CLASSIFICATION OF CLINICAL TWEETS USING APACHE MAHOUT

A THESIS IN
Computer Science

Presented to the Faculty of the University
of Missouri—Kansas City in partial fulfillment of
the requirements for the degree

MASTER OF SCIENCE

by
LI WANG

B.S. Wuhan Textile University, Wuhan, China, 2010

Kansas City, Missouri

2015

CLASSIFICATION OF CLINICAL TWEETS USING APACHE MAHOUT

Li Wang, Candidate for the Master of Science Degree

University of Missouri - Kansas City, 2015

ABSTRACT

There is an increasing amount of healthcare related data available on Twitter. Due to Twitter's popularity, every day large amount of clinical tweets are posted on this microblogging service platform. One interesting problem we face today is the classification of clinical tweets so that the classified tweets can be readily consumed by new healthcare applications. While there are several tools available to classify small datasets, the size of Twitter data demands new tools and techniques for fast and accurate classification.

Motivated by these reasons, we propose a new tool called Clinical Tweets Classifier (CTC) to enable scalable classification of clinical content on Twitter. CTC uses Apache Mahout, and in addition to keywords and hashtags in the tweets, it also leverages the SNOMED CT clinical terminology and a new tweet influence scoring scheme to construct high accuracy models for classification. CTC uses the Naïve Bayes algorithm. We trained four models based on different feature sets such as hashtags, keywords, clinical terms from SNOMED CT, and so on. We selected the training and test datasets based on the influence score of the tweets. We validated the accuracy of these models using a large number of tweets.

Our results show that using SNOMET CT terms and a training dataset with more influential tweets, yields the most accurate model for classification. We also tested the scalability of CTC using 100 million tweets in a small cluster.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Computing and Engineering have examined a thesis titled “Classification of Clinical Tweets Using Apache Mahout”, presented by Li Wang, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Praveen R. Rao, Ph.D., Committee Chair
Department of Computer Science and Electrical Engineering

Yugyung Lee, Ph.D.
Department of Computer Science and Electrical Engineering

Yongjie Zheng, Ph.D.
Department of Computer Science and Electrical Engineering

CONTENTS

ILLUSTRATIONS

Figure Page

TABLES

Table Page

ACKNOWLEDGEMENTS

I would never have been able to complete this thesis without help, support and encouragement from many people.

Firstly I would like to express my deepest appreciation to my advisor: Professor Praveen Rao, a respectable, responsible and resourceful scholar, for his invaluable guidance, patience and encouragement through my graduate study. Dr. Rao has not only provided me with enlightening instruction to my research, but also has been a true mentor for me. I also thank the National Science Foundation under Grant No. 1115871 for support.

I shall extend my thanks to my committee: Dr. Yugyung Lee and Dr. Yongjie Zheng, for their insightful advice and comments to improve my work.

Last but not least, I would like to express my sincere gratitude to my family and relatives, especially my beloved parents, for their unconditional support, encouragement and love.

CHAPTER 1

INTRODUCTION

With the increasing popularity of social network, more and more users adopt it in daily life or for business. Twitter is a one of the leading social network sites in the U.S. The microblogging service provided by Twitter attracts not only normal users, but also medical professionals like doctors, nurses and healthcare organizations. Various topics are posting on Twitter, including technology, politics, sports as well as healthcare. People use Twitter as an alternative way to publish, discuss and communicate healthcare related messages. This trend turns Twitter into a large data corpus of healthcare information. Even more, Twitter provides various useful RESTful APIs for developers to download tweet samples from Twitter server. Thus these features together make it easy to collect tweets for analysis.

When the tweets are able be collected and stored into local machine, the next step is to find an effective method to facilitate clinical analysis. A typical method towards extracting healthcare related data is to classify clinical data to related categories. Currently, there are two popular types of classifications, supervised classification and unsupervised classification. Unsupervised classification is useful when the categories are not clearly identified, while supervised classification turns out to be a good solution to classify data into well-defined categories.

Traditional approaches for supervised classification tend to process small dataset so that the classification can be done in a single machine. But when the datasets become larger and larger, classification can no longer be performed with single commodity hardware. Many traditional methods and tools then tend to fail to handle large datasets because the implemented classification system cannot scale up well. One promising solution is to adopt Hadoop MapReduce framework to enable machine learning algorithms to scale up on large-scale datasets in distributed systems.

Apache Mahout is a collection of machine learning algorithms, which are used to perform clustering, classification and recommendation. Mahout is becoming more and more popular because it is a new open source project that is able to run on top of the Hadoop framework. Many algorithms are available for classification, such as Naïve Bayes, SVM, Logistic Regression and Markov Models. Naïve Bayes classifier (NBC) uses Bayes theorem with independence assumption for classification decisions, which is widely used for supervised classification. Therefore, this Mahout/Hadoop integration is a promising approach to solve related issues of classification on large-scale dataset.

1.1 Problem Statement

With the increasing number of social media users, the data generated on social network becomes larger and larger. When it comes to the huge amount of data, new demands of efficiency, accuracy and scalability raise up in data classification. Even more, how to satisfy all these requirements is becoming more difficult. For example, some implemented algorithms need to load all the data into memory which makes it impossible to scale up to large datasets. To overcome the related problems, we propose CTC using Apache Mahout NBC and Hadoop MapReduce framework along with SNOMED CT and implemented tweet influence algorithm. To evaluate the thorough performance of this approach, several experiments were conducted to measure different aspects.

1.2 Objective

With 42 million tweets collected in the local machine using Twitter RESTful Stream API, the aim of this project is to provide an appropriate way to improve classifying large clinical tweets into related categories in distributed system for better healthcare data analysis on social network. In this thesis, we defined 6 categories for classification including brain, heart, lung, stomach, kidney and colon. Several issues need to be addressed to achieve this goal. First, the tweets were randomly collected without any filtering criteria, which means the data is not clean enough for classification. Some tweets might construct with non-English languages, which should be removed. Second, the collected tweets are not restricted to healthcare topics. A schema is required to distinguish clinical data from nonclinical data. Third, the ability to classify clinical tweets into correct categories with high accuracy is required. Last but not the least, the approach must be able to deal with classification over large-scale datasets.

1.3 Solution

To achieve the goal and address above issues, we propose CTC using Mahout Naïve Bayes algorithm and Hadoop MapReduce framework along with SNOMED CT and implemented tweet influence algorithm. Mahout NBC is extremely useful in this situation. One of the most significant advantages of Mahout NBC is that it is compatible with Hadoop MapReduce framework which means it is able to scale up to handle large-scale dataset. Currently many classification algorithms are available and for our project we decide to use Naïve Bayes algorithm, which is widely used for supervised classification.

When it comes to supervised classification, how to build the training dataset is the first concern because the classification accuracy is largely affected by the quality of the training dataset. CTC builds the training dataset by referring to SNOMED CT to improve classification accuracy. The reason we choose SNOMED CT is that it is a well-recognized comprehensive clinical healthcare terminology and each term is identified with a unique code. With this feature, it can be used as a common language communicated for healthcare professionals. For each organ, 4 most common diseases are selected from SNOMED CT. Then we use them to query against the collected tweets dataset to create clinical dataset by matching the keywords.

In addition to referring to SNOMED CT, we have implemented an algorithm to calculate the tweet influence to help improve classification accuracy. The tweet influence is measured to evaluate the capacity of one user to have an effect on others. The higher influential tweets tend to be more valuable, accurate, informative, formal and trustworthy. Instead of using existing tools or methods, we implemented an original algorithm to better fit our data. Many factors were taken into account to produce the influence value ranging from 0 to 100. After that, we associate the calculated value with each tweet and use it to rank the tweets. Then two subsets of the training dataset, top 50% influential tweets and bottom 50% influential tweets were created to compare for classification accuracy.

To examine the scalability of NBC in classifying large amount tweets in distributed system, we setup a physical Hadoop cluster including one master node and five slave nodes for experiments. Several classification MapReduce jobs with different data sizes were conducted to evaluate the scalability performance. Since we already had 42 million tweets collected in the local machine, we could easily pull a small portion to generate 1 million or 10 million tweets for classification. In order to evaluate the scalability with larger dataset, we need to replicate the whole dataset twice to generate 100 million tweets. Since there is no data cache during Map and Reduce tasks, this replication would not affect the performance evaluation.

1.4 Outline

The rest of this thesis is organized as follows. Chapter 2 introduces the background and the related work of this field. Chapter 3 presents a detailed illustration of the design and architecture of CTC. Chapter 4 provides the experiments setup, results as well as thorough evaluation of CTC. Finally, we draw the conclusion in Chapter 5. New approach to increase classification accuracy in distributed system will continue be explored. Better algorithms and technologies are desired to deal with large dataset.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Twitter

Twitter is an online social networking and microblogging service that enables users to send and read short 140-character text message, called “tweets”. Registered users can read and post tweets, but unregistered users can only read them [1]. There are various ways for users to access Twitter, like official website interface, SMS, or mobile device app.

On Twitter, users can add a friend by searching for his or her user name. After the friend relationship is generated, users are able to start to view the updates of their friends. Another relationship is called “follower”. A user becomes someone’s follower when he starts following someone on Twitter, which means the user is subscribing to updates of users that he follows. This helps users to build relationship with those who share the same interests. Friends and followers together facilitate the users interactions [2] through both one-way and mutual way relationships.

Many features are provided for users to have a better communication. The “reply” button is similar to reply function in email. By clicking “reply”, a user can respond to the tweets. Direct tweets can be sent to dedicated users by using “mention”. Simply compose a tweet containing “@” followed by the username and the tweet will be sent as a message to the mentioned user. In addition, the mentions show up as links in the tweet. Retweet is another important content-oriented interaction feature of Twitter. If a user likes the tweet and would like to share it with others, he can click the “retweet” button. After that all his followers will see the tweet in the updates. The retweeted tweet starts with RT to indicate that this is a retweeted tweet. Hashtags [3] are used to categorize tweets by keyword. People use the hashtag symbol # [4] before a relevant keyword or phrase in their tweet to categorize those Tweets and help them show more easily in Twitter Search. Clicking on a hashtagged word in any message shows you all other tweets marked with that keyword. Hashtags can occur anywhere in the tweet, at the beginning, middle, or end. Hashtagged words that become very popular are often Trending Topics.

According to Twitter statistics in 2013, there are 271 million monthly active users and 100 million daily active users [5]. 500 tweets are posted per day [5]. 29% users check Twitter multiple times a day [5]. 52 million users live in US [5]. Projected number of Twitter users by 2018 will be 400 million [5]. Currently Twitter supports more than 35 languages.

For reliability, Twitter sets some technical limits to reduce downtime and error functions. First, the maximum length of the tweet content is limited to 140 characters including the links. Second, all text should be converted to UTF-8 before sending to twitter to avoid errors. Third, 250 direct messages are allowed per day. Forth, the daily update limit is 2400 per day including both tweets and retweets. Fifth, 1000 following times per day are allowed to prohibit aggressive following behavior. Sixth, once an account is following 2000 other users, additional follow attempts are limited by account-specific ratios. Seventh, in version 1.1 of the API, an OAuth-enabled application could initiate 350 GET-based requests per hour per access token [6].

2.1.1 Twitter REST API

Representational state transfer (REST) is an abstraction of the architecture of the World Wide Web. More precisely, REST is an architectural style consisting of a coordinated set of architectural constraints applied to components, connectors, and data elements, within a distributed hypermedia system. REST ignores the details of component implementation and protocol syntax in order to focus on the roles of components, the constraints upon their interaction with other components, and their interpretation of significant data element [7].

An application programing interface (API) is a set of programing instructions and standards specifies how to access a web based software application or web tool. The REST APIs enable any interactions with HTTP, such as reading data, posting (create and update) data and deleting data. Therefore, REST implements all four CRUD (Create Retrieve Update Delete) operations by sending HTTP POST, GET, PUT and DELETE requests.

A resource is exposed via a fixed Universal Resource Identifier (URI). The consuming client of a RESTful application needs to know the persistent URI to access it. All future actions should be discoverable dynamically from hypermedia links included in the representations of the resources that are returned from that

URI. A media type description is needed to define hypermedia access and specify what methods are available for the resources of that type.

Twitter is not only a useful online social tool it also provides a comprehensive array of REST APIs. Developers can use these APIs to make applications, websites, widgets, and other projects that interact with Twitter. Current version 1.1 offers three main APIs, the normal REST API, the search API and the stream API. Each of the APIs represents one facet of Twitter.

The REST APIs constitute the core of the Twitter API. It enables developers to access and manipulate all of Twitter main data including timelines, status updates, and user profiles. Timelines on Twitter are collections of Tweets, ordered with the most recent first. In addition, users can use the APIs to generate and post tweets back to Twitter, favorite certain tweets, retrieve statuses, send direct messages, retweet certain tweets.

The search API exposes a way for users and developers to look up keywords within twitter content to filter query. It will return a collection of related tweet objects matching a given query with HTTP GET method. Additionally, hashtag query is supported. This function enables users to view tweets beyond their friends or followers. Furthermore, trending topics can be discovered with the help of search API.

Stream API offers a low-latency, high-volume and near-real time access to various subsets of public and protected Twitter data. This API is only accessible to authorized users. Three main streaming products are supported: streaming API, user streams and site streams. First, streaming API returns public tweet objects matching one or more query schemas. It also supports returning a small random subset tweet objects of all public updates. Second, user streams return a stream of data dedicated to the authenticated user, and are mainly used to update to the client. Third, site streams allow multiplexing of multiple user streams over a Site Stream connection. Only a preliminary number of calls are allowed to Twitter API.

2.1.2 Tweet Influence

As online social networking becomes more and more popular, many studies have been done to discover valuable information from it, among which social influence has drawn a lot of attention. Influence has long been studied in many fields and the findings about influence contribute a lot in advertising and marketing. On social networks, such as Twitter, the influence refers to the ability of a user to have an effect on others or the capacity to drive action. The influence can also be interpreted as the respond of one user to the activity of another user on a social network. Similarly, studying the influence on Twitter also provides new insights in social networking.

On Twitter, a small group of users who excel in spreading information is called influencers. The common characters of influential users include a larger number of audiences, more frequent updates and higher activities. In addition, the influential tweets are more likely to be retweeted than those of others. The users who have high influence tend to gain more attention than those with low influence.

Many of them are celebrities or leaders, e.g. President Obama is ranked as No.3 on Twitter [8]. He has 44,275,975 followers and created 12,164 updates. Furthermore, he is given a score of 99 out of 100 by Klout [9], which is a famous online Twitter user ranking service. Celebrities like President Obama with a tremendous number of followers can be more effective at spreading information than others. Another example is the famous photo taken during the Oscars. Ellen DeGeneres asked other actors and actresses to take a photo and upload it to Twitter. Now the photo has been retweeted over 3 million times and becomes the most retweeted photo ever. From these examples, the most influential users show their abilities to boost the rapid diffusion of opinions, promote news quickly and disseminate the popularity of political parties. In addition, studying the influence pattern can help people have a better understanding of trending flows.

How to come up with a proper approach to characterize or quantify the influence [10] on Twitter becomes an issue. Many theories have been applied to study the influence. Traditional view focuses on the influential users and regardless the role of ordinary users. In contrast, the modern theory states the users are more likely to be affected by their peers. There are both advantages and disadvantages on each theory.

Direct links [11], e.g. follower and friend relationship, represent the way information flows on social networks. Thus, in general the number of followers and friends is an important indicator of user influence. A review of MIT Technology [12] compares three different ways to spot the most influential spreaders based on the number of followers, degree, PageRank algorithm and K-core. After comparing the advantages and disadvantages of each method, the author draws the conclusion that the sum of the number followers of each direct follower of a user would be the best way to predict the most influential spreader. However this work has its own limitation. The number of followers for each direct follower must be known which does not suit every case. In addition, to predicate how widely the information would spread based on the larger number of followers and friends is biased. To get the measure [13] of the influence, many other factors should also be taken into account.

The retweet times indicate the quality of the content and the pass-along value. The more times a tweet got retweeted, the more value it will have. In addition to retweet influence, the frequency of updates is also a significant factor. The frequency of updates points whether a user is active or passive. Followers tend to lose interests in those less active. Moreover, the use of hashtag could also add value to the influence. In general, the hashtag is used to specify certain topic or keyword in a tweet. It will gain more attention compared to other words. Now hashtag becomes more and more popular due to the ease of use. Besides, the number of mentions represents the value of the user name.

Deciding the factors is just the first step towards creating an approach to measure the tweet influence. After that, the proper weight for each factor should be determined and coordinated in order to achieve a comprehensive ranking. Each defined weight indicates a different importance of the factor while composing the tweet influence.

2.2 Healthcare Information on Twitter

Within 140 characters, users can tweet whatever they like. Topics range from political opinions, comments on news events, daily life to healthcare. According to USNEWS [14], more and more medical professionals like doctors and nurses adopt social network tools like Twitter to monitor and interact with the patients. Physicians could be friends with patients online which is good for maintaining a robust relationship between them. Communication [15] via Twitter provides an alternative engagement beyond doctors' offices or hospitals. Social network has played a significant role in changing the nature and the way of health care interaction between health care organizations and consumers.

Recently, many health care organizations have established official social network accounts, for example, US Food and Drug Administration (FDA) has multiple Twitter accounts to disseminate information. Whole Foods Market also uses Twitter to reach the consumers to promote new products and answer questions. A research shows, 90% of users from 18 to 24 years indicated that the medical information shared on social network is trustworthy [14]. Besides, lots of users are reported to use Internet including social network to seek health care information. Therefore, it is important for health care related organizations and individuals to maintain public reputation [16].

Everyday large amount of health related data are transmitted on Twitter. Users post about their own health experience, reviews of treatment, medications, hospitals or doctors, and symptoms. They seek for help as well as related tips, photos and videos. Sometimes patients may find out that they receive the same advice from doctors as from social network.

Effective cost and wide reach have made Twitter a new platform for health care information exchange. With huge amount of health care related data generate everyday, Twitter evolves into a potential source pool for health care. It can be used as a complementary source in addition to formal health care data. One obvious advantage is that the tweets are real time and more relevant to current trending topics. Some studies have been conducted to utilize Twitter data along with Epidemic Intelligence (EI) to analyze potential diseases outbreak [17]. A pilot study is gathering tweets including keywords relating to “flu” to analyze the trending disease activity.

Individual health related tweet might only provide limited informative value. However the aggregation of millions of tweet is large enough to provide some insights [18]. Moreover, the tweets are not separated events. Due to the created time and geography, many tweets are related to the same topic.

2.3 Data Mining and Knowledge Discovery

In general, data mining or knowledge discovery is a powerful new technology which refers to the process of extracting or discovering hidden insights

[19] and meaning from numerous sets of data beyond simple analysis [20]. In contrast to traditional statistical methods, complicated mathematical algorithms are applied to rapidly discover the patterns in data corpus and predict the probability of occurrences in the future. It provides powerful abilities for users to predict trends, analyze behaviors, make knowledge driven decisions and cluster large amount of data. Nowadays, data mining is applicable to many fields, such as marketing, finance, communication as well as social networks.

One important prerequisite for data mining is massive data collection. With advancements in the capacity of storage [21], it becomes easier to store data in either distributed or centralized data storage. Data warehouse is a new technology, which aims to store, maintain and retrieve data. It has played a significant role in data mining for its ability of maximizing the efficiency in data accessing and analysis. A wide range of companies have deployed and maintained large data warehouses for data mining.

Data mining involves many knowledge and techniques. Among those modeling is the key to the process of data mining. Modeling refers to the act of building a model by applying certain algorithms to a specific dataset [22]. After that the model can be applied to new dataset in another situation for automatic discovery or trend prediction. Data mining is becoming increasingly popular because it helps to providing valuable insights to the data and it can be applied to various fields.

2.4 Apache Hadoop

Apache Hadoop [23] is an open source distributed framework for large scale batch processing on either standalone machine or across clusters commodity servers. It is licensed under Apache Software Foundation. Nowadays, with the dropping of hardware cost and advance in storage capacity and dramatic increasing in data size, there is a clear need for efficiently processing such huge amount of data [24]. Under this circumstance, Hadoop with the ability to horizontally scale to large datasets on thousands of commodity nodes is widely adopted by both industry and academic [25].

The Hadoop framework is implemented with Java. Thus it can be easily deployed across Windows, Unix and OS X operation machines. Currently, in additional to Java API, Python, C++, Ruby and PHP, other APIs are available for use. The programmers are free to choose the most suitable language to work with.

The MapReduce is the core of Hadoop program for parallel processing. It is this paradigm that enables large scale distribution across hundreds or thousands of commodity nodes within a cluster [26]. A MapReduce job actually refers to two distinct user defined functions. The first one is map job, which is in charge of taking input dataset and transforming to a set of intermediate key-value pairs. Then, the reduce job is applied to the output of map job to aggregate into a smaller set of tuples. The process follows a sequence that map job is always conducted before the reduce job. Hadoop MapReduce framework is shown in Figure 1.

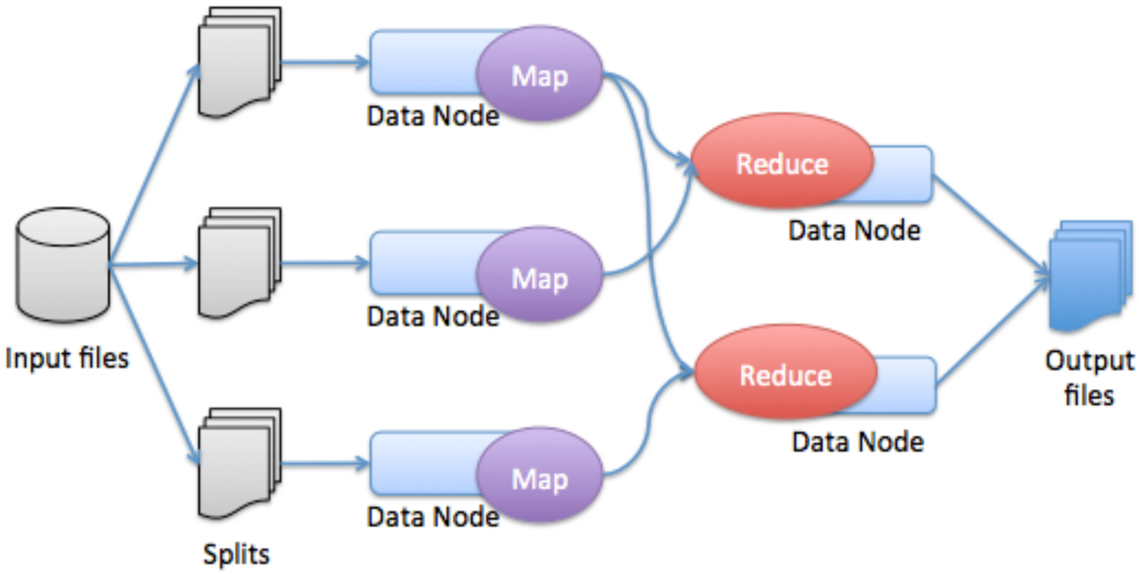


Figure 1 Hadoop MapReduce framework

2.5 Machine Learning and Apache Mahout

Machine learning is a sub-discipline of artificial intelligence which refers to the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases [27]. The major goal of machine learning is to apply learnt experience to new data. Machine Learning integrates many distinct fields such as data mining, probability theory, logic, combinatorial optimization, statistics, control theory, reinforcement learning and statistics [28].

Apache Mahout [29], which is written in Java, is an open source machine learning library built on top of Hadoop. It mainly integrates many algorithms to implements three use cases: collaborative filtering, classification and clustering [30]. Collaborative filtering is about referring recommendations based on user information, such as reviews, clicks and ratings. There are types of recommendations. One is user based, which means users who share similar tastes will be grouped together. Another is item based, in which similar items will be identified and classified. Clustering is unsupervised learning which targets to group a number of things that share the same similarity [31]. It is able to organize large number of data without requiring prior knowledge about the classification [32]. In contrary, classification that is supervised learning which requires the model to be trained before applying it to classify new instances. Both of them are useful and fit in different situations.

2.6 SNOMED CT

SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) is comprehensive clinical terminology which provides terms, synonyms, codes and definitions. The aim is to standardize the presentation of terms used in health information by codes. According to Wikipedia, "SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world." It mainly composes of concept codes, descriptions and relationships. Each term is well described and identified by a unique code. Besides, the terminologies are well structured according to logic-based representation of meanings. It provides an efficient way to index, store, retrieve and aggregate clinical data from organized, computerized health records. SNOMED-CT enables people to communicate in a common language, thus facilitate the quality of patient health data transmission across different healthcare providers. Besides, it helps record the patient clinical data in the electronic medical records. As the standard terminology is used across industries and hospitals, not only the transferring will be simplified but also the accuracy will be improved. The well-structured hierarchy navigation makes querying among related terms much easier. Now it is recognized as a useful resource in health care analysis. An example of IHTSDO SNOMED browser diagram for influenza is shown in Figure 2.

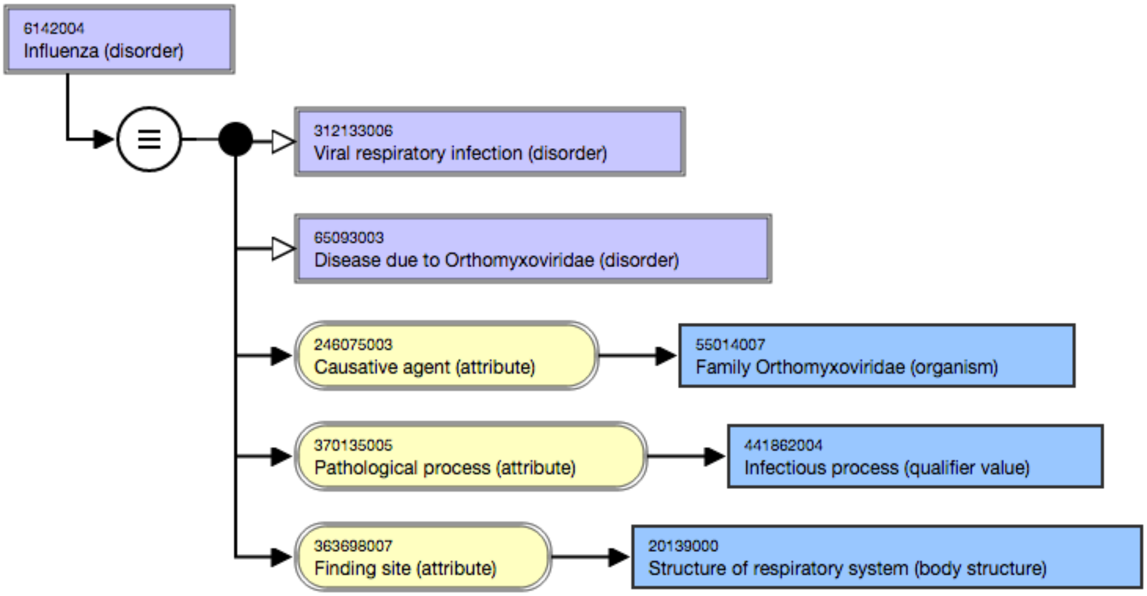

6142004
Influenza (disorder)
≡
312133006
Viral respiratory infection (disorder)
65093003
Disease due to Orthomyxoviridae (disorder)
246075003
Causative agent (attribute)
55014007
Family Orthomyxoviridae (organism)
370135005
Pathological process (attribute)
441862004
Infectious process (qualifier value)
363698007
Finding site (attribute)
20139000
Structure of respiratory system (body structure)

Figure 2 IHTSDO SNOMED CT browser diagram for influenza

2.7 Related Work

Klout [9] is a very popular online ranking service that measures social network users influence, for example, Twitter users influence. Each social network user has influence and Klout gives him a score ranging from 1 to 100 to represent the influence by measuring relevant data. More influential users will gain higher scores. Currently, Klout support many platforms user influence measuring, such as Facebook, Twitter and LinkIn. Many data are used to calculate the influence, such as follower count, friend count, retweet count. The score reflects the quality and volume of a user social activity. However, the exact calculation algorithm remains a secret. Although Klout is already available online, it is still helpful to implement the algorithm for the influence to achieve fine-grain control of the analysis procedure.

The paper [33] Twitter proposes an approach to analyze user influence in three perspectives, namely indegree measure, retweet measure and mention measure over topic and time. Each measure indicates a specific view of ranking users. Indegree displays the popularity of a specific user on Twitter. Retweets measure driven by content is strongly context-oriented. Mentions measure corresponds to the value of user name. The specific paper concluded that most users with high influence could have important influence over various topics. In addition, it is shown that influence is achieved through consistent passive social activities, but not accidently or spontaneously.

Kathy et al. [34] introduced a classification system to classify Twitter trending topics into generic categories to facilitate users to have better query experience with trending topics. Two approaches are presented to address the issue, text classification and network-based classification. 768 manually labeled topics with varying number of tweets are classified into 18 well-defined classes.

Different algorithms, like Naïve Bayes, Support Vector Machine and ZeroR are used in the classification to get the one with he best accuracy. From the results, Naïve Bayes Multinomial turns out to have the highest accuracy for the text classification. In addition, they provide the comparison of classification accuracy with several different algorithms in network-based classification. The paper declares that network-based classification performs better in terms of classification accuracy.

In [35], the Ailment Topic Aspect Model (ATAM) is presented to discover public health topics from Twitter. The author has created a dataset of 11.7 million health related tweets for data mining by keywords filtering. Next, a corpus of 5128 labeled messages are created as training data for a supervised classification. After the SVM getting trained, it is used to classify the tweets into health relevant dataset and not relevant dataset. As a result, 1.63 million health related tweets are generated. Then the model ATAM is demonstrated to classify the health related tweets into many different topics as well as to group symptoms and treatments into related ailments. In the comparison with standard LDA model, ATAM is able to produce more identifiable ailments with higher coherence.

An application of Mahout Naïve Bayes classifier (NBC) to mine sentiment or opinion from massive dataset is presented in paper [30]. The authors implement a complete and simple system with integration of Hadoop framework to evaluate the scalability of NB classifier. Instead of using standard Mahout library, an implemented NBC for Hadoop program is used to evaluate the scalability over different data size. The virtual Hadoop cluster is set up in cloud. The experiment results are analyzed from three aspects: computation time, classification accuracy and the throughput of the system. The paper declares that the increase in the size of data would lead to improvement of classification accuracy. In addition, the results highlight that NBC is able to easily scale up no matter database exists or not. Instead of using virtual Hadoop cluster proposed in that paper, we measured the classification in physical Hadoop cluster which demonstrated more reliable results. The comparison between previous studies and this work is shown in Table 1.

Table 1 Comparison between previous studies and this work

	Ref. 34	Ref. 35	Ref. 30	This work
Classification goal	Twitter topic	Health tweet	Sentiment review	Clinical tweet
Algorithm	NBM, SVM, NB, ZeroR	SVM	NB	NB
Label	18 topic lists	Related/unrelated	Positive/negative	6 clinical categories
Data size	9000	11.7 Million	1 Million	100 Million
Evaluation metric	Different classifiers	Different classifiers	Different # of data	Different # of data
Classification accuracy	70.96%	Not available	82%	84%

CHAPTER 3

DESIGN AND FRAMEWORK

3.1 CTC Framework

We designed and implemented CTC on top of Hadoop MapReduce and HDFS to enable scalable classification of clinical content on Twitter. The main design requirement for CTC is to construct highly accurate models as well as enable efficient classification on large-scale datasets. CTC consists of four components, data collector, data parser, Apache Mahout NBC and Hadoop MapReduce framework. CTC utilizes Twitter Streaming API to collect tweet samples into local machine from Twitter server. To parse the downloaded tweets and construct high accuracy models, CTC leverages the SNOMED CT and a new user influence scoring schema. Apache Mahout NBC is used as the core classifier to enable clinical tweets classification. Hadoop MapReduce framework provides the ability to perform classification jobs on large-scale datasets in a deployed cluster. Figure 3 illustrates the framework of CTC and how it operates. The overall workflow is illustrated as follows:

1) Model training: When the training dataset has been prepared and is ready in HDFS, the first step is to start the training job to build a model. After the model has been successfully trained, several files will be generated in the output directory, like model files, tfIdf vectors and dictionary files which are needed to perform the classification jobs in the future. Since the training dataset is usually small, this training job is performed on a single machine.

2) Model testing: Once the model has been trained with training dataset, the next step is to evaluate its performance using the testing dataset. The label assigned to each tweet by the model will be compared to the associated correct label and the results for all tweets will be presented. The percentage of correctly classified instances and a matrix will be generated for validate the model.

3) Applying trained model: Finally, the validated model can be applied to classify new data. It reads each line and simultaneously computes the probability of each tweet for all the categories. Then the label with the highest relevance score will be assigned to the tweet. When the all the classification jobs complete, CTC will write the final results to local file system. The classification jobs can be conducted either on a single machine or in the cluster based on the data size.

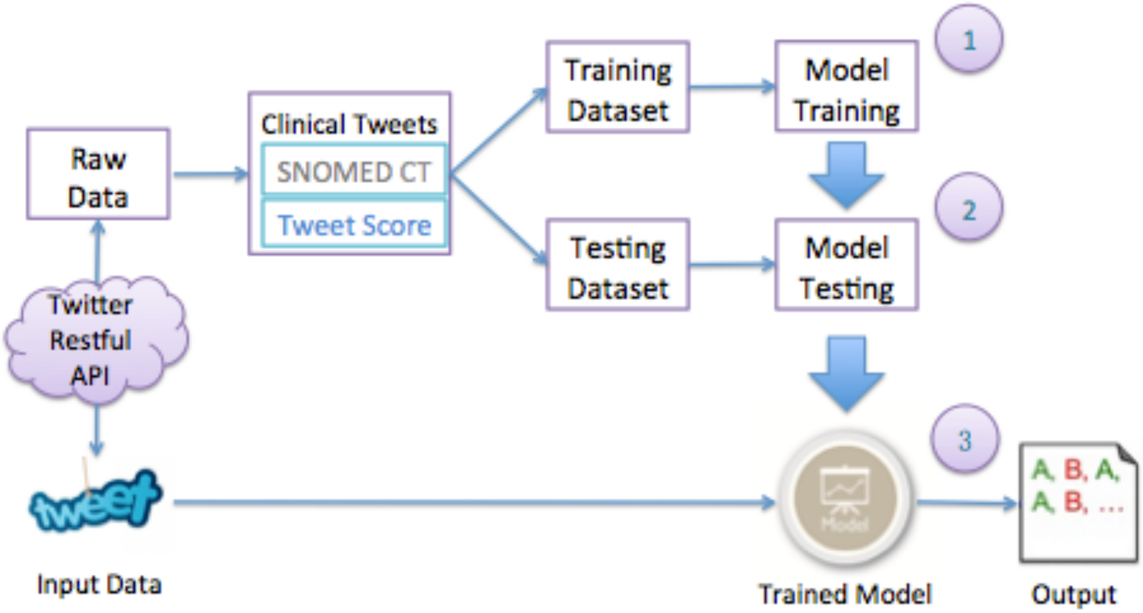


Figure 3 Overview of the framework

An abstract CTC classification representation is shown in Figure 4.

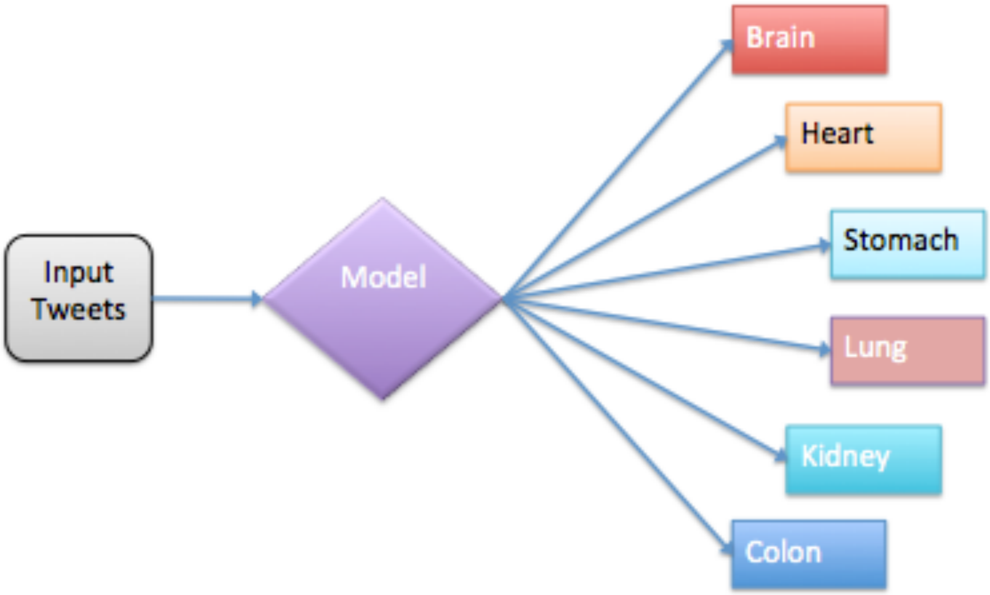


Figure 4 Abstract CTC classification

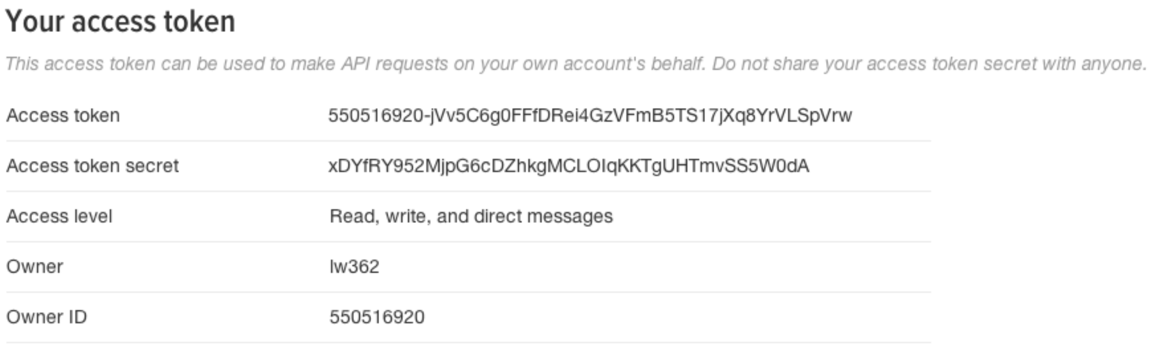
3.2 Twitter Data Collection

Apache Maven [36] is Java-based project build and manage tool that is developed and hosted by Apache Software Foundation. The main goal of Maven is to not only describe the building of software, but also to mange the dependencies. It is different from Apache Ant because of the use of “convention over configuration”. A Project Object Model (POM) is the description file in Maven. It is an XML file that contains information about how the projects are built and configured. In addition, because of its simplification and standardization, it has been widely used by industry and academy for project development.

Twitter4J [37] is a Java library for Twitter APIs. It is a useful tool which can be integrated to develop java application using Twitter APIs. Currently, it supports many functions, for example authentication with OAuth and streaming. Moreover, it has Maven dependencies available.

A Java application was developed using Twitter4J dependencies to collect public sample tweets.

First, to grant the application to access a Twitter account, the application must be registered at official Twitter developer web site [38]. After successful registration, generate access token for application authentication with OAuth. Access token used in this application is shown in Figure 5.



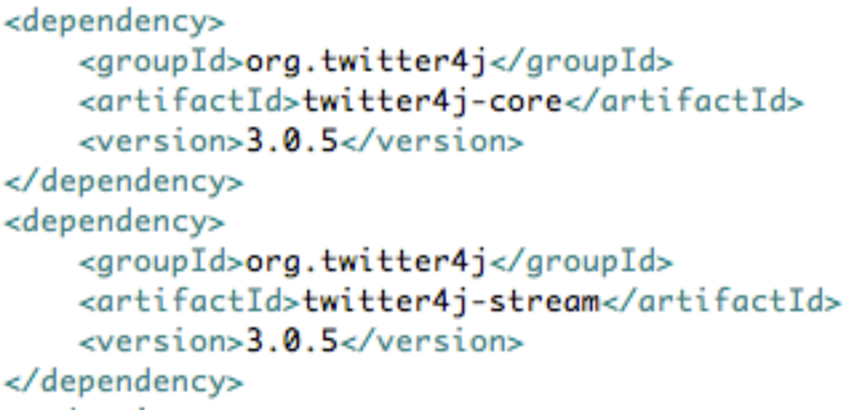
Your access token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access token	550516920-jVv5C6g0FFfDRei4GzVFmB5TS17jXq8YrVLSpVrw
Access token secret	xDYfRY952MjpG6cDZhkgMCLOIqKKTgUHTmvSS5W0dA
Access level	Read, write, and direct messages
Owner	lw362
Owner ID	550516920

Figure 5 Access token for application authentication

Second, after OAuth authentication with generated access token, implement StatusListener class in Twitter4J to use Twitter streaming API to consume the public sample statuses streaming. Tweets with all topics were collected, not limited to any specific topics, keywords or queries. Twitter4J maven dependencies used in the application are shown in Figure 6.



```
<dependency>
    <groupId>org.twitter4j</groupId>
    <artifactId>twitter4j-core</artifactId>
    <version>3.0.5</version>
</dependency>
<dependency>
    <groupId>org.twitter4j</groupId>
    <artifactId>twitter4j-stream</artifactId>
    <version>3.0.5</version>
</dependency>
```

Figure 6 Twiter4J maven dependencies

The data collected since September 2012 is shown in Table 2

Table 2 Collected data description

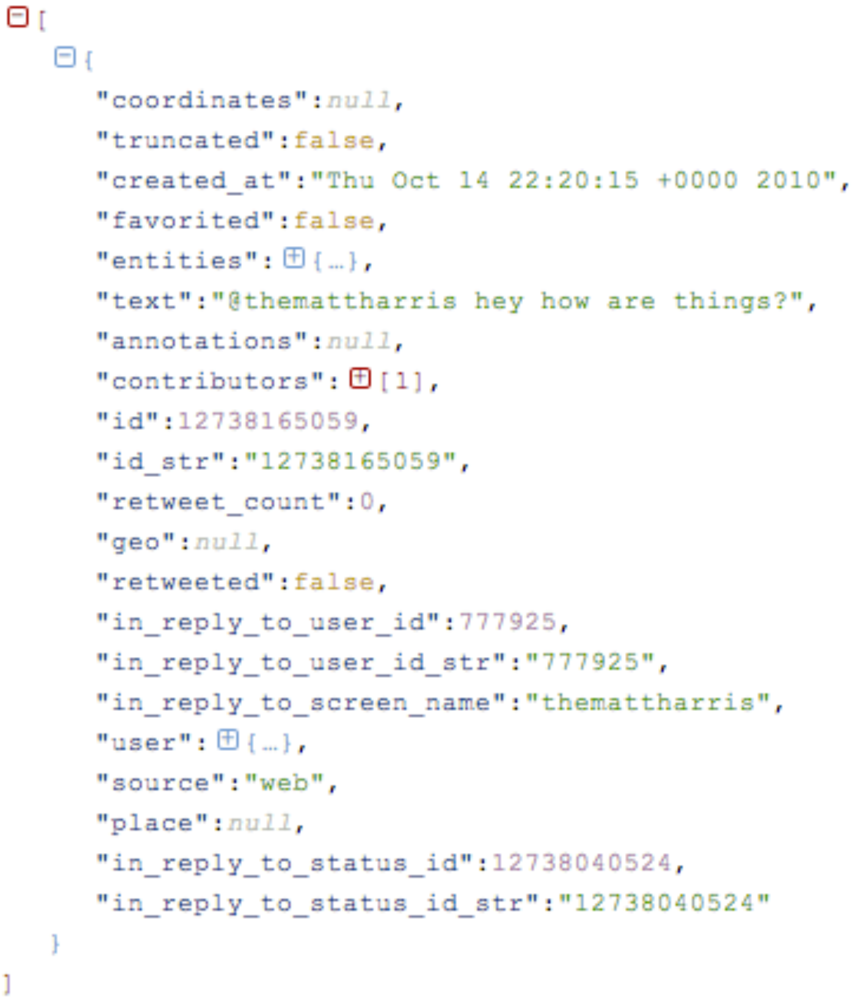
Month	File Size (GB)
September 2012	12
October 2012	23.7
November 2012	22.7
December 2012	23.2
January 2013	26.8
February 2013	23.6
September 2013	14
October 2013	23
February 2014	56
March 2014	62

Month	File Size (GB)
April 2013	57

3.3 Twitter Data Parsing

The data structure that Twitter uses for storing tweet information is JSON format. JSON [39] (JavaScript Object Notation) is an open standard format that stores human readable text organized, easy-to-access key-value pairs. Due to its language-independent and ease of generating and parsing, JSON now is widely adopted for data storing and transmitting.

The complete information of a tweet is stored in a JSON object. Within the object, many nested JSON objects and arrays are used to represent the data about a tweet, for example Twitter ID, create time, user name, language, content, follower count, friend count, list count, retweet count, hashtag, geo-location and mention. A sample JSON tweet is parsed and represented with an online JSON parser [40] in Figure 7.



```
⊟[
   ⊟{
      "coordinates":null,
      "truncated":false,
      "created_at":"Thu Oct 14 22:20:15 +0000 2010",
      "favorited":false,
      "entities":⊞{…},
      "text":"@themattharris hey how are things?",
      "annotations":null,
      "contributors":⊞[1],
      "id":12738165059,
      "id_str":"12738165059",
      "retweet_count":0,
      "geo":null,
      "retweeted":false,
      "in_reply_to_user_id":777925,
      "in_reply_to_user_id_str":"777925",
      "in_reply_to_screen_name":"themattharris",
      "user":⊞{…},
      "source":"web",
      "place":null,
      "in_reply_to_status_id":12738040524,
      "in_reply_to_status_id_str":"12738040524"
   }
]
```

Figure 7 Sample tweet in JSON format

We create a Java program to parse and retrieve several important fields from the JSON format tweet. The first job is to filter the tweets by language. We only analyze tweets in English. Tweets in other languages are all filtered out. Next, tweets including unrecognized characters are removed from the dataset. Finally, extract the metadata needed from the JSON file, as shown in Table 3.

Table 3 Extracted fields from tweets in JSON format

Fields Name in JSON	Extracted Fields
name	User name
create_at	Create time
text	Tweet content
hashtags	Number of hashtags
retweet_count	Number of retweets
status_count	Number of statuses
follower_count	Number of followers
friend_count	Number of Friends
list_count	Number of Lists

After cleaning and parsing, the data are presented in 11 files. Each of them represents a one-month tweets collection, as shown in Table 4.

Table 4 Parsed data description

File	Row Number (MM)	File Size (MB)
1	2.3	155
2	4	281
3	3.6	257
4	3.8	264

File	Row Number (MM)	File Size (MB)
5	4.3	300
6	3.8	258
7	1.7	113
8	2.9	200
9	5.2	353
10	5.4	364
11	4.9	332

Some of the extracted data are used to calculate user influence which will be discussed in the next section.

3.4 Tweet Influence Algorithm

The influence is a comprehensive evaluation on user activities on social network. It can be viewed as the reputation of the user. Information from users with high influence could be spread widely and quickly over the social network due to their impact and connections. Relatively, the influential tweets would be considered more reliable and trustworthy. In addition, studies indicate that Twitter is rather a content-oriented information platform than an individual conversation platform. In this section, we present an algorithm to calculate tweet influence score in the range 0 to 100, shown as the equation below.

$$Influence(T) = V_{\mathrm{h}}\times W_{\mathrm{h}} + V_{\mathrm{r}}\times W_{\mathrm{r}} + V_{\mathrm{s}}\times W_{\mathrm{s}} + V_{\mathrm{f}}\times W_{\mathrm{f}} + V_{\mathrm{d}}\times W_{\mathrm{d}} + V_{\mathrm{l}}\times W_{\mathrm{l}} \qquad (3-1)$$

Where

V_h = Value of hashtag field; W_h = Weight of hashtag field

V_r = Value of retweet field; W_r = Weight of retweet field

V_s = Value of status field; W_s = Weight of status field

V_f = Value of follower field; W_f = Weight of follower field

V_d = Value of friend field; W_d = Weight of friend field

V_l = Value of list field; W_l = Weight of list field

Each field is illustrated in Table 5.

Table 5 Explanation of fields in the algorithm

Fields	Explanation
Hashtag	Indicate the topic and keyword
Retweet	Indicate the content-oriented quality
Status	Indicate the frequency of updates
Follower	Indicate the volume of information receiver
Friend	Indicate the volume of information receiver
List	Indicate the popularity of the account

The above six fields are used to form the tweet influence algorithm. The most important attributes contribute to the influence are the number of followers and friends. Each of them weights as 0.25 as shown in Table 6. Other fields' weights are assigned due to their correlations to tweet influence.

Table 6 Weight for each field

Field	Hashtag	Retweet	Status	Follower	Friend	List
Weight	0.1	0.5	0.2	0.25	0.25	0.15

Calculating the value for each field becomes an issue because of that value varies in large range, for example, one user may have 25 followers whereas another user may have more than 1 million followers. It is difficult to form an equation to calculate and assure the result to be in the range from 0 to 100. To overcome this issue, we introduce an approach that is relative ranking in several intervals. We use retweet value calculation for illustration. First we count the percentage of the number of retweet in five intervals, as shown in Table 7.

Table 7 Percentage of the number of retweets in different intervals

Interval	[0,10]	(10, 100]	(100, 1k]	(1k, 10k]	(10k, ∞)
Percent	92.12%	3%	3.37%	1.28%	0.23%

Then we form an equation to calculate the retweet value based on this range statistics. See equation 3 - 2 below, n is the number of retweets.

$$V_{\mathrm{r}} = \begin{cases} (n * 92.12)/10, & 0 \le n \le 10 \\ 92.12 + (n - 10) * 3/90, & 10 < n \le 100 \\ 95.1 + (n - 100) * 3.37/900, & 100 < n \le 1000 \\ 98.47 + (n - 1000) * 1.28/9000, & 1000 < n \le 10000 \\ 100, & 10000 < n \end{cases} \tag{3 - 2}$$

92.12% of the data are centralized in the interval 0 to 10. To keep the calculation method consistent in all the fields, we do not apply other methods to this situation. Similarly, we count the percentage and form the equation for status (3 – 3), follower (3 – 4), friend (3 – 5) and list (3 – 6).

Table 8 Percentage of the number of statuses in different intervals

Interval	[0,10]	(10, 100]	(100, 1k]	(1k, 10k]	(10k, 100k)	(100k, ∞)
Percentage	1.39%	4.28%	15.54%	45.31%	32.52%	0.96%

$$V_{\mathrm{s}} = \begin{cases} (n * 1.39)/10, & 0 \le n \le 10 \\ 1.39 + (n - 10) * 4.28/90, & 10 < n \le 100 \\ 5.67 + (n - 100) * 15.54/900, & 100 < n \le 1000 \\ 21.21 + (n - 1000) * 1.28/9000, & 1000 < n \le 10000 \\ 66.52 + (n - 10000) * 32.52/90000, & 10000 < n \le 100000 \\ 100, & 100000 < n \end{cases} \tag{3 - 3}$$

Table 9 Percentage of the number of followers in different intervals

Interval	[0,10]	(10, 100]	(100, 1k]	(1k, 10k]	(10k, 100k)	(100k, ∞)
Percentage	4.35%	18.51%	64.5%	11.4%	1.1%	0.14%

$$V_{\mathrm{f}} = \begin{cases} (n * 4.35)/10, & 0 \leq n \leq 10 \\ 4.35 + (n - 10) * 18.51/90, & 10 < n \leq 100 \\ 22.86 + (n - 100) * 64.5/900, & 100 < n \leq 1000 \\ 87.36 + (n - 1000) * 11.4/9000, & 1000 < n \leq 10000 \\ 98.76 + (n - 10000) * 1.1/90000, & 10000 < n \leq 100000 \\ 100, & 100000 < n \end{cases} \tag{3 - 4}$$

Table 10 Percentage of the number of friends in different intervals

Interval	[0,10]	(10, 100]	(100, 1k]	(1k, 10k]	(10k, ∞)
Percentage	3.24%	15.31%	71.85%	9.26%	0.34%

$$V_{\mathrm{f}} = \begin{cases} (n * 3.24)/10, & 0 \leq n \leq 10 \\ 3.24 + (n - 10) * 15.31/90, & 10 < n \leq 100 \\ 18.55 + (n - 100) * 71.85/900, & 100 < n \leq 1000 \\ 90.4 + (n - 1000) * 9.26/9000, & 1000 < n \leq 10000 \\ 100, & 10000 < n \end{cases} \tag{3 - 5}$$

Table 11 Percentage of the number of lists in different intervals

Interval	[0,10]	(10, 100]	(100, 1k]	(1k, ∞]
Percentage	90.58%	7.63%	1.62%	0.17%

$$V_{\mathrm{l}} = \begin{cases} (n * 90.58)/10, & 0 \leq n \leq 10 \\ 90.58 + (n - 10) * 7.63/90, & 10 < n \leq 100 \\ 98.21 + (n - 100) * 1.62/900, & 100 < n \leq 1000 \\ 100, & 1000 < n \end{cases} \tag{3 - 6}$$

The number of hashtags is much smaller compared to the number of status or follower. For hashtag, we do not apply relative ranking approach. Just assign 25 for each distinctive use of hashtag. In addition, we limit the maximum value to 100. The equation to calculate hashtag value is shown in 3 – 7.

$$V_{\mathrm{h}} = \begin{cases} n * 25, & n < 4 \\ 100, & n \geq 4 \end{cases} \qquad (3-7)$$

3.5 Extract Clinical Data with Reference to SNOMED CT

We implemented a healthcare data analysis system to measure classification accuracy and evaluate the scalability using Mahout Naïve Bayes algorithm on top of Hadoop MapReduce framework as shown in Figure 4.

We introduce a new approach to extract clinical data by referring to SNOMED CT which is the most recognized clinical healthcare terminology. It maintains organized, identified and described clinical terms with unique code. Using the terms from SNOMED CT, we can benefit the communication in healthcare. In this project, we choose 4 most common diseases for each related organ by referring to SNOMED CT. The associated SNOMED ID is used to identify each disease, as shown in Table 12.

Table 12 Organs with related diseases and SNOMED ID

Organ	Disease	SNOMED ID
Brain	meningitis	7180009
	brain tumor	254941009
	stroke	25133001
	epilepsy	84757009
Heart	cardiovascular injury	282728007
	coronary disease	53741008
	myocardial infarction	22298006
	atherosclerosis	38716007
Stomach	gastric ulcer	397825006
	gastritis	4556007
	gastric cancer	276809004
	gastric polyp	78809005
Lung	pneumonia	233604007
	influenza	6142004
	asthma	195967001
	bronchitis	32398004
Kidney	nephritis	52845002
	renal failure	14669001
	nephrotic syndrome	52254009

Organ	Disease	SNOMED ID
	renal stone	62315008
Colon	appendicitis	74400008
	enteritis	78420004
	constipation	14760008
	diarrhea	62315008

CHAPTER 4

EVALUATION

4.1 Datasets

To classify healthcare related tweets into related categories, preparing training dataset is the first and most important step towards this classification because the classification accuracy depends on the quality of the training dataset. We have already collected more than 40 million tweets into local machine from Twitter using Stream API since 2012. But this data corpus was randomly collected without specifying particular topic. Various topics were covered in this data corpus. Thus how to retrieve clinical tweets from the data corpus becomes a problem. To address this issue, we used different features to extract clinical data. The features included hashtag organ, hashtag disease, keyword organ and keyword disease as shown in Table 13.

Table 13 Features to extract clinical data

Features	Method	Description
H_O	Match	Return tweets exact match with organ hashtag
$H_O + H_D$	Match	Return tweets exact match with organ or disease hashtag
$H_O + W_O$	Match	Return tweets exact match with organ keyword or hashtag

Features	Method	Description
$H_O + W_O + H_D + W_D$	Match	Return tweets exact match organ or disease keyword or hashtag

The clinical data were extracted by matching one or multiple features. Each training dataset was stored in a separate file in which clinical tweets were separated by line. The training datasets were named D1, D2 D3 and D4, as illustrated in Table 14.

Table 14 Training datasets and information

Dataset	Number of Tweets	Feature
D1	684	H_O
D2	1433	$H_O + H_D$
D3	132742	$H_O + W_O$
D4	141684	$H_O + W_O + H_D + W_D$

4.2 Workloads

In order to evaluate the classification performance of classifier models built by different training datasets, we used three workloads, namely W1, W2 and W3. Experiments in W1 and W2 were conducted on single node because the aim was to evaluate training dataset. Experiments in W3 were conducted in Hadoop cluster to explore the scalability.

Workload W1 is shown in Table 15 and used to measure the classification accuracy of different models with different training dataset sizes. To perform a reliable examination, we used different percentage combinations of actual training dataset and testing dataset to perform classification. Mahout supports random dividing training dataset to actual training dataset and testing dataset. For the experiments, we used three types of combination: 90%, 70%, 50% and 30% for training dataset and 10%, 30%, 50% and 70% for testing dataset. The models trained by different training datasets are illustrated in W1.

Table 15 Workload W1

Classification	Model	Training set percentage	Testing set percentage
C1	M1	90%	10%
C2	M1	70%	30%
C3	M1	50%	50%
C4	M1	30%	70%
C5	M2	90%	10%
C6	M2	70%	30%
C7	M2	50%	50%
C8	M2	30%	70%
C9	M3	90%	10%
C10	M3	70%	30%
C11	M3	50%	50%
C12	M3	30%	70%
C13	M4	90%	10%

Classification	Model	Training set percentage	Testing set percentage
C14	M4	70%	30%
C15	M4	50%	50%
C16	M4	30%	70%

Workload W2 is shown in Table 16 and used to compare the classification accuracy between classifiers trained with higher influential tweets and classifier trained with lower influential tweets.

First, for each training dataset, we randomly chose a small amount of data and use them as testing dataset. Next, we ranked the remaining tweets by implemented tweet influence algorithm. Then we divided the training dataset into two subsets with equal data size. One dataset contained top 50% influential tweets. Another one contained bottom 50% influential tweets. Finally, we used these two datasets to train the model separately and measured the classification accuracy against the same testing dataset.

Table 16 Workload W2

Classification	Model	Training set	Testing set
C17	M1	Top 50% influential tweets	80
C18		Bottom 50% influential tweets	
C19	M2	Top 50% influential tweets	150
C20		Bottom 50% influential tweets	

Classification	Model	Training set	Testing set
C21	M3	Top 50% influential tweets	13000
C22		Bottom 50% influential tweets	
C23	M4	Top 50% influential tweets	14000
C24		Bottom 50% influential tweets	

Workload W3 is used to measure the scalability of Mahout Naive Bayes algorithm in distributed system, Hadoop cluster. We established a physical Hadoop cluster in our lab with 6 nodes, 1 Master node and 5 Slave node. The nodes were connected via one router and one switch. With this network configuration, all the traffic went through the inner network. The Hadoop cluster configuration is shown in Table 17.

Table 17 Hadoop cluster configuration

Node	OS	Memory size	Hard drive size	Hadoop version
Master	Ubuntu 12.04	4 GB	350 GB	1.2.1
Slave	Ubuntu 12.04	2 GB	80 GB	1.2.1

We started all Hadoop daemons from Master node which would invoke Task Tracker and DataNode daemons on Slave nodes. The training dataset what was used to build the model and the data to be classified were all stored on Hadoop HDFS. Once starting the classification jobs, the MapReduce framework split the classifying dataset into multiple chunks based on the block size and dispatched the map and reduce tasks to each Slave node. The default block size (64 MB) was applied to all the experiments. We used both Model 3 and Model 4 to evaluate the scalability with different to-be-classified data sizes. Workload W3 is shown in Table 18. The data sizes and the number of map reduce tasks with respect to the number of tweets are shown in Table 19.

Table 18 Workload W3

Classification	Model	# of tweets to be classified
C25	M3	1 Million
C26	M3	10 Million
C27	M3	100 Million
C28	M4	1 Million
C29	M4	10 Million
C30	M4	100 Million

Table 19 MapReduce jobs

# of tweet	Data size	# of map task	# of reduce task
1 million	82.8 MB	2	1
10 million	842 MB	13	1
100 million	8.4 GB	126	1

4.3 Performance Results

The results for workload W1 are presented in Figure 8. The average classification accuracy is measured for each model with different ratios of training dataset size to testing dataset size. Each accuracy number was the average of three trials. As we can see that the classification accuracy increased as the number of applied features increased. The model trained with disease features referring to SNOMED CT turned out to have the highest accuracy as we expected. For each model, the classification accuracy increased as the ratio of training dataset size to testing dataset size increased.

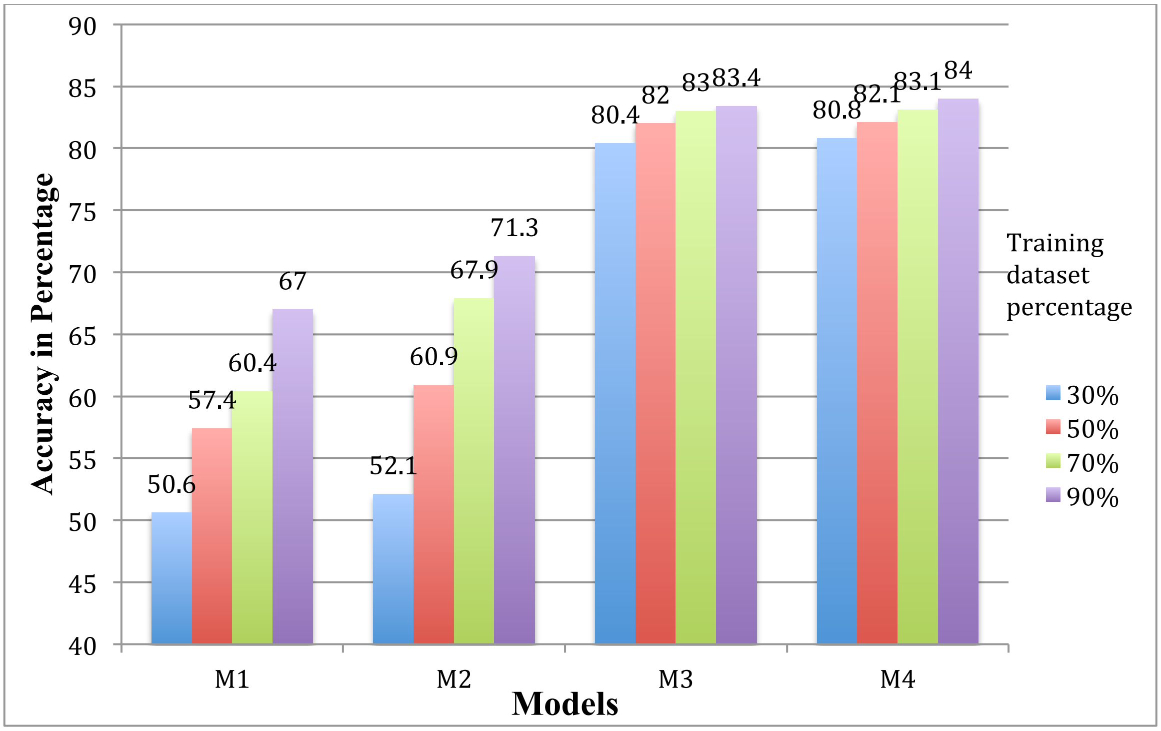


Figure 8 The accuracy for each model in workload W1

The average time to complete the above classifications for each model is measured as shown in Figure 9. This included time to upload training set, transform, split, classify and test the model. As we can observe that, the consumed time increased as the training dataset grew. But the time was not significantly affected by the data size. The time taken to complete each classification was under 10 minutes.

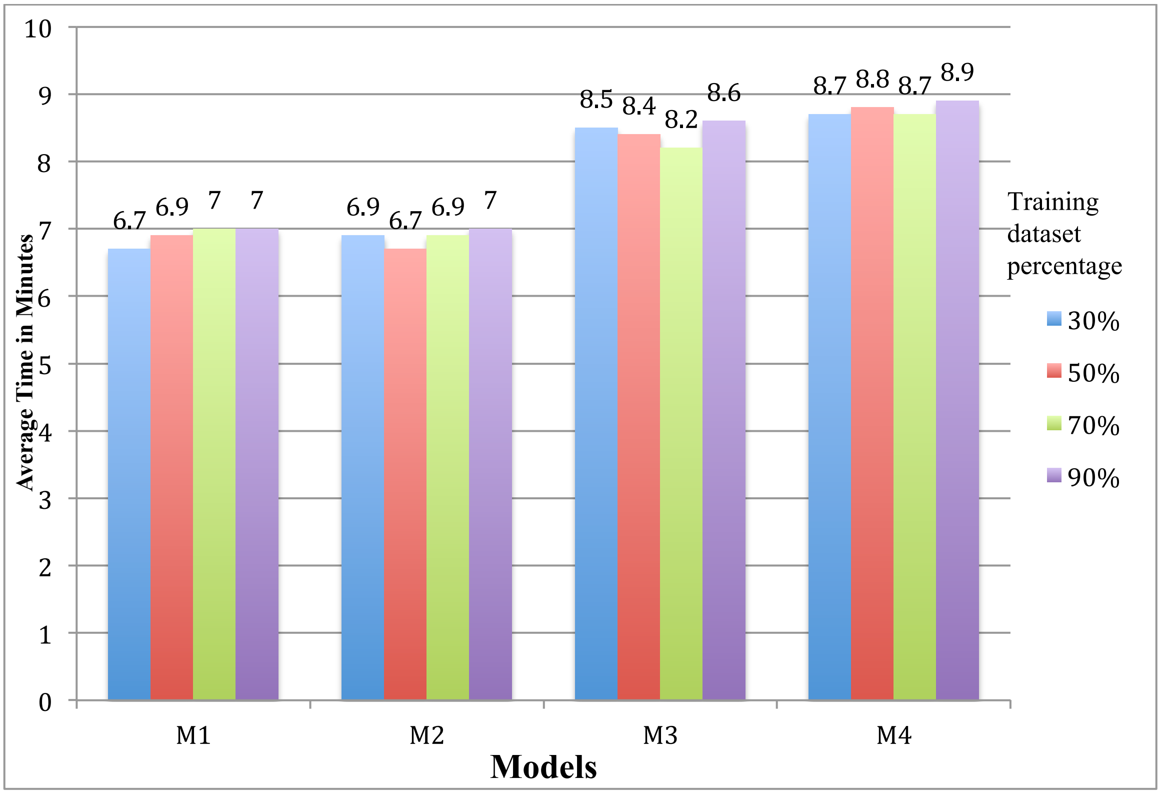


Figure 9 Time for classification in workload W1

Next, we present a classification comparison between classifier with top 50% influential tweets and the one with bottom 50% influential tweets for each model. Also the accuracy number is the average of three trials. For each model, both top 50% and bottom 50% influential training datasets were tested on the same testing dataset. From each pair of top/bottom classifications, we can see that the classifier trained with top 50% influential dataset always produced higher accuracy than the one trained with bottom 50% as we expected. The largest difference was 6% in Model 1. Clearly, this demonstrated that the training dataset created from with higher influential tweets were more informative, reliable and accurate which lead to higher classification accuracy. The results are shown in Figure 10.

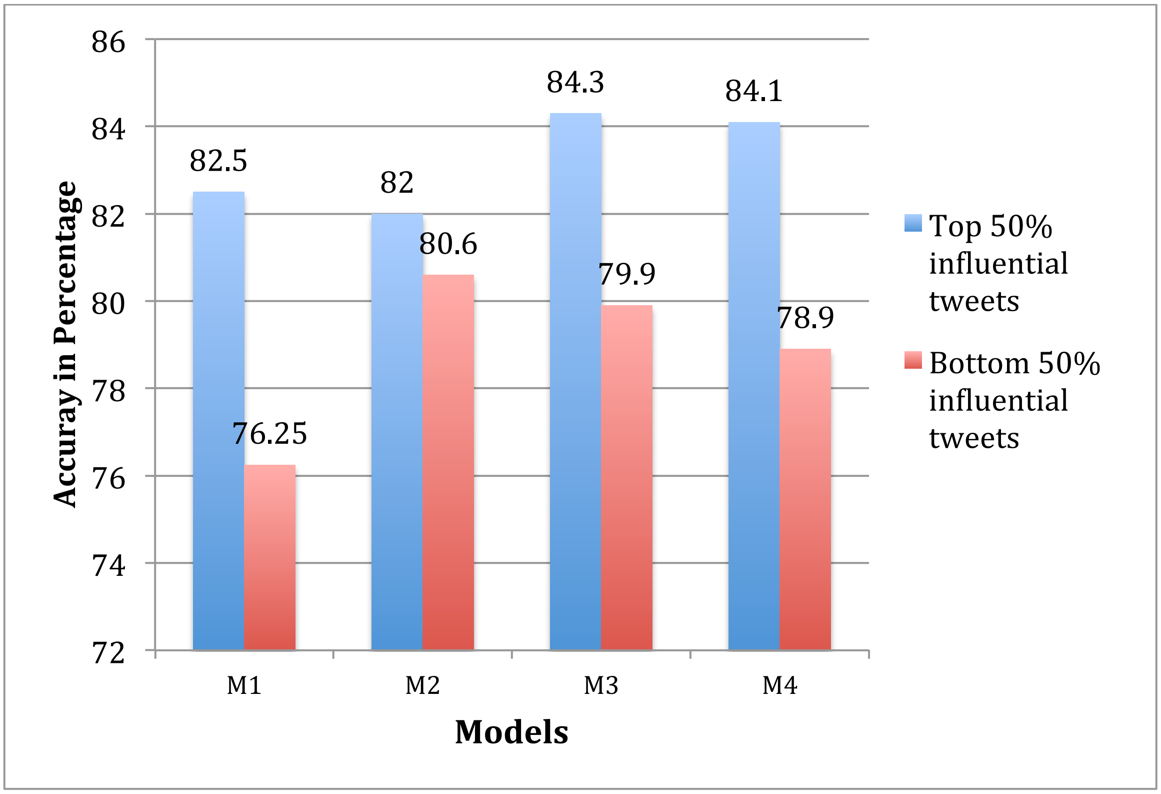


Figure 10 Accuracy for classification in workload W2

We measured the average time to finish the classifications in workload W2 and present the results in Figure11. The time difference was very close for each pair because the top training dataset had the same number of tweets as the bottom one had. The largest time difference was less than 1 minute. In addition, the execution time increased as the training dataset grew, but not by much.

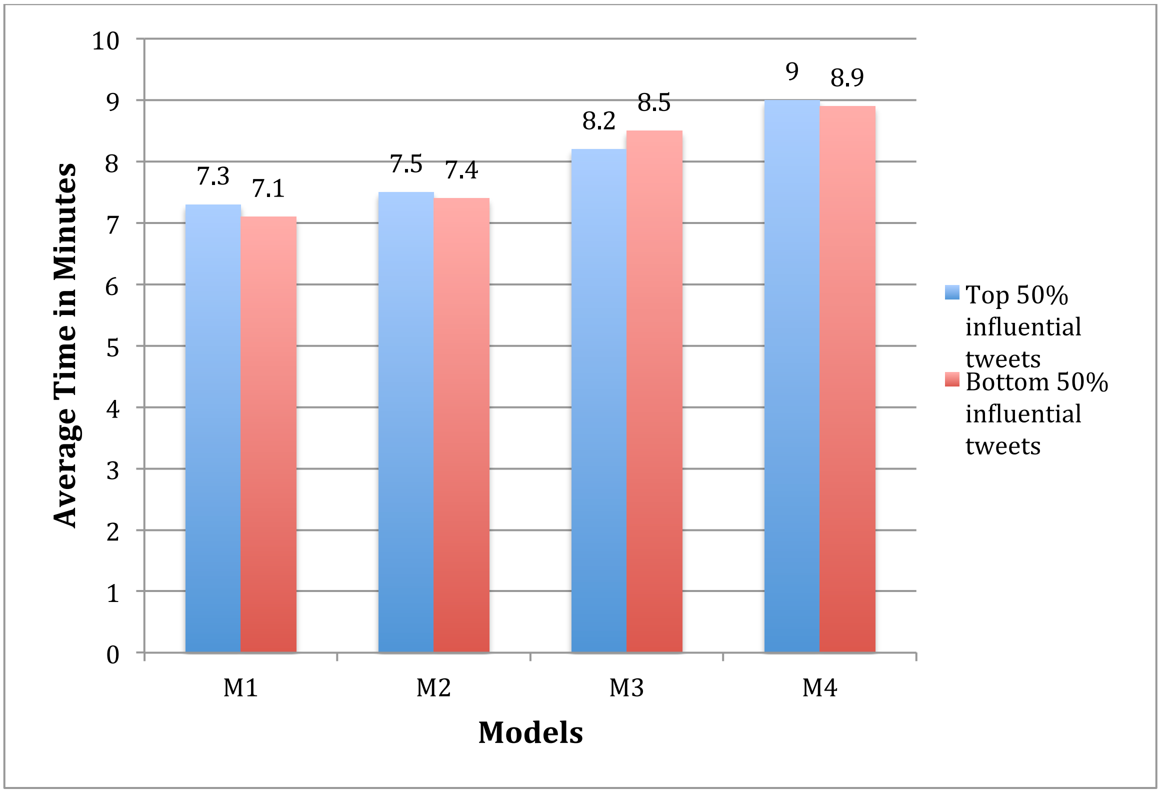


Figure 11 Time consumption for classification in workload W2

4.4 Scalability

To further exploit the scalability of CTC, we measured the average time taken to complete classifications with different data sizes in physical Hadoop cluster using workload W3. To achieve a reliable evaluation, we used both Model 3 and Model 4 to perform classifications on 1 million, 10 million and 100 million tweets respectively. At this time, the time consumption of training data model and uploading input file were excluded. We only focused on the time consumption of classification job. The processing time for workload W3 is shown in Figure 12.

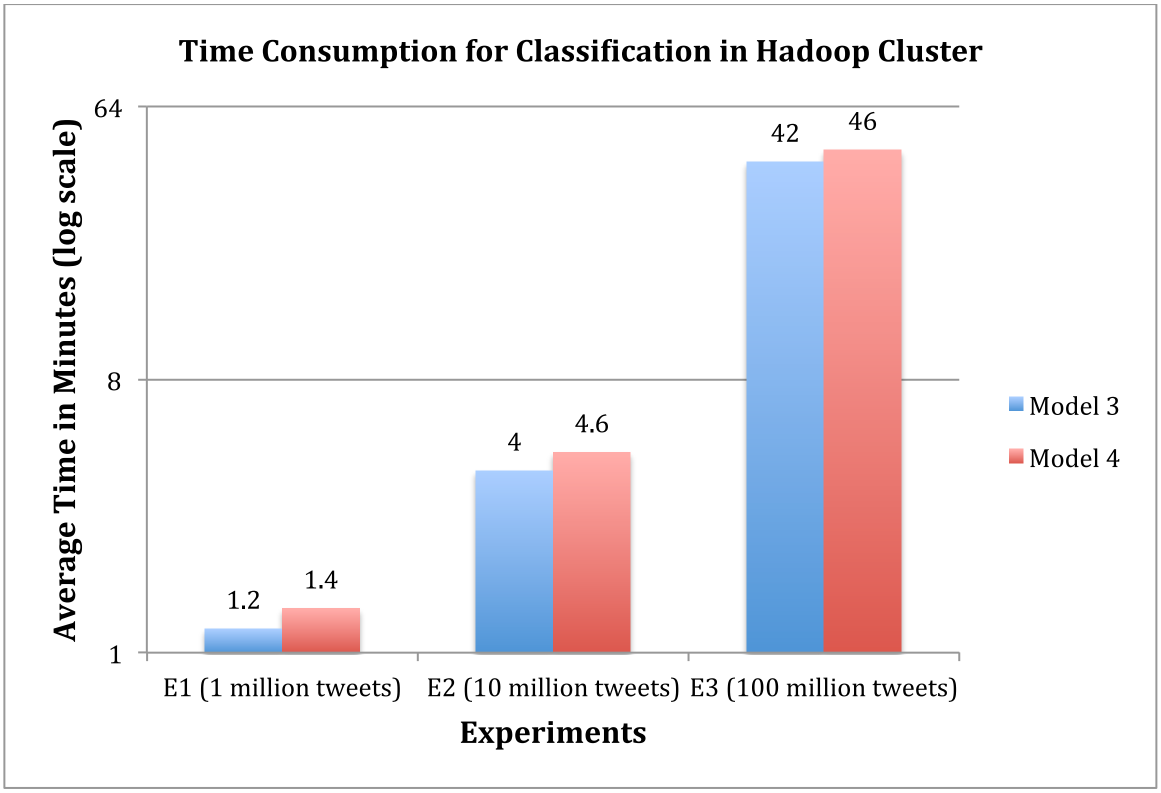


Figure 12 Time consumption for classification in workload W3

The results showed that for each model, the average time to finish classifying 1 million tweets was less than 2 minutes. When the dataset increased to 10 million, the average time consumption was less than 5 minutes. When the dataset reached 100 million, each classification consumed less than 50 minutes.

As we expected, the time consumption increased with respect to the dataset size. From the comparison between E1 and E2, we can see that when the dataset increased 10 times, the time consumption did not increase 10 times. This happened because there were only 2 map tasks in E1 which did not fully take advantage of the cluster with 6 nodes. Four nodes have been idle while 2 nodes were performing tasks. However, in E2 the input file has been split into 13 blocks. Each node in the cluster has been assigned at least two tasks to perform.

From the comparison between E2 and E3, the time consumption increased linearly along with the increase in the data size because the cluster has already been taken full use since the dataset increased to 10 million. Clearly, these dataset benefited from the parallelization of Hadoop since it was larger than the default block size in HDFS.

Overall, the main evaluations of CTC are summarized as follows:

- We trained four models based on different feature sets and validated the accuracy of each model. Both Model 3 and Model 4 produced very high accuracy. Model 4 leveraged by SNOMED CT produced the best accuracy of 84%.
- For each model, we split the training dataset into two subsets with equal size based on the influence score of the tweets. Our results demonstrate that the training dataset with more influential tweets always performed better accurate classification than the one with less influential tweets.
- We used Model 3 and Model 4 to test the scalability of CTC with different number of tweets in a physical cluster. The results showed that CTC could easily and efficiently scale up to classify 100 million tweets in less than 50 minutes.

CHAPTER 5

CONCLUSION AND FUTURE WORK

There are quite a lot of requirements for today's data classification: how to build an efficient model to improve the classification accuracy is a critically significant problem. In addition, the scalability is an urgent demand due to the increasing large amount of data. Speed performance is another concern when there is a time requirement. How to make a balance of these aspects, especially in distributed systems is important.

With this project, we presented an efficient Clinical Tweets Classifier CTC for healthcare information analysis by classifying clinical tweets into related categories using Apache Mahout and Hadoop. CTC applied SNOMED CT as well as implemented tweet influence algorithm to prepare better training dataset to achieve higher classification accuracy. Multiple classification models were built by applying different features to prepare training datasets. The results showed that the classifier integrated with SNOMED CT produced the highest classification accuracy among all models. Besides, the models trained with top 50% influential tweets turned out to have higher classification accuracy compared with those trained with bottom 50% influential tweets.

In addition, by adopting Naïve Bayes classifier with Hadoop MapReduce framework, CTC can easily and efficiently scale up in distributed system to handle large amount of data. Our experiments demonstrate that the Naïve Bayes classifier could easily scale up to 100 million tweets as well as efficiently perform classification jobs in our physical Hadoop cluster. CTC turned out to be an inexpensive solution for classification in distributed system environment.

We believe this work is just the beginning to analyze healthcare information on social network using Mahout machine learning algorithms along with Hadoop framework. The next step would be comparing classification with different algorithms. In addition to Naïve Bayes, Hidden Markov Models, Logistic Regression and Support Vector Machine (SVM) are alternatives to build classification model. A thorough comparison, including classification accuracy, speed and scalability among these algorithms would be studied to gain an insight. When the number of a field is not evenly distributed in multiple intervals, how to come up with a better method to calculate the value is also an interesting study. In addition, we plan to collect more data and add more nodes in Hadoop cluster to fully explore the underlying capability of Hadoop framework on large-scale dataset. The performance and quality would be measured with different sizes of Hadoop cluster.

REFERENCE LIST

[1] "Twitter." Wikipedia. http://en.wikipedia.org/wiki/Twitter. (accessed October 14, 2014).

[2] B. A. Huberman, D. M. Romero, and F. Wu. "Social networks that matter: Twitter under the microscope." *arXiv preprint arXiv:0812.1045*, 2008.

[3] "Hashtag Twitter." Twitter Help Center, 2014. https://support.twitter.com/articles/49309-using-hashtags-on-twitter. (accessed October 14, 2014).

[4] H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a social network or a news media?" *Proceedings of the 19th international conference on World wide web*, pp. 591-600, 2010.

[5] C. Smith. "By The Numbers: 215 Amazing Twitter Statistics." http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/ - .U-5fWLxdV_E.

[6] "Twitter Limits." Twitter Help Center. https://support.twitter.com/articles/15364-twitter-limits-api-updates-and-following. (accessed October 11, 2014).

[7] "Representational state transfer." Wikipedia. http://en.wikipedia.org/wiki/Representational_state_transfer. (accessed October 12, 2014).

[8] "The Twitaholic.com Top 100 Twitterholics based on followers." Twitaholic. http://twitaholic.com/. (accessed October 12, 2014).

[9] "The Klout score." Klout. https://klout.com/corp/score. (accessed October 12, 2014).

[10] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. "Everyone's an influencer: quantifying influence on twitter." *Proceedings of the fourth*

ACM international conference on Web search and data mining, pp. 65-74, 2011.

[11] M. Trusov, A. V. Bodapati, and R. E. Bucklin. "Determining influential users in internet social networks." *Journal of Marketing Research,* vol. 47, no. 4, pp. 643-658, 2010.

[12] "The Emerging Science of Superspreaders (And How to Tell If You're One Of Them)." MIT Technology Review, 2014. http://www.technologyreview.com/view/527271/the-emerging-science-of-superspreaders-and-how-to-tell-if-youre-one-of-them/.

[13] I. Anger, and C. Kittl. "Measuring influence on Twitter." *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*. ACM, pp. 31, 2011.

[14] A. Neuhauser. "Health Care Harnesses Social Media." U.S.NEWS, 2014. http://www.usnews.com/news/articles/2014/06/05/health-care-harnesses-social-media.

[15] C. Hawn. "Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care." *Health affairs,* vol. 28, no. 2, pp. 361-368, 2009.

[16] P. H. Keckley. "Social Networks in Health Care, Communication, collabortation and insights." Deloitte Center for Health Solutions, 2010. http://www.ucsf.edu/sites/default/files/legacy_files/US_CHS_2010SocialNetworks_070710.pdf. (accessed October 20, 2014).

[17] E. Quincey, and P. Kostkova. "Early warning and outbreak detection using social networking websites: The potential of twitter." *Electronic healthcare*, pp. 21-24, 2010.

[18] M. J. Paul, and M. Dredze. "You are what you Tweet: Analyzing Twitter for public health." *ICWSM,* pp. 265-272, 2011.

[19] J. Lin, and A. Kolcz. "Large-scale machine learning at twitter." *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 793-804, 2012.

[20] D. Alexander. "Data Mining." http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/ - 3. (accessed October 19, 2014).

[21] U. Fayyad, and R. Uthurusamy. "Evolving data into mining solutions for insights." *Communications of the ACM,* vol. 45, no. 8, pp. 28-31, 2002.

[22] K. Kranz. "Dig'in Social Media-The Data Mining of Social Media with Hadoop and his Friend Mahout." 2013. http://www.ca.com/us/~/media/Files/About%20Us/CATX/digin-social-media-kranz.pdf.

[23] "Apache Hadoop." The Apache Software Foundation. http://hadoop.apache.org/. (accessed November 10, 2014).

[24] J. Dittrich, and J. A. Quiané-Ruiz. "Efficient big data processing in Hadoop MapReduce." *Proceedings of the VLDB Endowment,* vol. 5, no. 12, pp. 2014-2015, 2012.

[25] S. Konstantin, H. Kuang, S. Radia and R. Chansler. "The hadoop distributed file system." In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pp. 1-10, IEEE, 2010.

[26] "Hadoop Glossary: What is MapReduce?" IBM. http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/. (ccessed November 10, 2014).

[27] "Machine Learning." Wikipedia. http://en.wikipedia.org/wiki/Machine_learning. (accessed November 11, 2014).

[28] "What is Machine Learning?" Machine Learning Platform, SNN Adaptive Intelligence. http://www.mlplatform.nl/what-is- machine-learning/. 2011.

[29] "What's Apache Mahout." The Apache Software Foundation. https://mahout.apache.org/. 2014.

[30] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen. "Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier." *Big Data, 2013 IEEE International Conference on*, pp. 99-104, IEEE, 2013.

[31] L. Ma, E. Haihong, and K. Xu. "The design and implementation of distributed mobile points of interest (POI) based on Mahout." In *Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on*, pp. 99-104, IEEE, 2011.

[32] R. M. Esteves, and C. Rong. "Using Mahout for clustering Wikipedia's latest articles: a comparison between K-means and fuzzy C-means in the cloud." In *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, pp. 565-569, IEEE, 2011.

[33] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. "Measuring User Influence in Twitter: The Million Follower Fallacy." *Icwsm,* vol. 10, pp. 10-17, 2010.

[34] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, & A. Choudhary. "Twitter trending topic classification." *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 251-258, IEEE, 2011.

[35] M. J. Paul, and M. Dredze. "A model for mining public health topics from Twitter." *Health,* vol. 11, pp. 16-6, 2012.

[36] "Apache Maven Project." The Apache Software Foundation. http://maven.apache.org/. 2014.

[37] "Twitter4J." Twitter4J. http://twitter4j.org/en/index.html. (accessed November 12, 2014).

[38] "Twitter Developer." Twitter. https://dev.twitter.com/apps. (accessed November 12, 2014).

[39] "JavaScript Object Notation." Wikipedia. http://en.wikipedia.org/wiki/JSON. (accessed October 15, 2014).

[40] "JSON Online Parser." http://json.parser.online.fr/. (accessed November 18, 2014).

VITA

Li Wang was born on May 4th, 1988 in Wuhan, China. He graduated from Xiangyang No.4 High School in 2006. He was admitted into Wuhan Textile University in Wuhan, China in the same year. He received his Bachelor's degree of Computer Science in 2010.

He was accepted to School of Computing and Engineering at University of Missouri-Kansas City to pursue a Master of Science degree of Computer Science. His main field of interest is data mining. He worked as an intern in FutureWei Technologies Inc, CA and Accelerated Vision Group LLC, KS during his Master studies.