International Journal of Computer Applications (0975 8887)
Volume 111 - No. 9, February 2015

# Classifying Short Text in Social Media: Twitter as Case Study

Faris Kateb
Computer Science Department
University of Colorado at Colorado Springs, USA and
King Abdulaziz University, Saudi Arabia

Jugal Kalita
Computer Science Department
University of Colorado at Colorado Springs, USA

## ABSTRACT

With the huge growth of social media, especially with 500 million Twitter messages being posted per day, analyzing these messages has caught intense interest of researchers. Topics of interest include micro-blog summarization, breaking news detection, opinion mining and discovering trending topics. In information extraction, researchers face challenges in applying data mining techniques due to the short length of tweets as opposed to normal text with longer length documents. Short messages lead to less accurate results. This has motivated investigation of efficient algorithms to overcome problems that arise due to the short and often informal text of tweets. Another challenge that researchers face is stream data, which refers to the huge and dynamic flow of text generated continuously from social media. In this paper, we discuss the possibility of implementing successful solutions that can be used to overcome the inconclusiveness of short texts. In addition, we discuss methods that overcome stream data problems.

## Keywords:

Social Media Mining, Short Text Classification, Stream Data

## 1. INTRODUCTION

By the term social media, we mean Internet-based applications that include methods for communication among their users. One of the fastest growing social media applications is Twitter[1]. Currently, Twitter is gaining 135,000 new users every day, with a total of 645,750,000 users in 2013[2]. Social networks have received attention of analysts and researchers because decision makers rely on statistics such as summaries of people's opinions that can be obtained from analysis of social media. We focus on Twitter as a case study in this paper because it has become a tool that can help decision makers in various domains connect with changing and disparate of consumers and other stakeholders at various levels. The reason is that Twitter posts reflect people's instantaneous opinions regarding an event or a product, and these opinions spread quickly [39].

As researchers, we concentrate on Twitter for three reasons. The first reason for choosing Twitter is its popularity. Enormous numbers of people constantly post on Twitter regarding many varied topics. Topics could be politics, sports, religion, marketing, people's opinions or friends' conversations. Being a constantly updated huge repository of facts, opinions, banters and other minutiae, Twitter has received a large amount of attention from business leaders, decision makers, and politicians. This attention comes from the desire to know people's views and opinions regarding specific topics [71]. The second reason for using Twitter is the structure of its data, which is easy for software developers to deal with. The data is structured in such a way that all information regarding a tweet is rolled into one block using the *Json* file format. A block consists of many fields regarding user information, tweet description and re-tweet status. This type of structure eases difficulties in mining for specific information such as tweet content while ignoring other details such as user or re-tweet status. Finally, Twitter provides data filtering based on features such as retrieving tweets in a specific language or from a certain location. This flexibility in retrieving data encourages developers to perform research and analysis using Twitter. However, there is a limit on the size of retrieved data within a certain period of time. To retrieve more than 5% of all tweets, developers need special permission from Twitter.

The discussion above provides good motivation for finding approaches to apply data mining techniques on messages or posts in social networks. Researchers and analysts face two main challenges when they work on extracting information from social media. The first challenge is due to the brevity of Twitter posts. The second challenge is the huge amount of data streamed from Twitter. Fast algorithms are crucial in some cases, such as opinion mining in stock markets, which need real-time analysis.

Hotho et al. present four major challenges in classification of text generated by social media. These challenges are short texts, abundant information, noisy phrases and time sensitivity [29]. Short texts contain very few words, leading to a lack of features for classification, negatively affecting results. Twitter produces around 500 million tweets per day [56] in a variety of languages, containing spam as well as personal conversations. This means that we have to filter the tweets for spam and noise before analysis. However, since the number of tweets is high, pre-processing and classification require extensive time to perform, which make it

[1]http://www.twitter.com
[2]http://www.statisticbrain.com

difficult to perform real-time classification.

In this paper, we focus on short text and stream data for two reasons. First, we do not need to separate issues due to time sensitivity and stream data because they can be treated as one challenge. If a fast algorithm can be developed, both issues are solved simultaneously. Second, the use of noisy words and phrases can be partially handled by normalization. Normalization has been addressed successfully by many methods using dictionaries and NLP techniques.

The rest of the paper is organized as follows. Before discussing the challenges, we define important terms in the field in Section 2. In Section 3, we present the motivation behind this work. Section 4 includes techniques that have been used for classifying text. In Section 5, we introduce short text classification challanges and we discuss the published solutions for this issue. Section 6 presents challenges of processing stream data and solutions. Section 7 has a brief discussion and thougths regarding social media text classification.

## 2. IMPORTANT TERMS

In this section, we provide short description of terms that will help understand the rest of the paper. We define the keywords that we think need to be clarified.

The term **Short Text** refers to a short message and is usually less than 200 characters long, such as mobile SMS, online chat records and some blog comments [20]. They identify three features in a short text message: sparsity, immediacy, and words with unrecognizable format. Sparsity refers to a very short text that contains few words and makes information extraction difficult. Immediacy refers to messages that are generated in real time.

**Twitter Posts (tweets)** are the messages that people contribute on Twitter. Each tweet includes detailed information regarding the user such as name, location, language, creation time and date stamp in addition to the content of the tweet. Each tweet has 140 characters, which may be considered a sentence or approximately 18 words [30]. Our focus on this paper is information retrieval from Twitter posts.

The term **Stream Data** refers to data that is retrieved continuously from the Internet or other sources to be analyzed or for any other purpose. If the results of analysis are required quickly, we call its applications a real-time application. Obvious examples of stream data are stock market data and online TV channels. Twitter is also an example of stream data since it provides the user with continuous data from people around the world.

**Information Retrieval (IR)** according to Manning [35], Information Retrieval is the process of discovering information hidden within a big data set. The extracted information cab be named entities, relationships among entities or a a summary whereas the big data set could be unstructured data saved on many computers on the cloud. In other words, it is the process of inferring information from large data files, in our case large streamed sequences of twitter posts. Some researchers call this process Information Extraction (IE) when natural language processing (NLP) is involved.

**Machine learning (ML) and Data Mining (DM)** Machine learning is defined, according to Samuel [52], as collection of methods that can be used by machines to learn from data or prior experience without explicit programming. According to Harrington, the techniques used in machine learning come from many disciplines such as computer science, engineering, statistics, philosophy, psychology and other fields [27]. The power of machine learning ilies in its ability to turn data into information. Data mining may find hidden patterns, predict future trends, and solve time-consuming problems among large amount of data [67]. These are particularly important in the context of Twitter, which produce voluminous data that cannot be effectively handled by humans.

**Classification** is useful when given a set of classes, we seek to determine which class(es) or categories a given object belongs to. For example, in text classification, each class may correspond to a topic and objects correspond to documents. Each class has a set of documents that are pre-labeled. These documents constitute the training set (or data set). A trained classifier measures the similarity between the given document and the profiles of classes on which it has been trained and then chooses the class(es) to which the object belongs. For this reason, it is called a supervised learning method. The classification process has the following steps:

(1) Collect data,
(2) Normalize data,
(3) Analyze the input data,
(4) Train the algorithm,
(5) Test the algorithm, and
(6) Apply on the target data.

One application of classification methods is the context of Twitter for opinion mining. A service of a product can be evaluated based on people's opinions, where the classifier decides positive and negative opinions.

**Clustering** is unsupervised learning, where no label or target value is given for the data. Clustering is a method of gathering items or (documents) based on some similar characteristics among them. It performs categorization of data items exclusively based on similarity among them. Most clustering algorithms need to know the number of categories in advance. Some researchers use clustering instead of classification in topic detection because it hard to find data set for new topics.

## 3. MOTIVATION

Processing of short text is not a new problem that has suddenly come to the fore in the past few years. It has been a problem since people started using text messages at least twenty years ago. However, the immense popularity of social media has increased the need to provide efficient techniques to overcome the challenges posed by publicly available short texts. Social media has received attention of both researchers and decision makers due to the high rate of increase in the number of active and engaged users. The analysis of short text has seen some progress in areas other than social media as we going to show in next section of Section 6.
In addition to short texts, stream data also causes many issues in analysis of social media. There are two main issues: the need to process large amounts of information and challenges of handling steam data. Visualization helps handle stream data, but requires many challenges to overcome [48]. However, in social media,

researchers usually prefer to work with static data instead of dynamic data since static data is more stable. Static data does not require real-time analysis and presentation, so there is no time concern for filtering and preprocessing.

Our motivation is to improve the classification of social media texts by taking advantage of techniques that have been designed to solve these issues in other fields. We highlight these solutions and discuss the feasibility of applying these techniques to text classification in the context of social media. We focus on Twitter due to its popularity in research, the simple structure of Twitter data, and open data availability.

## 4. CLASSIFICATION TECHNIQUES

In this section, we show how we can choose a technique that satisfies specific requirements. Then, we review popular techniques that have been used for text classification in general.

### 4.1 Choosing the right algorithm

Choosing a classification algorithm to apply is confusing since there are many. In this section, we present issues that need to be considered before choosing an algorithm.

*4.1.1 Define the research goal.* What is the purpose of the data analysis? Is it to predict a particular event or action? Should we use supervised learning? This is like forecasting weather based on the past several years of recorded weather information for the same location. If we are looking for yes/no values, classification is the choice. If the values to predict are in a number range, we use regression. We choose clustering if the outputs are separate groups. If no targeted values are needed, unsupervised learning is the appropriate option [27].

*4.1.2 Does speed matter?.* An important question for researchers who work on mining stream data is whether the classifier result is immediately required or not. If results are immediately required, the classifier needs to focus on performance fewer and faster similarity computation, leading to faster classification. One method to perform high-speed classification is to compare with fewer documents from the data set, leading to quicker results[14]. This technique, however, affects the quality of the classifier because the comparison is done incompletely.

*4.1.3 What is the size of the data?.* A new area that has grabbed attention recently is working with big data. There is a lot of research underway to classify huge amounts of data such as data coming from scientific instruments that continuously generate large amounts of data. Data generated by millions or billions of mobile devices, and sensor technology that uses large number of devices to track objects and large scale computer networks can generate voluminous amounts of traffic data. This amount of data needs efficient classifier algorithms that provide reasonable processing time.

### 4.2 Text Classifiers and Short Text

Several techniques are commonly used for text classification. Classification is different from clustering in two important manners: how the data set is used and whether the number of categories is known a priori. First, classification uses a training set of documents $D = \{d_1, d_2, \ldots, d_n\}$, each document from the set $D$ has been assigned to a set of classes e.g., $C = \{Sport, Politics, Entertainment, \ldots\}$. The classification process develops a model of the classes and may use all documents to do so. Later it uses the model to find the closest class to a given document to be classified. Clustering does not use pre-categorized documents; it alternatively measures similarity among documents, using a similarity measure, and organizes the most similar documents together into a number of categories. The number of categories has to be given a priori based on the data set or other criteria. Clustering usually requires more processing time to compare all documents with each other. However, when the data is streamed, it means the size of the data is immeasurable and dynamic in content and topics. This makes clustering hard to apply in social media text classification. In the next section, we present a few different techniques for classification and how they have been used in the content of social media. We also discuss how researchers have overcome the problem of short texts when they use these techniques.

*4.2.1 Naïve Bayes.* A technique used widely is Naïve Bayes classification, which uses a probabilistic algorithm that takes into account the frequencies of appearance of terms in documents as well as in labels. Therefore, it assigns class $C$ to document $d$ by counting terms and computing the probability that a document $d$ belongs to class $C$ using the Bayesian formula. Researchers have tried to improve Naïve Bayes classification through methods such as reducing the number of probabilities by using index term selection methods [55]. Using a combination of Expectation-Maximization and a Naïve Bayes classifier is another way to increase the accuracy of the classification [43].
Naïve Bayes is an easy technique to apply for text classification. Examples for situations where it has been used are recommendation systems [53], topic detection [58], finding trending topics [33], spam detection [41], and summarizing social media-blogs [57]. Another way to improve Naïve Bayes classification process is to use unlabelled documents for training in order to get high correlation between a word $w$ and a class $L$ [13].

*4.2.2 K-Nearest Neighbor.* In contrast to Naïve Bayes classification, which is based on a probabilistic algorithm, the K-Nearest Neighbor classifier [10] relies on mathematically measuring the distance between documents that are presented. All documents are pre-labeled as belonging to its appropriate class. The cosine similarity measurement is commonly used with $K$-$NN$. If the given document $d$ is similar to a document that is labeled to class $C$, we can assign document $d$ to this class as well. Often, the classification accuracy of $K$-$NN$ can be improved significantly if Large Margin Nearest Neighbor or Neighborhood component analysis is applied [29]. The use of K-NN for text classification is widespread; however, it needs a significant pre-processing stage. This pre-processing stage transforms documents and words into sparse numeric vectors which use linear algebraic operations [30], which are necessary for most classification techniques as well.

*4.2.3 Decision Trees.* Quinlan represent a method to automatically generate rules that can be applied on data points by training on labeled data [47]. One starts with one rule to split data into two or more sections, then each section is again split by a different rule. The process is repeated until a point is reached such that data is classified into homogeneous groups. Consequently, the Decision Tree assigns the given document to that class, which is supposed to be the mo st similar. The advantage of using this approach is that it is faster compared to other techniques [29]. However, Decision Trees have the disadvantage of relying only on a few terms of comparison.

*4.2.4 Support Vector Machine.* A Support Vector Machine (SVM) [11] is a supervised classification algorithm that splits data

into classes based on the widest margin between points in the classes. Linear SVM, the most commonly used, separates classes using a hyperspace given by $w \times x - b = y$. $Y$ is referred to as a linear separator which is trapped between upper class margin $y = 1$ and lower margin $y = -1$. A binary SVM algorithm takes positive and negative examples of the training set and draws a hyper-plane to separate two classes [11].
In social media, SVM has been used to solve many problems, one of them being opinion mining. O'Connor et al. extract tweets to measure people's satisfaction regarding a product [44]. The approach relies on SVM to divide Twitter posts into positive and negative classes based on the appearance of sentiment words. They compare classification techniques for text and determine that SVM produces the most accurate results among the methods tested [17]. Similarly, Zubiaga et al. classify Twitter posts to infer trending topics [73]. In addition to four types of trending topic triggers known as news, events, memes, and memoratives, Soboroff et al. use fifteen additional straightforward features or characteristics to predict the spread of trending topics [59]. The authors use SVMs to classify the trending topics and categorize them automatically.

### 4.3 Mining Stream Data Using Text Classifiers

Babcock et al. address technical challenges regarding the speed of classification algorithms, unexpected sizes of data, and processing on the fly [3]. The algorithm should be fast and efficient, which is hard to achieve at the same time with stream data. The incoming data need to be measured in order to choose the best fit algorithm [69]. Because of the many challenges of applying data mining techniques, some researchers have come up with new techniques or sometimes modifications of standard techniques to overcome these challenges.

An example is trending topic identification by Lee et al. who classify trending topics into general categories using text-based modeling and network-based modeling [33]. The authors limit trending topics to be only the top five terms that come from the most influential users' tweets and call it $C5$. The classifier shows best accuracy when $C5$ is used instead of $K\text{-}NN, SVM, LogisticRegression$ [7], or $ZeroR$ in the same research.

Multinomial Naïve Bayes is used widely since it is fast for text classification [40]. It considers a document as a bag of words. It computes the frequency of each word in the collection of training documents of that class and it obtains the probability of a word occurring in documents of the class. The probability of class $c$ given a test document is calculated as follows.

$$P(c|d) \propto p(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (1)$$

Equation (1) Calculating the probability of class $c$.
$P(c|d)$ is the probability of a document $d$ in class $c$, $p(c)$ is the prior probability of a document (tweet) occurring in class (topic) $c$, and $P(t_k|c)$ is the conditional probability of term $t_k$ occurring in a document of class $c$. Pak and paroubek use classification based on the multinomial Naïve Bayes classifier for sentiment analysis [45]. Similarly, Bifet and Frank claim that multinomial Naïve Bayes provides higher accuracy classification when it is used in topics trending discovery.

## 5. THE CHALLENGE OF SHORT TEXT CLASSIFICATION

In this section, first we review current approaches to overcome issues in classifying short texts in the context of Twitter posts. Each approach has advantages that make it perform well in some situations. However, these approaches show some drawbacks, especially when working on datasets such as Twitter data.

### 5.1 Current Approaches

We review the techniques that have been applied in text classification to overcome problems created by the short length of Twitter posts.

*5.1.1 Many Tweets in a Single Document.* One of the methods to handle tweets is to combine many Twitter posts as a single document, based on common features. For example, a document may contain all posts that are obtained when we search for a keyword. The process of analysis then becomes easy since we deal with a single document for each keyword. An example is using keyword appearances during a specific period of time to summarize Twitter posts [57]. Before collecting documents, Sharifi and Kalita use a Naïve Bayes classifier to remove spam and irrelevant tweets. They also remove non-English and duplicated posts. Then, They collect all Twitter posts based on the appearance of a keyword or a hashtag into a single document. Their algorithm analyzes the tweets to generate one sentence that summarizes the content of a document. Given a number of posts within a document, they centralize the common keyword as the root in the middle to build the sentence and generate an ordered path of words on both sides of the root considering the position of words in posts within the document. The algorithm weights paths of words from the beginning to the end through the central keyword, by taking into account weights of individual words along the path. Word weight is calculated by the frequency with which a word appears in a tweet divided by frequency of its appearance in the training set (relevant posts). After calculating the weight for all paths, the algorithm chooses the path with greatest total weight among all generated paths.

Wing and Baldridge split the collected tweets so each user's tweets are in a single document [66]. Then from its content, they predict the user's location based on the textual content and a geodesic on earth. The prediction relies on a simple supervised method based on three distributions. First is $\kappa$ which is a standard multinomial distribution of vocabularies over a map grid. Second distribution is $\theta$, the distribution of single document. Third, $\rho$ distribution that is the reverse of the first distribution.

Another example of using a single document for a single user is topic modeling [28]. Hong and Davison classify Twitter users and their messages into topical categories, obtained from Twitter suggestions. The users are picked from a trusted list from $wefollow.com$ and 150 messages from each user are aggregated. They compare three schemes called MSG, USER, and TERM schemes. The difference among the schemes is the training method–whether it trains on messages, user profiles, or messages that contain a specific term, respectively. The USER scheme provides the best result (82% accuracy for message classification) when it is used with $TF\text{-}IDF$ (term frequency inverse document frequency) weighing [51] and topics $T = 50$ Jensen-Shannon (JS) divergence [18] is used to measure the similarity among the docu-

ments probability distributions.

$$D_{JS} = \frac{1}{2} D_{KL}(P||R) + \frac{1}{2} D_{KL}(Q||R) \quad (2)$$

Equation (2) Calculating (JS) divergence.

$$R = \frac{1}{2}(P + Q) \quad (3)$$

Equation (3) Calculating R.
Where $D_{KL}(P||R)$ represents the KL divergence between document $P$ and $R$. Here

$$D_{KL}(A||B) = \sum_{n=1}^{M} \phi_{na} log \frac{\phi_{na}}{\phi_{nb}} \quad (4)$$

Equation (4) Calculating KL divergence.
Where $M$ is the number of distinct term types, and $\phi_{na}$ is the probability of term $n$ in topic $a$. This approach has been leveraged by other researchers to improve a recommendation system based on the same methodology [26]. They consider a user's profile as a source that is used to calculate weights of terms and IDs so it can be used in their recommendation system "Twittomender". A user profile includes user's tweets, followers, followees, followers' tweets, and followees' tweets. They use $TF$-$IDF$ as a weighting metric, which is shown in these equations:

$$TF.IDF(t_i, U_T, U) = tf(t_i, U_T) \bullet idf(t_i, U) \quad (5)$$

Equation (5) Weighing Metric.

$$tf(ti, U_T) = \frac{n_{i,T}}{\sum_k n_{k,T}} \quad (6)$$

Equation (6) Calculating term frequency.

$$idf(t_i, U) = log \frac{|U|}{|\{d : t_i \epsilon d\}|} \quad (7)$$

Equation (7) Calculating inverse document frequency.
where $t_i$ is a term, $U_T$ is user who wrote the tweets, $U$ is a set of users, and $T$ is a set of tweets. $d$ is referred to as the document profile and $k$ is number of documents. Therefore, in 6, $n_{i,T}$ is the number of a specific term in document $i$ divided by the total number of term $t$ occurrence on all documents. They compare nine recommendation strategies within two groups. The first group contains content-based strategies that represent users by their tweets, followees' tweets, followers' tweets, or combination of all tweets. The second group contains collaborative filtering style strategies that represent users by IDs of followers, followees, or both. The last two strategies are combination of some previous strategies. They found that the recommendation strategies perform equally well. However, performance correlates with the number of recommended users inversely and the best number is around 5 users.

Information about the profile of a Twitter author may help handle issues in classification when many tweets are treated as a single document. Sriram et al. classify tweets using author information as domain-specific features [60]. The author information consists of the authors' profile and posts written by the author, where categories include news, events, opinion, deals, and private messages. They use simple bag of words and Naïve Bayes classification with author topics so they can narrow them to the author topics only. The author selection features are presence of slang and abbreviation, event phrases, mentions from other authors, currency signs, opinion terms and word emphasis. This approach relies on analyzing these features and building a learning model so the algorithm can be trained using a training set and then categorize tweets automatically based on the learning method. For example, if slang and abbreviation words do not appear in a tweet, it might mean it is news-related. The advantage of this approach is the ability to exploit features that are included in the tweet itself without using outside sources such as Wikipedia.

*5.1.2 Collecting Tweets Based on Time Frame.* In order to measure a topic's popularity over time, tweets may be saved in daily documents. Then, one can find relations among the topics that appear in all documents to show the topics' lifespan and strength. Weng and Lee measure topic popularity by applying approaches for time series analysis to track topic appearance in each document [64]. A time series is usually used to predict the future based on past statistical data on topics such as weather. The approach's drawback is the difficulty of tracking many topics if documents are very large. For example, a daily document may contain up to 5% of 500 million Tweets per Day (TPD) that developers and researchers are allowed to retrieve from Twitter.

Some researchers use shorter time frames such as earthquake detection system [61], Sakaki et al. were looking for specific keywords regarding earthquakes in specific areas in order to detect the occurrence of earthquakes. A time frame of 10 minutes is used to measure the appearance of the required keywords. This system implements a near real-time method (update every ten minutes) and it takes advantage of the use of short texts to measure keywords' appearance. This research classifies Twitter posts using a linear kernel SVM to either a tweet includes keywords that are related to earthquake occurrence or it just a normal conversion.

*5.1.3 A Single Tweet as Single Document.* The third approach that researchers rely on to handle the shortness of texts in Twitter, is to treat each tweet as a single document as usual [45]. Pak and Paroubek conduct sentiment classification and opinion identification using this technique. The reason for applying this method is that each tweet contains emotional symbols that signify message sentiment to indicate if it is positive, negative or neutral. Therefore, gathering tweets into a single document may not provide the same result. In other words, if we gather emotional symbols form various tweets in a single document, the emotions expressed will be mixed up. Because these emotional symbols are assigned to terms within the tweets, the document may provide incorrect results when tweets are combined.

## 5.2 Alternative Solutions

In this section, we discuss techniques that may help overcome difficulties in classifying short texts. However, the techniques we discuss have not been used for classification on social media texts yet to the best of our knowledge. The methods have been used in clustering such as in topic detection.

*5.2.1 Enrich Twitter Posts.* One way to overcome the problem of classifying short texts is to make them longer. Therefore, we need to find related content that is similar to the short text we have. The relation may be with content in documents obtained from either internal sources or external sources. Using an internal sources means that information inside the tweets themselves is used in the enriching process by extracting word synonyms from tweets. Hu

et al. use WordNet to extract synonyms of each noun in a snippet of text [31]. To improve clustering for short text, they use these synonyms and Wikipedia named entities in addition to the original text. The purpose of clustering is to understand these snippets by using external knowledge.

Another method to leverage word semantics is by Romero et al. who generate a semantic analysis of each term, entered by a user in a tweet by finding its synonyms, and then by performing web search with these words [49]. This technique is also used in search engine algorithms used by Google and Bing [35], which do not show the exact match words only but also include the synonyms of these words in performing matches. Weng and Lee go beyond this by stemming words and using all possible prefixes and suffixes in the search process [64]. This approach improves search engine results. In classifying social media texts, we may consider limited semantic analysis as a secondary tool when classifying.

In contrast, external sources indicate using content beyond the tweet's contents in the enriching procedure. An example of using external content is [1], where they use news articles to enrich the contents of tweets. This approach links a tweet to the content of news articles found at the URL included in a tweet. The purpose is to understand the meaning of hashtags or ambiguous content in the tweets. They measure the similarity between a tweet $t$ and news article $s$ with $TF\text{-}IDF$ score.

*5.2.2 Compare two short texts.* In contrast to gathering tweets in one file, another solution is to limit the comparison to short texts. In other words, make data set documents short to make the comparison fair between a given document and the documents in the data set. This is used by Sahami and Heilman who simply convert two texts to queries before comparison using Google as the search engine [50]. They next rank the suggested queries using $TF\text{-}IDF$ to weight the result as in this equation:

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \quad (8)$$

Equation (8) Ranking queries using $TF\text{-}IDF$.
Where $tf_{i,j}$ is the frequency of term $t_i$ in document $d_j$, and N is the set of documents in the corpus. $df_i$ is the total number of documents that include term $t_i$. The queries have similar information except they are written in a different format or in different words. The comparison then becomes between the original text with top queries on one side and the second text with its top queries on the other side. The common features between this method and the previous methods is that both rely on using outside resources. However, this cannot be applied in the case of Twitter posts classification because the average length of tweets is longer than an average length of a query which lead to inaccurate result. If the query is long, the outcome extend to include unnecessary result.

Thus, we may limit the query to the most important keywords in a tweet. We pick these keywords based on weight calculation. The calculation aims to find the most popular words among many tweets after filtering the tweets words for a certain topic. We may also apply this method on other social media that uses smaller vocabularies such as in Instagram. The text in Instagram is more pliable since posts describe a picture and may include hashtags.

*5.2.3 Manipulate Classifier Rules.* Specifying rules based on word usage for classification could improve performance. [20] develop an algorithm using rules and statistics to classify short texts. They claim that classification performs best if rules and statistics are added compared to regular techniques of classification. The essence of this algorithm is to use statistical calculation $X^2$ that weights words as following:

$$W_{weight} = \frac{N \times (A \times D - B \times C)^2}{((A+B) \times (C+D)(A+C) \times (B+D))} \quad (9)$$

Equation (9) Calculating $X^2$ weight.
where $N$ is the total number of documents, $A$ is the number of documents in category $C_i$ that word $w$ appears in, $B$ is the number of documents in other classes that $w$ appears in, $C$ is the number of documents in category $C_i$ that $w$ does not appear in, and $D$ is the number of documents in other classes that $w$ does not appear in. In Twitter we may use the same calculation with documents that include tweets regarding a specific topic for example.

*5.2.4 No Need to Solve.* With all these complications with classification of short texts, there is an opposing opinion for working with short texts. A few researchers consider short text an advantage, citing the famous quote "The least said the better." For this reason, Bermingham and Smeaton claim that sentiment classification for short reviews is easier than for regular long reviews [5]. The reason is that when the judgment is made on a review, the decision is made based on a few term appearances. In short texts, once the terms appears, the decision is made quickly. However, in longer documents such as movie reviews include many appearances of opposite terms, which may lead to inaccurate judgment. The classifier used by Bermingham and Smeaton are Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB). The accuracies were around 70% for microblog classification, around 60% for blog post classification, and around 80% for movie reviews classification using unigrams, bigrams and trigrams. This research has not been applied to Twitter posts, but the problem is quite similar in tweets.

## 6. THE CHALLENGE OF MINING STREAM DATA

Stream data means that the data is retrieved continuously. Twitter provides for retrieving posts through their API with some limitations. The challenge of analyzing this kind of data lies in reducing the difference between the time of retrieval and the time of analysis. A delay in processing one of these tasks creates a cumulative gap between the analysis time and the retrieval time. Classification is also affected by the huge number of short tweets spread out over many different topics, and hence, with many features.

Static data is useful and easy to analyze. However, there are situations where continuous monitoring is required. For instance, decision-making, trend detection, event tracking and monitoring require real-time analysis; otherwise an analysis is useless [19]. Furthermore, for stakeholders the information should be comprehensible and quick to obtain. Therefore, visualization is one of the tools that we need in order to provide the information in a comprehensible manner.

In stream data, time limitations is hard to control. One of the goals of this survey is to present techniques that are able to handle stream data well. We can then determine if these techniques can be applied to classify social media posts as well. For example, stream data could be saved and later analyzed within a flexible time period such as an hour or 24 hours. Although the choice of a time window is flexible, the analysis process is performed the same. The only difference is the size of retrieved data, which is sometimes

critical. In other words, classification time may become substantial with stream data. The time includes retrieval time, filtering and noise removal, extraction of information, and classification time. Retrieving data from social media differs from retrieving regular information from a website. The enormous amount of data coming from social media needs to be processed on the fly since saving such data is costly. For example, in context of newswires articles, Buehler et al. use a database buffer to save the text stream from newswires to overcome the costly storage problem and make the approach fast [2]. The cost was high because the visualization is continuous for stream text. So instead of saving text in database, they save the statistics regarding the text that is visualized.

Rohrdantz et al. address the challenge of visualizing stream data and identify three scenarios where it is necessary. The first is emergency management and response, which includes mining user generated text [48]. This may help track changes in stream data in order to detect and monitor unplanned emergency events, e.g., hurricanes, accidents, and attacks. For instance, Sakaki et al. use Twitter as a monitor to detect and track earthquakes using the users as sensors [61]. News and the stock market are the second scenario domains that need fast and real-time information analysis according to Rohrdantz et al. [48]. Timely and useful information helps stakeholders prevent financial loss and gain competitive advantage in pursuit of profit. The third scenario is server administration and log management, where data is analyzed for security purposes.

Static data is useful and easy to analyze. However, there are situations where continuous monitoring is required. For instance, decision-making, trend detection, event tracking and monitoring require real-time analysis; otherwise it is useless [19]. Therefore, visualization is one of the tools that we need in order to provide the information in a comprehensible manner.

For marketing, knowing the total numbers of followers for an individual is useful to accurately measure influence of a tweets author. Measuring influence of Twitter posts has received more attention than the analogous problem for any other social network because of the current popularity of Twitter. Consequently, research is focusing on finding more accurate ways of measuring the influence of Twitter posts based on different variables. Recently, news sources and websites have begun to measure the power of a speech or an event by the tweets per minute (TPM) metric[3]. The beginning of this metric dates back to the 2012 Olympics in London. TPM has become a popular choice because it is a quick way to get a barometer on the opinion of the masses. Even though all social media sites have the ability to apply such a technique, Twitter's use of such a metric has been impactful [34]. However, when people mention a user in their posts, The TPM measurement shows the number of people who are talking about an event, but it does not necessarily reflect how much people are interested in the topic because the content may include different cognitive dimensions of emotions.

Tsur and Rappoport introduce an algorithm that determines whether a hashtag trend contains positive or negative tweets [62]. The algorithm calculates the most frequent words that appear within the hashtag and maps it to the 69 Linguistic Inquiry and Word Count (LIWC)[4]. If the most frequent word is positive, it means the talk regarding the topic is positive. Not all people use the same vocabulary on Twitter, and the noise included in these tweets reduces the accuracy of any result. For example, the #CNN hashtag brings up the most recent news from the news source CNN, so people can get updates of breaking news by following the hashtag. However,these posts do not contain positive or negative words because the hashtag is not for specific event.

### 6.1 Current Approaches

Many approaches for stream data in social media rely on processing saved data as a first step. Consequently, time becomes critical with THE saving and analyzing processes. Developers save the information they want to analyze and classify. The drawback of these approaches is the time constraint. However, if classification works continually on the stream data, the result is fast. A state of the art approach to detect earth quacks that uses this technique is the earthquake detecting system provided by [61]. The system, as we mentioned earlier, like a physical earthquake detector, considers Twitter users as sensors that provide information to the detector. The time frame for calculating theses keywords is 10 minutes or less. What makes it state of the art is the quickness of providing earthquake alarm by analyzing stream data.

Predicting events of all sorts from social media has become a topic of great interest in the last few years [54]. Much work has been done to forecast information based on people's opinions, starting from disasters and emergencies such as predicting flu trends around the world [12]. Researchers have also attempted to predict the stock market based on Twitter users' opinions and how much their opinions are reflected in the movement of the stock market itself [72]. Zhang et al. approach is simply tracking certain emotional words, and tag tweets based on them such as being fearful, happy, or hopeful. Zhang et al. found that the more emotional words that appear in Twitter the more it is that people are worried, even when there are positive emotional words. Measuring mood may be helpful in this situation. For example, Bollen et al. attempt to predict the Dow Jones Industrial Average (DJIA) by developing two tools: OpinionFinder and Google-Profile of Mood States (GPOMS) [6]. OpinionFinder determines if a sentence is positive or negative using Part of Speech POS information and a dictionary list of words created by Wilson et al. [65]. Google-Profile of Mood States (GPOMS) has 6 dimensions of mood measurement $(Calm, Alert, Sure, Vital, Kind, Happy)$, which are derived from the Profile of Mood States (POMS-bi) [42]. Unfortunately, Bollen's results are not promising, as they claim, because emotional analysis in Twitter was not able to predict the stock market movement [6].

Sellers and creators also would like to measure opinions about a product they sell or a service they provide. For example, movie makers attempt to measure people's opinions via Twitter users. Wong et al. investigate the ability to predict whether there is a relation between users' opinions and Oscar nominations for films [68]. First, Wong et al. determine if a tweet is relevant to a movie. Then, they classify tweets' sentimentality as "negative", "positive", or "mention" (does not contain positive or negative terms). Then, they regulate the temporal context of a tweet: whether it is written before, after, or concurrent to the time of release a movie. They use SVM in the classifier, which perform classification more efficiently

[3]https://blog.twitter.com/2013/behind-the-numbers-tweets-per-minute

[4]http://www.liwc.net

than using Naïve Bayes. They found that Twitter users have more positive reviews than other rating sites and they show that there is no relationship between box office gains and social media reviews. Another group of people who are interested in prediction from social media are politicians. Gayo-Avello et al. attempt to predict winners of election in the 2010 US congressional elections race [23]. They use a polarity lexicon to determine positive, negative, and neutral terms. Gayo-Avello conducts a survey of a large amount of political prediction work done based on social media, especially Twitter [22]. Gayo-Avello claims that most current work is weak and does not present a reliable prediction methodology in the context of social media. Therefore, he provides some suggestions to improve techniques such as choosing the duration and method of collection, data cleansing measures, and performance evaluation. In conclusion, at this time all research done to predict stock market or elections from social media are unsuccessful or do not perform as expected [54].

A problem that can benefit from real-time analysis of social media text is spam detection. McCord and Chuah work on detecting spammers by divide the approach into collection followed by classification [41]. The collected data is saved for different time durations, making it easier to analyze. The drawback of this process is the costly side effect of saving data physically. This technique has not been applied on Twitter for the general public so people can detect spammers. Consequently, the problem of credibility of Twitter content still exists for regular users of Twitter [9]. This motivated Gupta et al. at Qatar Computing Research Institute (QCRI) and the Indraprastha Institute of Information Technology (IIIT) to develop an evaluation tool. The tool, called "tweetcred," [25] is publicly available for download from the Google Chrome Extensions store [5]. The tool is integrated with the Twitter website to show a meter, beside a username, with a scale of seven stars that shows an evaluation of the user's credibility. They measure credibility based on information within the tweet's content. There are around 45 features such as URL appearance, number of followers, and tweet's length. In addition to all these features, tweetcred learns to improve evaluation since it receives feedback from users. The speed of the process, according to the paper, is 6 seconds to show the evaluation. The classifier used is SVM-rank [32] trained on data from six high impact crisis events of 2013.

Topic detection in social media allows an analyst to identify topic of high interest in social media. Topic detection helps reveal what topics people are spreading and, how fast the spread is. According to Petrović et al., celebrity death is one of the fastest spreading news topics on Twitter. With high growth in the number of tweets, it is difficult to have an algorithm which works efficiently in real time [46]. Wang and Lee [64], from HP Labs, demonstrate that trending topics can be performing statistical computing on the words present in the posts [39, 24]. Benhardus and Kalita develop an algorithm that identifies trending topics by computing term frequencies and the inverse document frequency [4]. They start with normalized term frequency within document $d_j$ as shown in this equation:

$$tf_{norm_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} * 10^6 \tag{10}$$

Equation (10) normalize term frequency.
Where $n_i, j$ presents the appearance of term $t_i$ and the summation is the number of total words in document $d_j$. Then use trend scoring for each word using this equation:

$$ts_{i,j} = \frac{tf_{norm_{i,j}}}{atf_{norm_{i,s}}} \tag{11}$$

Equation (11) Trend Scoring
in which

$$atf_{norm_{i,s}} = \sum_{S=\{s_1,\ldots,s_p\}} \frac{tf_{norm_{i,s_k}}}{P} \tag{12}$$

Equation (12) Appearance of Term Frequency. Where $S$ is the set of $p$ baseline documents to which the test document was compared. Another technique for topic detection is by Gao et al. who use the detected topics to create a summary of Twitter posts [21]. They detect sub topics for each main topic in order to build the summary. They develop the Offline Peak Area Detection (OPAD) algorithm that detects a peak area $P$ and its appearance period. The algorithm uses Transmission Control Protocol (TCP) congestion detection [36] that computes mean (the sum over every possible value weighted by the probability of that value) and variance (measure the distance among numbers and how they spread out) to find the being of peaks $P$ in Twitter streams $S$ and the number of tweets.

## 6.2 Alternative Solutions

In this section, we provide some suggestions to overcome the issues with stream data that we have discussed earlier. First, we look at how to deal with the infinite length of stream data. How much data should be saved to get the best result when resources are limited? Another challenge is limited time, especially when results are demanded quickly. We review some solutions that have been used in various areas, so that we can learn from them if we want to improve classification of social media texts.

There is not much prior research on text classification in real-time. Instead most research performs analysis in almost real-time by saving data and then analyzing it. Buehler et al. visualize stream data coming from consumers to facilitate recognizing system bugs [2]. To overcome the time constraint challenge when receiving vast numbers of emails from customers regarding bug report, they visualize only the most important keywords, where importance is determined by term frequency. If the term frequencies of words in an email are high, the term is included in the display graph. They also build connections among terms that are present in the same email, and the length of the connection between two terms becomes shorter if they appear together more frequently. Therefore, we may be able to apply this technique to identify the most important Twitter accounts that need quick responses in an emergency. Tyshchuk et al. recommend that it is necessary to engage social media in developing emergency plans for all organizations, and suggest developing data mining techniques to detect emergency situations from analysis of social media [63].

Dubinko et al. investigate the impact of using different sizes of buffers to perform classification [16]. They observe that when the data being classified is static, the size of the buffer does not matter. However, if classification requires automatic update for data set training to work with the dynamic of data, the buffer size affects the result. Statistical information is impacted by buffer size because the size is computed over long periods. However, they did not determine ideal size for buffering since every situation has different requirements. In classifying social media texts, buffering is important especially with the huge number of posts on Twitter. Another

[5]https://chrome.google.com/webstore/detail/tweetcred/fbokljinlogeihdnkikeeneiankdgikg

important aspect that requires attention when classifying social media text is measuring a user's influence. If a famous celebrity has millions of fans (followers), they see their tweets and spread news in seconds while if the same tweet is written by a user with low number of followers; it will not have the same impact. Therefore, the buffering system should be used wisely since the sizes of buffers impact analysis results in situations such as measuring relationships amongst networks or users in social media. Cha et al. track relationships and interactions among users such as the number of followers, retweet totals, and mention numbers [8]. They find that, among 6 million users, there are only 233 users who have garnered massive amount of mentions and retweets. They use Spearman's rank correlation coefficient as a measure of the strength between two rank sets:

$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{N^3 - N} \quad (13)$$

Equation (13) Calculating Spearman's rank.
Where $x$ and $y$ are the ranks of users based on one of influence measures in database of $N$ users. Hence, they suggest focusing on the most influential users instead of focusing on all users.

Recently, visualization has become the focus of attention for researchers in many areas because it makes it easier to show results to users and researchers in the presence of massive amount of data. Stanford University has just built a facility for collaborative scientific visualization[6]. This lab helps researchers visualize their data by taking advantage of supercomputers and high definition screens. However, because the lab is new, there is no facility for text visualization yet. Efforts have been made to involve data visualization in business as well. Pivotal[7] is a company that provides its members with tools to visualize data. It is simple and easy to apply since it uses the Hadoop[8] framework written in library. The company provides an example[9] that presents Twitter hashtags using their application that simply runs statistical analysis over real time posts and presents them in terms a number of circles where sizes of circles indicate the number of hashtag appearances to adapt the tool to show popular keywords in Twitter as well. These two examples show the trend in present data visualization instead of just showing numbers and graphs, but no much effort has been made in visualizing social media text.

To visualize text in social media, [15] concentrate on visualizing people's conversations during significant events by tracking hot keywords using term frequencies within a time window. Another feature is providing related photos for an event. This feature is already provided by Twitter when a user searches for a keyword related to an event or a hot topic. Dubinko et al. visualize tags from Flickr (flickr.com), a social network to share images in order to detect hot topics [16]. Xu et al. compare the popularity of two Twitter topics as a case study: the 2012 United States presidential election and the Occupy Wall Street movement using agenda setting to characterize the dynamics of topic competition and impact of opinion leaders [70].

There is another significant problem for mining stream data, referred to as concept-drift. According to Masud et al., concept-drift indicates change in the topics being discussed in stream data, requiring generation of new data sets for the classifier to get accurate results [38]. Masud et al. propose an algorithm that uses a decision tree with additional ability to determine whether new data belong to a particular known class or belong to a novel class automatically. They apply the algorithm on classifying data for Network Intrusion Detection (KDD Cup 99)[10] and Forest Cover[11], showing greater accuracy and speed compared with $K$-$NN$. In their follow up work, Masud et al. overcome concept-drift by using an adaptive decision boundary and the Gini coefficient [37]. They build a graph to detect multiple novel classes and determine the connections between included components. This is needed in social media since the topics and users are changeable rapidly and affected by many conditions and situations. In order to apply this technique to classify social media text, we need to obtain values for the two parameters that Masud et al. use in their technique. It may be advantageous to use various data sets for each chunk of data, following their lead.

## 7. DISCUSSION

Twitter does not have a long history but it has received a great deal of attention since its founding in 2006. It is easy to get in a conversation with other people and exchange short messages. Twitter provides open source access for researchers, analysts or anyone interested who want to work on their posts. This has encouraged a large number of research publications on various topics related to Twitter posts. All of the above has encouraged us to dig into classification and follow the state of the art of the work that has been done regarding social media, concentrating on Twitter. The first challenge, short text classification, has received a good deal of attention and was very active between 2009 and 2011 since it twitter was born. Some solutions have been successfully found as mentioned earlier.

Mining streamed text data has recently become popular as daily posts on Twitter reached 250 million in 2012 and 500 million in 2014. Many papers have been published dealing with the two issues of mining stream data: incremental classify and infinite length.

We have seen many researchers focus on each challenge separately and ignore other challenges. The researchers need to work on solving all challenges simultaneously to achieve research purposes such as prediction or topic detection. Lastly, there are many other areas such as text visualization, incremental data analysis that need to be studied in order to get a handle on classification and analysis of the fast growing number of Twitter posts.

## 8. CONCLUSIONS

In this paper, we review text classification techniques that have been used to overcome the challenges posed by the short length of messages like those found in social media sites such as Twitter. We also present techniques that have not yet been used to classify short text in the context of social media. The second challenge in classifying social media text is that the data is streamed. We review stream data classification techniques that have not been used for classifying social media text. As future work, we would like to implement the suggested solutions to simultaneously overcome challenges caused by short text and stream data. We will evaluate

[6] http://news.stanford.edu/news/2014/june/hive-open-house-060514.html
[7] http://blog.pivotal.io
[8] http://hadoop.apache.org
[9] http://blog.gopivotal.com/pivotal/products/spring-xd-for-real-time-analytics
[10] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[11] http://www.wri.org/applications/maps/forest-cover-analyzer

the performance of these techniques. In addition, We want to extend the research to include classifying all kinds of streamed text such as news feed instead of focusing only on social media.

## 9. REFERENCES

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter Leenheer, and Jeff Pan, editors, *The Semanic Web: Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 375–389. Springer Berlin Heidelberg, 2011.

[2] C. Albrecht Buehler, B. Watson, and D.A. Shamma. Visualizing live text streams using motion and temporal pooling. *Computer Graphics and Applications, IEEE*, 25(3):52–59, May 2005.

[3] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 1–16, New York, NY, USA, 2002. ACM.

[4] James Benhardus and Jugal Kalita. Streaming trend detection in twitter. *Int. J. Web Based Communities*, 9(1):122–139, January 2013.

[5] Adam Bermingham and Alan F. Smeaton. Classifying sentiment in microblogs: Is brevity an advantage? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1833–1836, New York, NY, USA, 2010. ACM.

[6] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[7] S. Le Cessie and J. C. Van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):pp. 191–201, 1992.

[8] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, volume 14, pages 10–17, 2010.

[9] Adrian Chen. Can an algorithm solve twitter's credibility problem?, 2014. [Online; posted 5-May-2014].

[10] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[11] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[12] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York,NY,USA, 2010. ACM.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal Of The Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[14] Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 71–80, New York, NY, USA, 2000. ACM.

[15] M. Dork, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1129–1138, Nov 2010.

[16] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Visualizing tags over time. *ACM Trans. Web*, 1(2), August 2007.

[17] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, pages 148–155, New York, NY, USA, 1998. ACM.

[18] D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49(7):1858–1860, July 2003.

[19] Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[20] Zhou Faguo, Zhang Fan, Yang Bingru, and Yu Xingang. Research on short text classification algorithm based on statistics and rules. In *Electronic Commerce and Security (ISECS), 2010 Third International Symposium on*, pages 3–7, 2010.

[21] Dehong Gao, Wenjie Li, Xiaoyan Cai, Renxian Zhang, and You Ouyang. Sequential summarization: A full view of twitter trending topics. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 22(2):293–302, February 2014.

[22] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, 31(6):649–679, 2013.

[23] Daniel Gayo-Avello, Panagiotis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. *CWSM*, 2011.

[24] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[25] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. TweetCred: A Real-time Web-based System for Assessing Credibility of Content on Twitter. *arXiv preprint arXiv:1405.5490*, May 2014.

[26] John Hannon, Mike Bennett, and Barry Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 199–206, New York, NY, USA, 2010. ACM.

[27] Peter Harrington. *Machine Learning in Action*. Manning Publications Co., Greenwich, CT, USA, 2012.

[28] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[29] Andreas Hotho, Andreas Nrnberger, and Gerhard Paa. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20(1):19–62, May 2005.

[30] X. Hu and H. Liu. *Text Analytics in Social Media*. Springer, 2012.

[31] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th*

*ACM Conference on Information and Knowledge Management*, CIKM '09, pages 919–928, New York, NY, USA, 2009. ACM.

[32] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[33] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 251–258, Washington, DC, USA, 2011. IEEE Computer Society.

[34] Lila MacLellan. Tweets per minute social media, 2012. [Online; posted 6-Sep-2012].

[35] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[36] 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.

[37] M.M. Masud, Qing Chen, L. Khan, C.C. Aggarwal, Jing Gao, Jiawei Han, A. Srivastava, and N.C. Oza. Classification and adaptive novel class detection of feature-evolving data streams. *Knowledge and Data Engineering, IEEE Transactions on*, 25(7):1484–1497, July 2013.

[38] M.M. Masud, Jing Gao, L. Khan, Jiawei Han, and Bhavani Thuraisingham. Classification and novel class detection in concept-drifting data streams under time constraints. *Knowledge and Data Engineering, IEEE Transactions on*, 23(6):859–874, June 2011.

[39] Diana Maynard, Kalina Bontcheva, and Dominic Rout. Challenges in developing opinion mining tools for social media. In *Workshop Programme*, pages 15–, 2011.

[40] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naïve bayes text classification. pages 41–48. Citeseer.

[41] M. McCord and M. Chuah. Spam detection on twitter using traditional classifiers. In *Autonomic and Trusted Computing*, volume 6906 of *Lecture Notes in Computer Science*, pages 175–186. Springer Berlin Heidelberg, 2011.

[42] Douglas McNair, Maurice Lorr, and Leo Droppleman. Profile of mood states (poms). Profile of Mood States, 1989.

[43] Kamal Nigam, AndrewKachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2-3):103–134, 2000.

[44] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, 11:122–129, 2010.

[45] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *The International Conference on Language Resources and Evaluation (LREC)*, 2010.

[46] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: 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.

[47] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[48] Christian Rohrdantz, Daniela Oelke, Miloš Krstajic, and Fabian Fischer. Real-time visualization of streaming text data: Tasks and challenges. In *Workshop on Interactive Visual Text Analytics for Decision-Making at the IEEE VisWeek*, volume 201, 2011.

[49] Francisco P Romero, Pascual Julián-Iranzo, Andrés Soto, Mateus Ferreira-Satler, and Juan Gallardo-Casero. Classifying unlabeled short texts using a fuzzy declarative approach. *Language Resources and Evaluation*, 47(1):151–178, 2013.

[50] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 377–386, New York, NY, USA, 2006. ACM.

[51] Gerard Salton, Edward A. Fox, and Harry Wu. Extended boolean information retrieval. *ACM Communication*, 26(11):1022–1036, November 1983.

[52] A.L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, July 1959.

[53] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 253–260, New York, NY, USA, 2002. ACM.

[54] Harald Schoen, Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, and Peter Gloor. The Power of Prediction with Social Media. *Internet Research*, 23(5):528–543, 2013.

[55] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Compututer Survey*, 34(1):1–47, March 2002.

[56] U.S Securities and Exchange. U.s securities and exchange commission report for 2013, 2013.

[57] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Automatic summarization of twitter topics. In *National Workshop on Design and Analysis of Algorithms, Tezpur, India*, pages 121–128, 2010.

[58] C. Shekar, S. Wakade, K.J. Liszka, and Chien-Chung Chan. Mining pharmaceutical spam from twitter. In *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference*, pages 813–817, Nov 2010.

[59] Ian Soboroff, Dean McCullough, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Richard McCreadie. Evaluating real-time search over tweets. *International Conference on Weblogs and Social Media ICWSM*, pages 943–961, 2012.

[60] 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.

[61] Takeshi 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.

[62] Oren Tsur and Ari Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM International Conference on Web sSearch and Data Mining*, pages 643–652. ACM, 2012.

[63] Y. Tyshchuk, C. Hui, M. Grabowski, and W.A. Wallace. Social media and warning response impacts in extreme events: Results from a naturally occurring experiment. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 818–827, Jan 2012.

[64] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, 11:401–408, 2011.

[65] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[66] Benjamin P. Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 955–964, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[67] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Amsterdam, 3 edition, 2011.

[68] Felix Ming Fai Wong, Soumya Sen, and Mung Chiang. Why watching movie tweets won't tell the whole story? In *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, WOSN '12, pages 61–66, New York, NY, USA, 2012. ACM.

[69] Pak Chung Wong, H. Foote, D. Adams, W. Cowley, and J. Thomas. Dynamic visualization of transient data streams. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium*, pages 97–104, Oct 2003.

[70] Panpan Xu, Yingcai Wu, Enxun Wei, Tai-Quan Peng, Shixia Liu, J.J.H. Zhu, and Huamin Qu. Visual analysis of topic competition on social media. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2012–2021, Dec 2013.

[71] Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. We know what @you #tag: Does the dual role affect hashtag adoption? In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 261–270, New York, NY, USA, 2012. ACM.

[72] Xue Zhang, Hauke Fuehres, and Peter A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear?"". *Procedia - Social and Behavioral Sciences*, 26(0):55 – 62, 2011. The 2nd Collaborative Innovation Networks Conference - {COINs2010}.

[73] Arkaitz Zubiaga, Damiano Spina, Víctor Fresno, and Raquel Martínez. Classifying trending topics: A typology of conversation triggers on twitter. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2461–2464, New York, NY, USA, 2011. ACM.