One of the directions of our research is to explore and investigate other Named Entity Recognition tools to improve recognizing entities in social media content. In addition, we can investigate the usefulness of advanced NLP techniques such as morphological analysis, dependency parsing, etc. Another directions of our research is the use of word embeddings to resolve ambiguity in the meaning of words or phrases. In particular, the idea of using semantic features such as word embeddings (the vector representation of terms) is very attractive, especially in that we are dealing with short text which are not necessarily parseable and contain many slang expressions. We can also implement word2vec word embeddings and/or the GloVe algorithm. We recognize the comparison between word2vec word embeddings and the distributional semantics approaches, as well as other traditional approaches which leverage word embeddings of different dimensionalities. We will certainly consider using the semantic features in the future study.

Clustering and multiple events: Our model includes the assumption that every single instance belongs to a single cluster at any point of time and does not assume multi-

output clustering, where a post may be associated with multiple clusters. Furthermore, we assume that two or more events (such as protests or fire incidents) do not occur simultaneously at the same location in approximately the same time-window. Although this assumption is reasonable for such cases, it might not hold for other, bigger events such as political campaigns or concerts or conferences. To realize multiple event detection, we must consider and construct advanced probabilistic models that will allow hypotheses of multiple event occurrences [127].

Feature selection and analysis :

Temporal features: Temporal features in social media sites are significant and revealing source of information for detecting events, particularly trending events. In our temporal analysis section, we found that events in general exhibit temporal patterns. But, we assumed that events contained in a tweet occurred on the day that the tweet is posted; however, tweets may describe events that occurred days, weeks, or even years ago using linguistic expressions (e.g., 2 days ago A storm ...). This information needs to be taken into account when identifying real incidents. Note that temporal metadata is mostly used for detecting changes in the frequency tweets are created or words are used [132]. Some approaches also include temporal metadata in the clustering process [76]. Furthermore, [132] makes use of automatic named entity and temporal expression recognition, which are tokens or phrases in a text that serve to identify time intervals such as "yesterday". We plan to explore these techniques further in future work. The identification of temporal expressions and temporal boundaries and ranking them is another interesting task. Ranking documents according to temporal information is useful not only in clustering but also in event summarization.

Spatial features: As we mentioned, many studies have concluded that a person's geographical location (geolocation) significantly affects her/his social connections and activities in the offline world. Many researchers have also found evidence to show that offline geography significantly affects user interactions on social media. However, we found that the spatial features are weak event indicators. One reason may be

that the methods used in this research to automatically obtain the geolocations of real-world events need to be improved. Hence, more sophisticated and novel geolocation approaches should be proposed to extract geographical locations from social media streams.

A promising direction of future work related to event location would be to identify people on the "ground location" e.g. people on the ground near a disaster or a crisis who can report incidents directly and can be treated as primary and reliable sources of information. The geographical proximity between a user's location and the event's location may, of course, provide additional predictive power with respect to different event categories (topics).

Network features and visual features: Another direction is to consider more features in the clustering stage, such as network features (community detection) and visual features. Network features and detecting communities will allow us to discover groups of interacting nodes and the relations between them, which may lead to a better understanding of the real-time events, due to the high correlation between the influence of a social network (e.g., network size and social ties) and user interactions with civic events, either directly or indirectly [55]. This will not only enable people to detect events and their characteristics but should also help to identify communities with properties or attributes in common. Another direction for future work would be to learn the visual features from images and videos, which would be useful to the real-time event detection task in data streams.

Sentiment analysis: Regarding future work about sentiment analysis, we will attempt to improve the performance of the system through linguistic processing, despite the poor grammar of the short informal text messages analyzed. A promising future approach is the incorporation of context about the reasons why sentiment is used, such as differentiating between intention, arguments, and speculation and/or questions, idioms, and sarcasms [145].

Summarization: There are many interesting directions for future work on the sum-

marization task. Although our summarization can effectively process and summarize several languages, each summary from each cluster is in the same language as the input cluster. Therefore, one of the main directions is to produce multi-sentence or multi-post summaries or even to go further and form a coherent multi-sentence summary. This multi-document summarization can be composed from different languages aimed at generating high-quality multilingual summaries. Another direction is to extend our investigation of the multilingual summarization and conduct more experiments on larger data sets as well as adding a range of other languages such as German, French, and Russian. We will study the effect of these different languages in greater detail in the near future and we will also investigate language models so as to construct abstractive summarization models. In addition, we selected the top 10 event clusters per day when evaluating various summarization techniques, this could mean that we only focus on the main events per day. In the future, we will compare our summarization techniques against a technique that selects clusters randomly (Random).

Spam and spammers identification: In Chapter 1, we mentioned the spam and spammers in the sense that they introduce a great challenge for researchers in the social media domain. These spam/social spammers are filtered in three stages of our framework (pre-processing, classification and clustering). The short tweets that include spam are filtered in the pre-processing step, as discussed in the paper. Posts that are very short are removed because they are less likely to contain useful information. Most of the spam messages are removed in the classification step as non-events. In the online clustering step the remaining spam will be clustered as very small clusters and will not affect relatively large clusters - hence they will be filtered out. We plan to study and refine successful techniques for identifying social spam and spammers who target social media systems.

Rumors detection: In Chapter 4, we investigated various textual features, in addition to temporal and spatial features to guarantee the detection of small-scale events. Shown in Chapter 4 (when we discussed the Near-Duplicate measure)are some exemplary

cases of how textual content can be applied to effectively increase the confidence level of an event; for example, if several users independently tweet about an event, this would effectively increase the confidence level.

Another example is the co-occurrence of URLs in a cluster or sharing links from news websites, which would confirm that these tweets refer to the same event and improve the level of confidence of reports about it. Moreover, the higher the volume of tweets from nearby coordinates, the higher the level of confidence in the location of the event will be. In addition, one of the retweets applications is estimating rumors in social media by analyzing the retweet path of rumor-tweets. All of these features may facilitate the detection of rumors in the social media. Another method that we might consider is to analyze the distinctive characteristics of rumors/false information and the way in which they propagate in the microblogging communities.

Visualization: Using the visualization tools, we were able to visualize complex relationships around real-time events and analyze the interaction patterns in each cluster over a stream of data. We present an interactive tool for visualization of different summarization systems. The results of the visualization tool are shown in the Appendix. One of the main drawbacks of our visualization is that our online clustering algorithm clearly favors upward trends in the topic stream. Additionally, although the current results include the starting time for each event, we want to improve the event detection algorithm to include downward trends in the topic stream, so that the event cycle is complete. In addition, we aim to build the timeline of events with the complete context of these events and their future repercussions as well as the connection between some of them. Our current event detection framework does not exploit the relationships between users and treats each event independently, which is clearly reflected in our visualization output. Specifically, the question deserves further study to see how analyzing the relationships between users can contribute to event detection.

Interface system (dashboard): In the near future, we plan to develop a simple user-friendly interface requiring a minimum of user expertise. The interface will be de-

signed to facilitate the visualization and the analysis of events particularly disruptive events in real time. The end users (police or other decision makers), who will interact directly with the dashboard, will give only the location (ex. city name or neighbourhood name) as well as the time period. If many users agreed on a tweet that is classified as an "event" or clustered as a "disruptive event" and would like to find similar events, the system will add that tweet to the *dynamic* annotation corpus.

In the future and to ensure that we do not classify tweets related to event as non-event, we will use a corpus that has both a *static* and a *dynamic* component. The *static* corpus is made up of a large collection of annotated tweets labeled as events and another large collection of tweets that are non-events. In addition to the static corpus, we will build a *dynamic* corpus of tweets, labeled as events or non-events, which is periodically obtained from highly end users' agreement on tweets. Currently, we use only the static corpus which contains labelled tweets. The idea here is to include the dynamic corpus which will ensure identifying tweets about future events. Similarly, we will build a static and a dynamic corpora of disruptive events in order to identify disruptive events tweets on topics that we have not encountered previously or that we may encounter in the future.

We accept the limitations of our system and will explore improvements in the near future. One limitation is in the data collection and annotation for both classification and clustering and their biases toward event-related posts. Furthermore, validating our results against real-time official reports or from news stream is not feasible at this point, because we have yet to create a dataset of events from the traditional media combined with official reports about, for instance, disruptive events. Even when we create such a dataset, the performance of our model will be lower, for many reasons; first, not all events reported on traditional platforms are reported in the social media and vice versa. Secondly, the presence and the fast propagation of rumours and false information in the online social networks. Last, we undoubtedly accept the limitations of our framework; while it is capable of capturing events (such as disruptive events) with few posts, it still

cannot identify events with too few messages.

The proposed framework in its current form can discover events only from one social media platform at a time. The integration between various platforms for detecting incidents and identifying events across social networking sites will be an interesting concept for future investigation. Moreover, we aim to reason events more with respect to the various types of event. For instance, sub-event detection and details search can be improved, but decision makers may additionally require further information about an event, such as the number of injured people, or affected buildings, which are compiled in the aggregated event clusters.

To conclude, the research in this thesis has made significant advances in web mining and content analysis methods and technology, particularly with respect to knowledge discovery and data mining in the social media. This thesis proposed the use of user-generated content as a rich source of information to identify a wide variety of events, including disruptive events. In particular, we provided important insights regarding the types of event that are reported in the social media and the characteristics of their associated content. We have proposed and illustrated methods and techniques for developing algorithms that can adaptively learn from data streams in real time. During the course of this thesis, we developed key methods for identifying events and their associated social media documents in four main components; event classification, online clustering algorithms, feature selection and event summarization. We have drawn insights from a broad range of experiments examining each framework component and reached some conclusions about the potential and scope for enhancements that can be brought about by the leverage of user-generated content for real-time event identification.

Appendix

Appendix A: Data Annotation Task

In this appendix, we present some additional example tweets and annotations that have been used in Chapters: 3, 4 and 5 in the classification, clustering and summarization tasks, respectively.

A.1: Data Annotation for the Classification Task

In Chapter 3, we provided a detailed description of the training and testing of the classifiers. We asked three human annotators to manually label 5000 randomly selected tweets in two classes, "Event" and "Non-Event". A set of instructions and examples was given to the annotators so that they can perform the annotation task. The annotation instructions along with some of the example tweets and annotations were shown in Chapter 3. This appendix provides some more example tweets and annotations for the classification task (Classes are: Event or Non-Event), which are shown in Table 6.1.

Table 6.1: List of example tweets and annotations that were provided to the annotators for the classification task (Classes are: Event or Non-Event).

Tweet	Event or Non-Event
#Breaking: At least 13 people have been killed and 25 wounded in a bombing of a church north of Cairo, an Egyptian official said.	Event
#Today: Two people have died in a horror crash in Melbourne's east. @andrew_lund is LIVE in #9News	Event
#TrafficUpdate : An #accident has been reported on SMBZ Rd After #GlobalVillage towards Sharjah, please be extra cautious	Event
.@RTA_Dubai holds 4th #Dubai International Project Management Forum in November 2013 http://tinyurl.com/n434h3p	Event
P&O Ports wins USD $ 336m 30-year concession for port of Bosasso in Puntland http://tinyurl.com/n6h22tx	Event
Smart #Dubai launches Dubai Careers, future generation of digital recruitment platforms	Event
. @DubaiCulture's 'Reading Box' initiative attracted high number of school students from the city http://tinyurl.com/l4fp347 #Dubai	Event
Dubai Land Department: AED 77 billion of #Dubai real estate transactions during the first quarter of 2014	Event
Whaaaat a game That performance gives additional self confidence for the upcoming tough games #FCBayern #FCBBVB #MiaSanMia #jb17	Event

Tweet	Event or Non-Event
Tourist dead and 20 hurt after hot air balloon horror crash at holiday spot http://bit.ly/2oV41HK	Event
Mohamed bin Zayed holds talks with Bahrain Crown Prince on ways to enhance fraternal ties, GCC coordination efforts, regional developments	Event
Real Madrid 1-1 Atletico Madrid FT:Real Madrid go 3 points clear of Barcelona in #LaLiga with 72 points total this season.	Event
Happy Birthday to my brother & one of my favorite collaborators EVER, @Pharrell. Enjoy your day, P!!! https://www.instagram.com/p/BShfSi0hA8O/	Non-Event
spring break has been amazingggg I don't want it to go back to school	Non-Event
Cute gift with purchase. I have no idea what to do with it. #idontcook	Non-Event
The root of all health is in the brain. The trunk of it is in emotion. The branches and leaves are the body. #ZenMoment #HealthyLiving	Non-Event
The good life is one inspired by love and guided by knowledge. - Bertrand Russell	Non-Event
Beach time tick! Nat Trust garden romp in the sun with kids next. +icecream. Let the hols Big Relax begin #teacher5aday	Non-Event
Wish you were here! It's a glorious Saturday morning in Dubai, time for a stroll before the gallery opens at 10am.	Non-Event

Tweet	Event or Non-Event
good morning im on my way home and my mum is singing along to the titanic song	Non-Event
don't ever be afraid to dream to big nothings impossible if you believe in yourself you can achieve it	Non-Event
Spending time doing things with people you care about is all that matters in life.	Non-Event
I've just realised I haven't taken a selfie in AGES. . so here's one I took earlier	Non-Event

A.2: Data Annotation for the Clustering Task

In Chapter 4, we provided a detailed description of the the online clustering algorithm as well as the feature selection method. In order to evaluate the clustering algorithm, we employed human annotators to manually label 1600 clusters. The task of the annotators was to choose one category from the following eight categories: politics, finance, sport, entertainment, technology, culture, disruptive event and other-event. The other-event category represents all other events which are not related to the above categories. To ease the annotation process, a set of instructions and examples was given to the annotators so that they can perform the annotation task. The annotation instructions along with some of the example tweets and annotations were shown in Chapter 4. This appendix provides some more example tweets and annotations for the clustering task, which are shown in Table 6.2.

Table 6.2: List of example tweets and annotations that were provided to the annotators for the Clustering task (Categories are: Politics, Finance, Entertainment, Sport, Technology, Culture, Disruptive Event and Other-Event).

Tweet	Categories
#Today: Two people have died in a horror crash, in Melbourne's east. @andrew_lund is LIVE in #9News #AD	Disruptive Event
#TrafficUpdate : An #accident has been reported on SMBZ Rd After #GlobalVillage towards Sharjah please be extra cautious	Disruptive Event
.@RTA_Dubai holds 4th #Dubai International Project Management Forum in November 2013 http://tinyurl.com/n434h3p	Finance
P&O Ports wins USD $ 336m 30-year concession for port of Bosasso in Puntland http://tinyurl.com/n6h22tx	Finance
Smart #Dubai launches Dubai Careers future generation of digital recruitment platforms	Technology
. @DubaiCulture's Reading Box initiative attracted high number of school students from the city http://tinyurl.com/l4fp347 #Dubai	Culture
Dubai Land Department:AED 77 billion of #Dubai real estate transactions during the first quarter of 2014	Finance
Whaaaat a game That performance gives additional self confidence for the upcoming tough games #FCBayern #FCBBVB #MiaSanMia #jb17 #UAE	Sport
Tourist dead and 20 hurt after hot air balloon horror crash at holiday spot http://bit.ly/2oV41HK	Disruptive Event

Tweet	Categories
Mohamed bin Zayed holds talks with Bahrain Crown Prince on ways to enhance fraternal ties, GCC coordination efforts, regional developments #AD #UAE #Bahrain #GCC	Politics
Real Madrid 1-1 Atletico Madrid FT: Real Madrid go 3 points clear of Barcelona in #LaLiga with 72 points total this season.	Sport
Adios Luis Suarez, I wish you well. You were a great addition to a great club. #LFC #YNWA #Liverpool	Sport
We're hosting a high-level int'l summit @CultureSummitAD, featuring world leaders addressing role of culture in today's time #InAbuDhabi	Culture
"Mars 2117" includes a major space sciences focus in our universities. We're building a space pioneering passion among our young people.	Technology
The new Samsung Galaxy launch tomorrow at #MWC Barcelona 23:00 AbuDhabi time.	Technology
#Mohamed_bin_Zayed receives Prime Minister of #Denmark http://wam.ae/en/details/1395302607961 #wamnews	Politics
Taiwan passes first law in Asia to ban the eating of cats and dogs #Animal #Dubai	Other-Event
Researchers discussed how to improve immunotherapy at #AACR could the common cold virus help? http://po.st/D6IstF #AbuDhabi	Other-Event
#dubaitoday German Foreign Minister calls for more humanitarian aid for Africa	Politics

Tweet	Categories
The Pulitzer Prize for Biography or Autobiography goes to Hisham Matar for The Return'. #AbuDhabi & #UAE	Other-Event
"The journey today is the start of an annual celebration embodies the values of diversity, tolerance and positive" Her Excellency Minister of happiness, #OhoodAlRoumi #worldhappinessday #HappinessJourney #goodtimes	Entertainment
World Health Day, celebrated on 7 April every year to mark the anniversary of the founding of WHO #World_Health_Day Please come along and visit us at Twam Hospital #AlAin	Entertainment

A.3: Data Annotation for the Summarization Task

In Chapter 5, we provided a detailed description of the three techniques that we developed for summarizing Twitter messages. We also described the annotation task which we have used in order to evaluate the summarization methods. In this appendix, we summarize the annotation task and then we provide the instructions that were provided to the annotations along with an example, which are shown in Table 6.3 and Table 6.4.

We selected the top 10 event clusters per day, with an average of 320 posts per cluster, using the online clustering algorithm outlined in Chapter 4. For each event cluster we selected the top 5 posts according to our proposed approach (whether by a Temporal TF-IDF, Retweet voting, or Temporal centroid method). We used three human annotators to label each post according to three desired goals; quality, relevance, and usefulness. The annotators labeled each message on a scale of 1-4 for each attribute, letting a score of 4 signify high quality, strong relevance, and clear usefulness, and a score of 1 signify low quality, no relevance, and no usefulness. We used the Crowd-

Flower crowdsourcing system (http://www.crowdflower.com) to annotate the tweets. The level of agreement between annotators was substantial to high, with kappa coefficient values = 0:92; 0:89; 0:61 for quality, relevance, and usefulness, respectively. After the annotators became familiar with the topics and the summarization task, each annotator was asked to summarize each cluster in order to generate gold standard summaries. The annotators where only provided with a subset of the posts from event clusters (the top 5 posts for each summarization approach, or a total of 15 posts per event cluster). Table 6.3 provides the set of instructions that was given to the annotators so that they can perform the annotation task and Table 6.4 shows an example tweet with a screenshot of the annotation task.

Table 6.3: The instructions provided to the annotators for the annotation task. Note: a topic consists of 15 tweets because we have selected the top 5 posts, for each event cluster, according to our proposed approaches (Temporal TF-IDF, Retweet voting, and Temporal centroid method). i.e. 5 posts per approach.

Instructions: Given a Twitter message and a topic (A topic is a set of 15 tweets about the same event, please read all 15 tweets before beginning the summarization task), Your task is to:
(A) First, assign each tweet with a score of 1-4 for each attribute (quality, relevance, and usefulness), letting a score of 4 signify high quality, strong relevance, and clear usefulness, and a score of 1 signify low quality, no relevance, and no usefulness.
(B) Second, summarize the topic using your own words within the maximal summary length constraint of 140 characters long.
Definitions:
Quality: refers to the textual quality of the messages, which reflects how well they can be understood by a human.
★ High-quality messages contain crisp and clear text that is easy to understand.
Relevance: how well a Twitter message reflects the information related to its associated event.
★ Highly relevant messages clearly describe their associated event.
Usefulness: the potential value of a post for someone who is interested in learning details about an event.
★ Useful messages should provide some insight into the event, beyond simply stating that the event has occurred.
Instructions:
(1) Spend at least five minutes on each topic, **please read all 15 tweets before beginning the summarization task.**
(2) Read the example tweets and the invalid scores 1-4 for each attribute (quality, relevance, and usefulness) before beginning assigning scores if this is the first time you are working on this annotation task.
(3) When summarizing a topic, please remain within the maximal summary length constraint of 140 characters long.

Table 6.4: An example tweet of the summarization annotation task

Tweet: **#TrafficUpdate : An #accident has been reported on SMBZ Rd After #GlobalVillagetowards Sharjah, please be extra cautious**

(A) Assign this tweet with a score of 1-4 for the following attributes:

Quality: (Low) ☐ 1 ☐ 2 ☐ 3 ☐ 4 (High)

Relevance: (Low) ☐ 1 ☐ 2 ☐ 3 ☐ 4 (High)

Usefulness: (Low) ☐ 1 ☐ 2 ☐ 3 ☐ 4 (High)

(B) Now, summarize the topic using your own words within the maximal summary length constraint of 140 characters long. Or write a tweet that describes the topic using your own words.

Your summary:

Appendix B: Visualization Tool (Case study)

Visualizing complex relationships around real-time events in social streams is important for acquiring insight and using it for decision making. Hence, we combine the three proposed methods and implement them in a visualization tool for detecting and summarizing events. Our goal with the visualization tool is to facilitate the discovery and increase the interpretability of Twitter summaries for decision makers. Therefore, this visualization tool can be used for identifying and exploring real-time events by decision makers (e.g., police, fire department, crisis management group) for a particular city or a country. We aim to support the exploration and identification of events in a particular location while also giving the administrators easier access for their searches and the ability to explore online communities. In this appendix, we implement this interface by means of the R tool (http://www.r-project.org) which is a free software environment for statistical computing and graphics. In particular, we use the package Gephi [149], an open source graph visualization manipulation software (available at

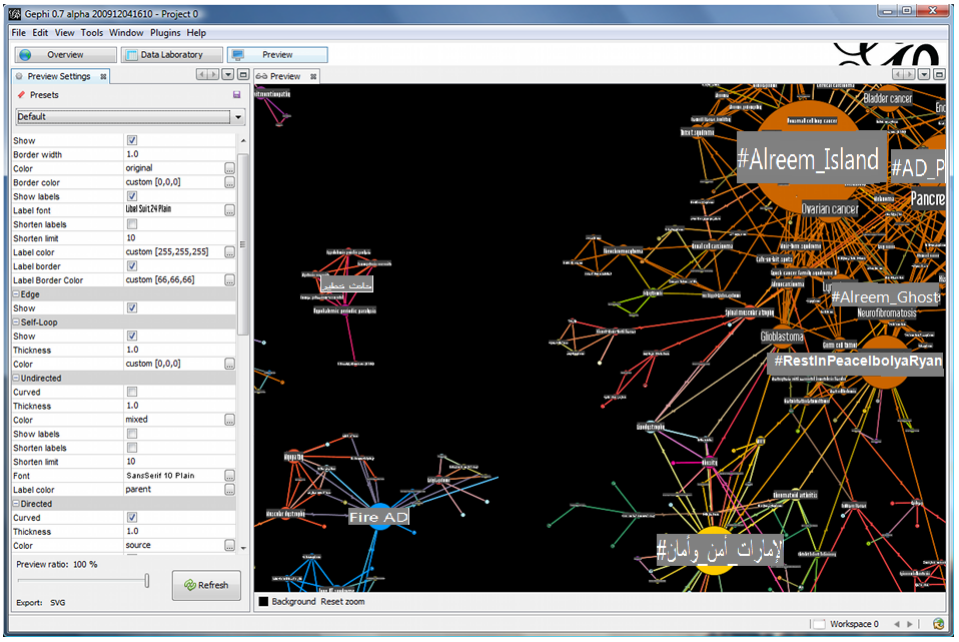


Figure 6.1: Results of combining summarization approaches after classification-clustering framework.

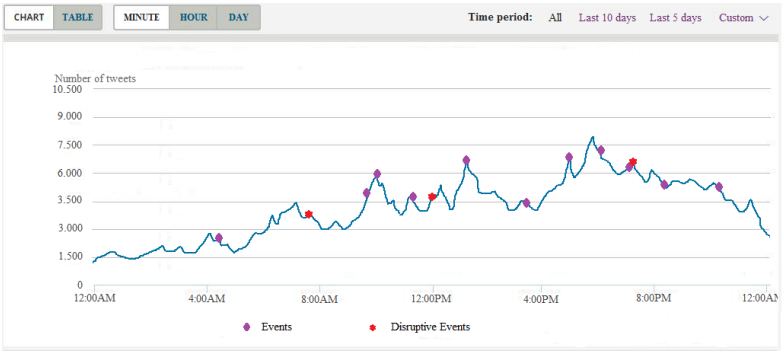
http://gephi.org/).

We use tweets for one day (3/12/2014) from our 2nd dataset 2.4.2 to visualize and identify the main events of the day, as shown in Figure 6.1. Figure 6.1 features a number of nodes, each representing a particular tweet. The color of each node represents the cluster that a tweet belongs to (Number of colors = Number of clusters (events)). The node size depends on two attributes: the TF-IDF value and the Change of Retweet Score - the larger the node, the higher the retweet count for this post. Lines between nodes specify communication and relationships between the exchanged messages; they also determine the centrality measures between messages and centroids. The visualization tool aims to visualize real-time events as a network of interactive events using Twitter data.

The visualization tool is also able to visualize event-related updates over time and space from tweets, giving a comprehensive view of events throughout a predefined period and an interpretation of these events. It supports the term-centric, temporal analytics

of event-related information in Twitter. To create the timeline we used the Annotated Time Line tool available as a Google Chart Tool (https://developers.google.com/chart/), as presented in figure 6.2.

In Figure 6.2, we present an example of the timeline of the number of tweets per hour (from the second dataset). Each of the peaks might be a candidate for an event, but, because we employ a classification-clustering framework, only event-related updates are detected; some of the peaks are actually related to events in general and others indicate *disruptive events* - events that threaten social safety and security, or could disrupt the social order. Identifying disruptive events from a social media stream is a useful source of information for improving situational awareness and decision support. The detected events and disruptive events are marked on the timeline, and are accompanied by a tag cloud description from one of several summarization systems.



Temporal TF-IDF	UAE News | @MohdAAhli محمد أهلي الداخلية تهيب بالجمهور الإتصال بشرطة دبي 8002626 للإبلاغ عن أية محاولة تخص مصرع سيدة امريكية داخل دورة مياه في احدى مولات جزيرة الريم 17 3
Retweet	MOI_UAE @AbuDhabiPolice Woman Dies after Public Toilet Fight on Reem Island fb.me/7fSmidVS9 30 12
Centroid	Gulf Eco News @Gulf_Eco_News Woman Dies after Public Toilet Fight on Reem Island mbcurl.me/W03A @AbuDhabiPolice

Figure 6.2: The timeline of identified events and disruptive events with examples of summaries using different summarization techniques.

Bibliography

[1] Fabian Abel, Claudia Hauf, Geert Houben, Richard Stronkman, and Ke Tao. Twitcident: Fighting fire with information from social web streams. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '14 Companion, pages 305–308. ACM, April 16-20 2012.

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[3] James Allan, Ron Papka, and Victor 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, pages 37–45, New York, NY, USA, 1998. ACM.

[4] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41, December 2007.

[5] Omar Alonso, Jannik Strötgen, Ricardo A. Baeza-Yates, and Michael Gertz. Temporal information retrieval: Challenges and opportunities. In *WWW2011 Workshop on Linked Data on the Web, Hyderabad, India, March 29, 2011*, pages 1–8, 2011.

[6] Nasser Alsaedi and Pete Burnap. Arabic event detection in social media. In *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14 _ 20 April*, CICLing '15, pages 384–401. Springer, 2015.

[7] Nasser Alsaedi and Pete Burnap. Feature extraction and analysis for identifying disruptive events from social media. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 1495–1502, New York, NY, USA, 2015. ACM.

[8] Nasser Alsaedi, Pete Burnap, and Omer Rana. A combined classification-clustering framework for identifying disruptive events. In *Proceedings of the 7th ASE International Conference on Social Computing, Stanford University, CA., USA*, SocialCom '14, 2014.

[9] Nasser Alsaedi, Pete Burnap, and Omer Rana. Identifying disruptive events from social media to enhance situational awareness. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 934–941, New York, NY, USA, 2015. ACM.

[10] Nasser Alsaedi, Pete Burnap, and Omer Rana. Can we predict a riot? disruptive event detection using twitter. *ACM Transactions on Internet Technology (TOIT)*, 17(2):18:1–18:26, March 2017.

[11] Nasser Alsaedi, Pete Burnap, and Omer F. Rana. Automatic summarization of real world events using twitter. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 511–514, 2016.

[12] Nasser Alsaedi, Pete Burnap, and Omer F. Rana. Sensing real-world events using arabic twitter posts. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 515–518, 2016.

[13] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Comput. Intell.*, 31(1):132–164, February 2015.

[14] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 13–22, New York, NY, USA, 2013. ACM.

[15] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[16] Hila Becker, Mor Naaman, and Luis 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.

[17] Hila Becker, Mor Naaman, and Luis Gravano. Selecting quality twitter content for events. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[18] Edward Benson, Aria Haghighi, and Regina Barzilay. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 389–398, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[19] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping Multidimensional Data - Recent Advances in Clustering*, pages 25–71. 2006.

[20] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Machine Learning Research*, 3:993–1022, March 2003.

[21] Alexander Boettcher and Dongman Lee. Eventradar: A real-time local event detection scheme using twitter stream. In *2012 IEEE International Conference on Green Computing and Communications (GreenCom)*, pages 358–367, 2012.

[22] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 7–13. Association for Computational Linguistics, 2013.

[23] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine.

[24] Pete Burnap, Omer F. Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. Detecting tension in online

communities with computational Twitter analysis. *Technological Forecasting and Social Change*, may 2013.

[25] Pete Burnap, Matthew L. Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining*, 4:206, 2014.

[26] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254, June 1996.

[27] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[28] Deepayan Chakrabarti and Kunal Punera. Event summarization using tweets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[29] Ling Chen and Abhishek Roy. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 523–532, New York, NY, USA, 2009. ACM.

[30] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, New York, NY, USA, 2014. ACM.

[31] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

[32] Soudip Roy Chowdhury, Muhammad Imran, Muhammad Rizwan Asghar, Sihem Amer-Yahia, and Carlos Castillo. Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management*, ISCRAM '10, 2013.

[33] Freddy Chong Tat Chua and Sitaram Asur. Automatic summarization of events from social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*, 2013.

[34] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[35] Mário Cordeiro. Twitter event detection: Combining wavelet analysis and topic inference summarization. In *Doctoral Symposium on Informatics Engineering, DSIE*, 2012.

[36] Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

[37] Kareem Darwish and Walid Magdy. Arabic information retrieval. *Foundations and Trends R in Information Retrieval*, 7(4):239–342, 2013.

[38] Dipanjan Das and Andre F.T. Martins. A survey on automatic text summarization single-document summarization. *Literature Survey for the Language and Statistics II course at CMU, Pittsburg (2007)*, pages 1–31, 2007.

[39] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM.

[40] Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. Unfolding the event landscape on twitter: Classification and exploration of user categories. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 241–244, New York, NY, USA, 2012. ACM.

[41] Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. Automatic tagging of arabic text: From raw text to base phrase chunks. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 149–152, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[42] Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard. Multiscale event detection in social media. *Data Min. Knowl. Discov.*, 29(5):1374–1405, September 2015.

[43] Marian Dörk, Daniel M. Gruen, Carey Williamson, and M. Sheelagh T. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, 2010.

[44] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[45] Jonathan L. Elsas and Susan T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 1–10, New York, NY, USA, 2010. ACM.

[46] Günes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479, 2004.

[47] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000, 1998.

[48] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Int. Res.*, 34(1):443–498, March 2009.

[49] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. In *Technical report, Stanford.*, 2009.

[50] Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 1045–1062, 2012.

[51] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[52] Sanda M. Harabagiu and Andrew Hickl. Relevance modeling for microblog summarization. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[53] Qi He, Kuiyu Chang, and Ee-Peng Lim. Analyzing feature trajectories for event detection. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 207–214, New York, NY, USA, 2007. ACM.

[54] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA, 2011. ACM.

[55] Yuheng Hu, Shelly Farnham, and Kartik Talamadupula. Predicting user engagement on twitter with real-world events. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*, ICWSM '15, 2015.

[56] Georgiana Ifrim, Bichen Shi, and Igor Brigadir. Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In *Proceedings of the SNOW 2014 Data Challenge co-located with 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014.*, pages 33–40, 2014.

[57] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, 47(4):67:1–67:38, June 2015.

[58] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 159–162. ACM, 2014.

[59] David Inouye and Jugal Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and*

2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011, pages 298–306, 2011.

[60] Akshaya Iyengar, Tim Finin, and Anupam Joshi. Content-based prediction of temporal boundaries for events in twitter. In *Proceedings of the 3rd IEEE International Conference on Social Computing*, pages 186–191, 2011.

[61] Anjali Ganesh Jivani. A comparative study of stemming algorithms. *IJCTA*, 2(6):1930–1938, 2011.

[62] Thorsten Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK, 1998. Springer-Verlag.

[63] Jeon Hyung Kang, Kristina Lerman, and Anon Plangprasopchok. Analyzing microblogs with affinity propagation. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 67–70, New York, NY, USA, 2010. ACM.

[64] Nattiya Kanhabua, Sara Romano, and Avaré Stewart. Identifying relevant temporal expressions for real-world events. In *SIGIR 2012 Workshop on Time-aware Information Access (TAIA'2012)*, 2012.

[65] Andrea Kavanaugh, Edward A. Fox, Steven Sheetz, Seungwon Yang, Lin Tzy Li, Travis Whalen, Donald Shoemaker, Paul Natsev, and Lexing Xie. Social media use by government: From the routine to the critical. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, dg.o '11, pages 121–130, New York, NY, USA, 2011. ACM.

[66] Houda Khrouf and Raphael Troncy. Eventmedia: a lod dataset of events illustrated with media. In *Semantic Web journal (Linked Dataset Descriptions)*, pages 1–6, 2012.

[67] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 91–101, New York, NY, USA, 2002. ACM.

[68] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press.

[69] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P. Gummadi. Geographic dissection of the twitter network. ICWSM '12, 2012.

[70] Shamanth Kumar, Fred Morstatter, and Huan Liu. *Twitter Data Analytics*. Springer Publishing Company, Incorporated, New York, 1st edition, 2013.

[71] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[72] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. *Mach. Learn.*, 59(1-2):161–205, May 2005.

[73] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 435–442, New York, NY, USA, 2010. ACM.

[74] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2nd edition, 2014.

[75] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 4–15, London, UK, UK, 1998. Springer-Verlag.

[76] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. TEDAS: A twitter-based event detection and analysis system. In *ICDE*, pages 1273–1276. IEEE Computer Society, 2012.

[77] Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma. A probabilistic model for retrospective news event detection. In *Proceedings of the 28th Annual Inter-*

national ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, pages 106–113, New York, NY, USA, 2005. ACM.

[78] Chen Lin, Chun Lin, Jingxuan Li, Dingding Wang, Yang Chen, and Tao Li. Generating event storylines from microblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 175–184, New York, NY, USA, 2012. ACM.

[79] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL workshop on text summarization branches out*, pages 74–81, 2004.

[80] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 71–78. Association for Computational Linguistics, 2003.

[81] Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. Automatically building a stopword list for an information retrieval system. *JDIM*, 3(1):3–8, 2005.

[82] Yue Lu, ChengXiang Zhai, and Neel Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 131–140, New York, NY, USA, 2009. ACM.

[83] Zongyang Ma, Aixin Sun, and Gao Cong. On predicting the popularity of newly emerging hashtags in twitter. *JASIST*, 64(7):1399–1410, 2013.

[84] Jim Maddock, Kate Starbird, Haneen J. Al-Hassani, Daniel E. Sandoval, Mania Orand, and Robert M. Mason. Characterizing online rumoring behavior using multi-dimensional signatures. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 228–241, New York, NY, USA, 2015. ACM.

[85] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, 2012.

[86] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, 2012.

[87] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, New York, NY, USA, 2011. ACM.

[88] Michael Mathioudakis and Nick Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[89] Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 646–655, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[90] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. *Proceedings of EMNLP*, pages 404–411, 2004.

[91] Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, March 2002.

[92] United kingdom: Metropolitan Police Service MPS. 4 days in august: Strategic review into the disorder of august 2011 - final report. `http://www.met.police.uk/foi/pdfs/priorities_and_how_we_are_doing/corporate/4_days_in_august.pdf`, 2012. Accessed: 2016-01-01.

[93] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy: Characterizing emerging trends on twitter. *J. Am. Soc. Inf. Sci. Technol.*, 62(5):902–918, May 2011.

[94] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. 2012.

[95] Jun-Ping Ng, Yan Chen, Min-Yen Kan, and Zhoujun Li. Exploiting timelines to enhance multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 923–933, 2014.

[96] Tu Ngoc Nguyen and Nattiya Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands*, pages 222–234, 2014.

[97] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pages 189–198, New York, NY, USA, 2012. ACM.

[98] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[99] Onook Oh, Manish Agrawal, and H. Raghav Rao. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers*, 13(1):33–43, mar 2011.

[100] Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings of the Advances in Digital Libraries Conference*, ADL '98, pages 12–, Washington, DC, USA, 1998. IEEE Computer Society.

[101] Andrei Olariu. Efficient online summarization of microblogging streams. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 236–240, 2014.

[102] Alexandra Olteanu, Carlos Castillo, Nicholas Diakopoulos, and Karl Aberer. Comparing events coverage in online news and social media: The case of cli-

mate change. In *Proceedings of the Ninth International Conference on Web and Social Media*, ICWSM '15, pages 288–297, 2015.

[103] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, ICWSM '14, 2014.

[104] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 994–1009, New York, NY, USA, 2015. ACM.

[105] Saša Petrović Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimptonand. Can twitter replace newswire for breaking news? In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM '13, 2013.

[106] Chi-Chun Pan and Prasenjit Mitra. Event detection with spatial latent dirichlet allocation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, pages 349–358, New York, NY, USA, 2011. ACM.

[107] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[108] Symeon Papadopoulos, Emmanouil Schinas, Vasileios Mezaris, RaphaÃ«l Troncy, and Ioannis Kompatsiaris. Social event detection at mediaeval 2012: Challenges, dataset and evaluation. In *In Proceedings of the MediaEval 2012 Workshop*, 2012.

[109] Symeon Papadopoulos, Emmanouil Schinas, Vasileios Mezaris, RaphaÃ«l Troncy, and Ioannis Kompatsiaris. Social event detection at mediaeval 2012: Challenges, dataset and evaluation. In *Proceedings of the MediaEval 2012 Workshop*, 2012.

[110] Ken Peffers, Tuure Tuunanen, Marcus Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *J. Manage. Inf. Syst.*, 24(3):45–77, December 2007.

[111] Sasa Petrović, Miles Osborne, and Victor Lavrenko. RT to win! predicting message propagation in twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[112] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[113] Swit Phuvipadawat and Tsuyoshi Murata. Breaking news detection and tracking in twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 120–123, 2010.

[114] M. F. Porter. Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[115] Martin F. Porter. Snowball: A language for stemming algorithms. Published online, 2001.

[116] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589–1599, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[117] Dragomir R. Radev, Sasha Blair-Goldensohn, and Zhu Zhang. Experiments in single and multidocument summarization using MEAD. *First Document Understanding Conference*, 2001.

[118] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: Computing word relatedness using temporal semantic ana-

lysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA, 2011. ACM.

[119] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA, 2011. ACM.

[120] Kira Radinsky and Eric Horvitz. Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 255–264, New York, NY, USA, 2013. ACM.

[121] Aniket Rangrej, Sayali Kulkarni, and Ashish V. Tendulkar. Comparative study of clustering techniques for short text documents. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 111–112, New York, NY, USA, 2011. ACM.

[122] Joel W. Reed, Yu Jiao, Thomas E. Potok, Brian A. Klump, Mark T. Elmore, and Ali R. Hurson. TF-ICF: A new term weighting scheme for clustering dynamic data streams. In *The Fifth International Conference on Machine Learning and Applications, ICMLA 2006, Orlando, Florida, USA, 14-16 December 2006*, pages 258–263, 2006.

[123] Timo Reuter, Philipp Cimiano, Lucas Drumond, Krisztian Buza, and Lars Schmidt-Thieme. Scalable event-based clustering of social media via record linkage techniques. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[124] Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. Weakly supervised extraction of computer security events from twitter. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 896–905, New York, NY, USA, 2015. ACM.

[125] Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *In Proceedings of the 1st International Workshop on Mining Social*, 2009.

[126] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pages 508–524, Berlin, Heidelberg, 2012. Springer-Verlag.

[127] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[128] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

[129] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. Twitterstand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, USA, 2009. ACM.

[130] Hassan Sayyadi and Louiqa Raschid. A graph analytical approach for topic detection. *ACM Transactions on Internet Technology (TOIT)*, 13(2):4:1–4:23, December 2013.

[131] Emmanouil Schinas, Georgios Petkos, Symeon Papadopoulos, and Y.Kompatsiaris. Certh @ mediaeval 2012 social event detection task. In *Proceedings of the MediaEval 2012 Workshop*, pages 6–7, 2012.

[132] Axel Schulz, Benedikt Schmidt, and Thorsten Strufe. Small-scale incident detection based on microposts. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, HT '15, pages 3–12, New York, NY, USA, 2015. ACM.

[133] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1177–1178, New York, NY, USA, 2010. ACM.

[134] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March 2002.

[135] Andrea Setzer. and University of Sheffield. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. University of Sheffield, 2001.

[136] David A. Shamma., Lyndon Kennedy, and Elizabeth F. Churchill. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events? 2010.

[137] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 685–688, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[138] Chao Shen, Fei Liu, Fuliang Weng, and Tao Li. A participant-based approach for event summarization using twitter streams. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Atlanta, Georgia, USA*, pages 1152–1162, 2013.

[139] Luke Sloan and Jeffrey Morgan. Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PLOS ONE*, 10(11):1–15, 11 2015.

[140] Irena Spasic, Pete Burnap, Mark Greenwood, and Michael Arribas-Ayllon. A Na\"{i}ve Bayes Approach to Classifying Topics in Suicide Notes. *Biomedical Informatics Insights*, 5(1):87–97, 2012.

[141] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 841–842, New York, NY, USA, 2010. ACM.

[142] Kate Starbird and Leysia Palen. "voluntweeters": Self-organizing by digital volunteers in times of crisis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1071–1080, New York, NY, USA, 2011. ACM.

[143] Kate Starbird and Leysia Palen. (how) will the revolution be retweeted?: Information diffusion and the 2011 egyptian uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 7–16, New York, NY, USA, 2012. ACM.

[144] Nicholas A. Thapen, Donal Stephen Simmie, and Chris Hankin. The early bird catches the term: Combining twitter and news data for event detection and situational awareness. *CoRR*, abs/1504.02335, 2015.

[145] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *J. Am. Soc. Inf. Sci. Technol.*, 62(2):406–418, February 2011.

[146] Oren Tsur, Adi Littman, and Ari Rappoport. Efficient clustering of short messages into general domains. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*, 2013.

[147] Lucy Vanderwende, Hisami Suzukia, Chris Brocketta, and Ani Nenkovab. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.*, 43(6):1606–1618, 2007.

[148] Konstantinos N Vavliakis, Andreas L Symeonidis, and Pericles A Mitkas. Event identification in web social media through named entity recognition and topic modeling. *Data and Knowledge Engineering*, 88:1–24, 2013.

[149] George G. Vega, orge Fabrega, and Joshua Kunst. rgexf: An r package to build gexf graph files. *In The Comprehensive R Archive Network*, 2012.

[150] Sarah Vieweg, Carlos Castillo, and Muhammad Imran. Integrating social media communications into the rapid assessment of sudden onset disasters. In *Proceedings of the 6th International Conference on Social informatics*, pages 444–461, 2014.

[151] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1079–1088, New York, NY, USA, 2010. ACM.

[152] Daniele Vitale, Paolo Ferragina, and Ugo Scaiella. Classification of short texts by deploying topical annotations. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pages 376–387, Berlin, Heidelberg, 2012. Springer-Verlag.

[153] Maximilian Walther and Michael Kaisser. Geo-spatial event detection in the twitter stream. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 356–367, Berlin, Heidelberg, 2013. Springer-Verlag.

[154] Xiaoyu Wang, Wenwen Dou, William Ribarsky, Drew Skau, and Michelle X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, pages 93–102, Washington, DC, USA, 2012. IEEE Computer Society.

[155] Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1301–1315, 2015.

[156] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2541–2544, New York, NY, USA, 2011. ACM.

[157] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, 2011.

[158] Matthew L. Williams and Pete Burnap. Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, pages 1–28, 2015.

[159] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[160] Wei Xu, Ralph Grishman, Adam Meyers, and Alan Ritter. A preliminary study of tweet summarization using information extraction. In *Workshop on Language in Social Media (LASM 2013), Conference of the Association of Computational Linguistics, Proceedings, June 13, 2013, Atlanta, Georgia, USA*, pages 20–29, 2013.

[161] Duan Yajuan, Chen Zhumin, Wei Furu, Zhou Ming, and Heung Y. Shum. Twitter topic summarization by ranking tweets using social influence and content quality. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 763–780, 2012.

[162] Xintian Yang, Amol Ghoting, Yiye Ruan, and Srinivasan Parthasarathy. A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 370–378. ACM, 2012.

[163] Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. Topic-conditioned novelty detection. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 688–693, New York, NY, USA, 2002. ACM.

[164] Jie Yin, Sarvnaz Karimi, Andrew Lampert, Mark A. Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness: Extended abstract. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI*, pages 4234–4239, 2015.

[165] Zicong Zhou, Roja Bandari, Joseph Kong, Hai Qian, and Vwani Roychowdhury. Information resonance on twitter: Watching iran. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 123–131, New York, NY, USA, 2010. ACM.

[166] Christos Zigkolis, Symeon Papadopoulos, George Filippou, Yiannis Kompatsiaris, and Athena Vakali. Collaborative event annotation in tagged photo collections. In *Multimedia Tools and Applications*, pages 1–30, 2012.

[167] Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. Towards real-time summarization of scheduled events from twitter streams. In *Proceed-*

ings of the 23rd ACM Conference on Hypertext and Social Media, HT '12, pages 319–320, New York, NY, USA, 2012. ACM.