Social Sensor Networks for News Mining

M. A. I. D. Fernando

2019

Social Sensor Networks for News Mining

A dissertation submitted for the Degree of
Master of Philosophy

M. A. I. D. Fernando
University of Colombo School of Computing
2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.
To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Students Name : M. A. I. D. Fernando

Signature: Date:28.05.2019

This is to certify that this thesis is based on the work of Mr./Ms. M. A. I. D. Fernando under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Main supervisor Name: Dr. T. N. K. De Zoysa

Signature: Date: 28.05.2019

Abstract

With the development of technology, many people tend to use the Internet which has resulted in an increase in usage of social networks and microblogs, inducing many organizations too to share their news in social networks and microblogs. News providers are such organizations that share large amount of news in social networks and blogs and Twitter is one such common social network, which is well known as a microblog. The short messages (Tweets) which are shared in Twitter can produce many important information. S2Net tool was developed in order to analyze these Tweets and generate useful information and present it in a suitable manner.

Situations where one is interested in the news topics rather than news groups. For such cases, the clustering technique was used, in which the news was clustered into news topics. Expectation–Maximization clustering (EM Clustering) and Hierarchical Clustering were the methods used in these situations. The results show that Hierarchical Clustering with Simple Linkage function performs better than EM Clustering. The Simple Linkage function can detect the small relationships between clusters. Because of the high dimension of the features, there will be many relationships which are hard to detect. Therefore using Simple Linkage function can improve the accuracy.

There can be situations where one is interested in the news topics rather than news groups. For such cases, the clustering technique was used, in which the news was clustered into news topics. EM Clustering and Hierarchical Clustering were the methods used in these situations. The results show that Hierarchical Clustering with Simple Linkage function performs better than EM Clustering. The Simple Linkage function can detect the small relationships between clusters. Because of the high dimension of the features, there will be many relationships which are hard to detect. Therefore using Simple Linkage function can improve the accuracy.

These two analyzing methods were evaluated using two evaluation techniques. The classification method was evaluated using F-measure. According to the F measure, it is clear that the Random Forest method performs well than the other methods. The clustering method was evaluated by getting review comments for the each cluster. The reviewer evaluates and marks the mismatches for each cluster. According t**o** their evaluations, EM clustering performs with 68.52% accuracy and Hierarchical clustering performs with 89.93%.

Acknowledgments

Dr. T. N. K. De Zoysa, Senior Lecturer-University of Colombo School of Computing and my main supervisor of the research project for the invaluable help, guidance and advice given throughout the research.

Dr. H. L. Premaratne, Senior Lecturer-University of Colombo School of Computing, for his immense support, guidance and suggestions throughout the research.

Mr. K. M. Thilakarathne, for advices and support given.

The researchers of WASN lab of University of Colombo School of Computing, for helping me by providing their precious time on analyzing the results. To all my staff members, for providing their help to achieve my goals at my work place while doing this research.

To all the staff of the University of Colombo School of Computing - my dear lecturers and teachers of the School, for the knowledge and advices provided to make me the person whom I am.

To all my friends at school, my university and the students of University of Colombo for the help provided in data collection and various other ways. Last but not least my dear husband, son and daughter, my dear father, my dear mother and my sister, for the invaluable support provided to me in various forms throughout my life.

Contents

List of Figures

List of Tables

List of Acronyms

Chapter 1

Introduction

1.1 Introduction

Recent developments in the field of web technologies make it easy for users to contribute contents to online communities causing an increase of user participation online. For example, social network sites such as Facebook and Twitter, mailing lists and discussion forums make users to communicate and share contents with other users online. When more users contribute contents, the data they input to the web will produce valuable information for social research.

Users tend to express opinions and views on current affairs ranging from gossips about celebrities, politicians and political events, religions, and latest releases of their favorite computer games. These are immensely important sources of information to gauge the moods and opinions of the society. Each contributing user can be considered as a tiny sensor in the society, knowingly or unknowingly leaking his/her view into the public network. However, with this massive amount of information, it is hard for humans to keep track of what is being said and done by each person and gather useful information to determine their moods and general opinion about the society.

In this research we propose a model to monitor textual information entered by online users and summarize them into topics. For example: education in Sri Lanka, crimes in Sri Lanka etc. By analyzing the WWW data, we will determine the actual situation (whether it is a popular topic or not) of a particular topic, by focusing about the particular events going on regarding that topic in given time period. Then that information will be used for further social research.

1.2 Motivation

With the development of WWW, people tend to use social networks actively and the society contributes large number of textual information to the web. One main source where the textual information comes is the news providers. The news which is published in social networks will be a good information source if summarized well. Once processed, it can be used for various types of researches such as social research, marketing etc. The motivation for the research is to provide a central unit where this processed information can be extracted easily. Thus, the suggested system can be an infrastructure to build reports for many researches.

1.3 Goals & Objectives

Our suggested way of summarizing the news was classifying and clustering them according to the news topics. By analyzing the textual information, we can determine the popularity of given news. Then that information can be used for further social research. This Social Sensor Network can be considered as an infrastructure to build other applications and it is not an application by itself. It is a means to make the humongous amount of information continuously flowing into the web, manageable for a specific purpose.

The main objective is to build a system where the news topics can be identified. This system should be able to fulfill the following 2 tasks.

- Classify the news into predefined groups
- Cluster the news into particular topics

1.4 Significance of the research

Many other text classification researches classify the text into one group at a time. The suggested system classifies the text into multiple groups. It has introduced a new feature selection method for classification. Then the system extracts the topics of the news. This was done by using clustering techniques.

1.5 Description of Data

As for many real time applications, we use sensors to gather data. In this case, it will be human sensors. Our focus is the information which the humans provide using social networks. Data will be extracted from a social network called Twitter. This will be done by creating a twitter user account and extracting tweets from publicly available accounts. Using python-twitter wrapper codes, the data will be extracted into a text file. A Java script will be developed in order to save the data into a database.

To avoid the complexity, we will use English text. The users are restricted and the text will be taken only from Twitter news providers as the texts which will be posted by news providers are well structured. The model will be developed for local news providers. Five active local news providers were chosen for the data gathering. The users were,

- Ada Derana

- Ceylon Today
- ITN
- Lanka Breaking News
- News First

The news was classified into 12 categories manually. The 12 categories were defined according to popular newspaper articles and news websites [1]. The 12 groups were,

- Economic-Business
- War-terrorist-crime
- Health
- Sports
- Development-government
- Politics
- Accidents
- Entertainment
- Disaster-Climate
- Education
- Society
- International

These tagged data were used for classification purpose.

1.6 Overview

The overview of the thesis is as follows. Chapter 2 describes the literature review for the research. The chapter 3 describes the methodology which was used for the research. There was a data gathering process in the research. Therefore, chapter 4 describes the data gathering and pre-processing stage of the research. The system used both classification methods and clustering methods. The chapter 5 describes the feature selection method for classification. The chapter 6 describes the classification process and the chapter 7 describes the clustering process. Finally, the chapter 8 describes the general discussion and the new system.

Chapter 2

Literature Review

2.1 Overview

The most important factor of the research was finding a way to classify the news and cluster the news. This chapter contains the details of literature on past researches carried out in text categorizing methods. Researchers have developed various text categorization methods in different situations such as document categorizing, open ended questionnaire categorizing etc. The most common techniques which they followed were Artificial Neural Network, Support Vector Machine, Decision trees etc. This chapter briefs out the strengths and drawbacks for each situation in each techniques.

2.2 Text Categorization

Text categorization is the process of sorting text documents into one or more predefined categories Basu et al. [2]. The phrase "text categorization" was mainly used in order to denote a system which is able to analyze large quantities of natural language text and extract information [3].

Text categorization is, classifying a document or a group of texts into a fix number of categories. Each text will belong to one category, multiple categories, or no category at all. Each category will be treated as a separate binary classification problem since categories can be overlapped [4].

Text categorization could be done by using text mining techniques and natural language processing techniques. When comparing with other data stored in databases, text is unstructured, amorphous and difficult to deal with algorithms [3]. Therefore, unlike data mining, text mining is difficult. Witten [3] had described the difference of text mining and data mining as, "data mining is about identifying the patterns in data and text mining is about identifying patterns in text".

Text classification techniques can be used in classifying news stories, to find interesting information on the World Wide Web, and to guide users search through hypertext [4]. Thus, it will be helpful in classifying sentences and short news into predefined groups.

2.3 Existing software packages for text categorizing

Alexa et al. [5] had conducted a research based on 15 existing software which are currently available for text categorization. The software is, AQUAD, ATLAS.ti, CoAn, Code-A-Text,DICTION, DIMAP-MCCA, HyperRESEARCH, KEDS,NUD*IST, QED, TATOE, TEXTPACK, TextSmart,WinMAXpro, and WordStat. These software packages are based on specialized dictionaries (rules). Each text fragment is assigned into a specific category if and only if they contain words matching those in the relevant category of the dictionary. Thus, the disadvantage is that, the dictionary has to be defined before the coding process begins and the dictionary needs to be developed regularly.

2.4 Data set

Basu et al. [2] had used Reuters News Data Set for the suggested classification algorithm. It contains 21,578 short news items (average 200 words in length), and was classified manually into 118 categories. As he used Artificial Neural Network techniques, the data set had to be divided into the training set and the testing set, the researcher used ModApte split to split the data.

Thus Basu et al. [2] used manual classification methods to create the training data set. Giorgetti et al. [6] shows that this manual system has a drawback as this process is likely to produce faulty encoding. They used data from NORCs General Social Survey in order to conduct the research.

Pak et al.[7] had used news portals which were collected from Twitter microblog. They gave four reasons for using Twitter microblog for their research.

- Microblogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions
- Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.
- Twitters audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interest groups.
- Twitters audience is represented by users from many countries. Although users from U.S. are prevailing, it is possible to collect data in different languages.

2.5 Feature extraction

Giorgetti et al. [6] used information retrieval (IR) as a tool to conduct document indexing. The text was typically represented as a vector of term weights which was computed by *tf* × *idf*, where *tf* is the term frequency and *idf* is the inverse document frequency.

Basu et al. [2] had stored data in KSS (Knowledge System Server) in order to create the dictionary. This contains 102,283 items. The IQ value which was provided by KSS measures the importance of the given item. The researcher had used two threshold values, 57 and 87 and chosen two item sets correspondingly 33,191 and 62,106.

The bag of words representation for a document was introduced by Witten [3]. This is basically to index every individual word in the document collection. Each document was represented as a bag of words. There are some practical issues in this method as to how to deal with numbers etc.

Joachims [4] had used the same approach by defining each distinct word as a feature, and number of times a word occurs in the document as its value. Words were considered as features only if they had occurred in training data at least 3 times in order to optimize the number of features.

Yu [8] follows the same approach and the researcher shows without the feature reduction, "a document vector" is often defined in a space of thousands of dimensions where each dimension corresponds to a word. The researcher suggests three tools in order to reduce the feature dimension. And they are stemming, stop word removal and statistical feature selection.

2.5.1 Removing stop words

Basu et al. [2] shows that the reduced feature set provides better performance than the full feature set. Yu [8] agrees to this suggestion in his research by stating that common words and functional words that are considered as synonyms should be removed from the feature set in order to achieve high performance.

2.5.2 Stemming algorithms

Frequently, a document may consist of similar words in different formats by adding prefix or postfix. This will cause to increase the feature dimension unnecessarily. Yu [8] had used Porter Stemmer [9] in order to stem the words. This may allow cutting off the prefix and post fixing which then results the actual word.

Yu [8] had used another algorithm which was developed by Baker and McCallum (1998) in order to measure the similarities of two words. This algorithm was named as averaged Kullback-Leibler Divergence (KLD). The smaller the KLD value, the similar the words are (minimum is 0).

2.6 News categories

Defining news categories is one of the major tasks in text categorization. Predefining news categories is essential in supervised learning. In unsupervised learning techniques, the divided clusters are the news categories. Bacan et al. [10] had conducted a clustering method in order to cluster the news items. The researchers had obtained several categories as the result. The obtained categories and the descriptions are given in table 2.1.

Lin et al. [11] had conducted a research in order to discover the information from web pages. Information would be categorized according to the content included. The researcher had divided the web pages manually into 12 categories as network investment, life, Taiwan news, investments, supplement, miscellaneous news, daily news, headlines, stock and financial, society, international and city. Some web pages were manually categorized into each category and other categories were identified using this manual categorization.

2.7 Data training

Data training can be done in two manners: using supervised learning techniques and unsupervised learning techniques. In unsupervised learning, the class tags are unknown. The system itself identifies the patterns of clusters using features and the system clusters the data. Supervised learning techniques use a set of training documents that have already been associated with a category to determine which feature set of the documents will produce the desired results [2]. Thus, predefining news categories for supervised learning techniques is a must.

2.7.1 Supervised learning techniques

Supervised learning techniques can be used to classify short messages into predefined categories. This method is more accurate than unsupervised learning methods.

Support Vector Machine

Support Vector Machine (SVM) is supervised learning techniques which are commonly used in text categorization. Joachims [4] had explored the benefits of SVM for text categorization. The researcher states one remarkable property of SVM is that its ability to learn is independent from the dimension of the feature space. Joachims [4] gives five reasons why SVM performs well in text classification.

Few irrelevant features

Using a bag-of-words method does not increase the dimension as using n-gram. However it increases the dimension for a considerable amount. The researcher had removed irrelevant features by removing common words and noise words. However, in text, there are very few numbers of irrelevant features. Thus, the dimension cannot be reduced to an acceptable level.

Table 2.1: Clustered news topics

Category name	Category description
Arts, culture and Entertainment	Matters pertaining to the advancement and refinement of the human mind, of interests, skills, tastes and emotions
Crime, Law and Justice	Establishment and/or statement of the rules of behavior in society, the enforcement of these rules, breaches of the rules and the punishment of offenders. Organizations and bodies involved in these activities.
Disaster and Accident	Manmade and natural events resulting in loss of life or injury to living creatures and/or damage to inanimate objects or property.
Economy, Business and Finance	All matters concerning the planning, production and exchange of wealth.
Education	All aspects of furthering knowledge of human individuals from birth to death.
Environmental Issue	All aspects of protection, damage, and condition of the ecosystem of the planet earth and its surroundings.
Health	All aspects pertaining to the physical and mental welfare of human beings.
Human Interest	Lighter items about individuals, groups, animals or objects.
Labor	Social aspects, organizations, rules and conditions affecting the employment of human effort for the economic support of the unemployed.
Lifestyle and Leisure	Activities undertaken for pleasure, relaxation or recreation outside paid employment, including eating and travel.
Politics	Local, regional, national and international exercise of power, or struggle for power, and the relationships between governing bodies and states.
Religion	All aspects of human existence involving theology, philosophy, ethics and spirituality.
Science and Technology	All aspects pertaining to human understanding of nature and the physical world and the development and application of this knowledge
Social Issue	Aspects of the behavior of humans affecting the quality of life.
Sport	Competitive exercise involving physical effort. Organizations and bodies involved in these activities.
Unrest, Conflicts and War	Acts of socially or politically motivated protest and/or violence.
Weather	The study, reporting and prediction of meteorological phenomena.

High dimensional input space

Because of few irrelevant features, still it remains large number of features, even after removing the irrelevant features. This will cause to have over fitting. However, when using SVM, the high dimensional space need not be dealt with directly [12]. To obtain the hyper plane, SVM do not deal with all data, it only considers support vectors. The support vectors were chosen as the vectors x_i where the i (Lagrange multiplier) is greater than zero. However, when applying Lagrange multiplier, a constant C will be applied as C should be greater than or equal to i (Lagrange multiplier). When this C becomes infinite, it will result in a more complex optimal hyper plane which will completely separate the data. Thus, having a constant C will allow having some misclassification, which eliminates over fitting.

Document vector are sparse

Since there are large number of features (words) and a short message relatively contains less amount of words, there are only few entries which are none zero (spare vectors) in the created dataset. Kivinen et al [13] showed that perceptron algorithm can take the advantage of sparse instances. Thus, using inductive bias like SVM will avoid the errors which additive algorithms generate.

Most text categorization problems are linearly separable

Joachims [4] shows that most text categorization problems are linearly separable. This researcher shows that if the problem is not linearly separable, it is due to misclassification of human indexes. Thus, the objective of SVM is a perfect match with text categorization problems.

Words of the short messages are not independent

Rennie et al [14] shows that the words are dependent on each other. Thus, the main assumption of Naive Bayes, that features should be independent, will be violated at this point. Thus, it will be appropriate to use SVM for twitter text classification, rather than using Naive Bayes classifier.

Basu et al. [2] had compared an Artificial Neural Network (ANN) algorithm with a Support Vector Machine algorithm in order to use as text classifications of new items. According to Basu et al. [2], the computational complexity for SVM is N m2 where N is the number of classifiers and m is the number of training examples. Thus, the performance is more sensitive to the number of training examples than to the number of classifications.

Yu [8] shows that existing studies indicate that SVMs are among the best text classification to date. Joachims [4] supports this result by providing 4 theoretical evidence (high dimensional input spaces, few irrelevant features, text categorization problems are linearly separable and document vectors are sparse which makes SVM more suitable.

Giorgetti et al. [6] had done the research using both SVM and Naive Bayes classifier in order to compare with dictionary-based approaches and obtain that SVM provides more than 26% accuracy than the dictionary-based method.

Nave Bayes classifier

Nave Bayes classifier is one of the most popular methods in Artificial Neural Network (ANN). Basu et al. [2] had chosen an ANN without a hidden layer (perception approach) as it is easy to construct. The researcher had taken only two categories per training network. In ANN, the classifier has one input layer of one node per item. The output layer contains one node either Zero or one to indicate two categories. The computational complexity for ANN is V^2N where V is the number of attributes in the item vector (number of input nodes), N is the number of classifiers. The major advantages of using Nave Bayes classifier is, it learns faster than other techniques Yu [8]. McCallum et al [15] used Multinomial Model of Naive Bayes, in order to deal with word frequencies.

Giorgetti et al. [6] had done the research using both SVM and Naive Bayes classifier in order to compare with dictionary-based approaches and obtain that Naive Bayes classifier provides more than 18% accuracy than the dictionary-based method.

Decision tree/ rule learner

Witten [3] had used rules and decision trees method. The researcher had defined several rules in order to assign a document to a particular category. These rules can be produced automatically using standard techniques of machine learning.

Random Forest

Rios et al. [16] had conduct a research regarding span detection. They had used SVM and Random Forest [17]. They had used R implementation of Breiman's Random Features. They had tried different number of trees and found that 500 trees seem to be quite adequate for the size of data which they had.

They had compared the performance of SVM and Random Forest with Naive Bayes classifier and figured out that both SVM and Random Forest performs better than Naive Bayes classifier.

KNN Classifier

Bacan et al. [10] suggests k-nn (k Nearest neighbor) method for classification news items. Each instance was assigned with a value which was named as weight of word i in j document. This weight was calculated based on frequency of the i^{th} word in document j. The following equation was used to calculate the weights.

$$W_{ij} = tf_{ij} \times log(\frac{N}{n})$$

The best result was obtained when k = 10. Therefore, the researcher obtains 10 categories in order to assign the news items.

2.7.2 Unsupervised learning

Unsupervised learning techniques can be used when the categories are unknown. The number of categories needs to be defined.

EM Clustering

Bhan[18] had Compared EM Clustering with K-means Clustering. The researcher had found that K-Means algorithm is very poor at handling overlapping data. He says that it is because it is only able to classify a point based on its distance from the estimated means. Researcher shows that EM does much better on the overlapping data, as the strength of EM lies in the fact that it is able to incorporate underlying assumptions about how the data was generated.

Hierarchical clustering

Zhao et al [19] experimentally evaluated nine agglomerative algorithms and six partitional algorithms to obtain hierarchical clustering solutions for document datasets. Their experimental results showed that partitional methods produce better hierarchical solutions than agglomerative methods.

K-Means Clustering

Steinbach et al. [20] had compared K-means and Hierarchical clustering. The less time complexity is the main advantage which they figure out when using k-means algorithm. However, they discovered that bisecting K-means algorithm performs well as Hierarchical algorithm.

2.8 Evaluation

In pattern recognition, precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both can be used in order to measure the accuracy of the developed system. However, to get a clear picture of the situation, both precision and recall values are important. Hence, Basu et al [2] had used the combined value of recall (r) and precision (p) in order to measure the accuracy. The following equation is used to compute the combined value.

$$F(r,p) = \frac{2rp}{r+p}$$

Rios et al [16] had used Receiver Operating Characteristic (ROC) to measure the performance where the ROC curve is the plot between true positive rate and false positive rate.

2.9 Summary

Text mining is different from data mining and analyzing text is harder than data. Thus, many researchers had used bag of words method to convert the text into an analyzable format. This was created by removing stop words and applying a proper stemming algorithm. The remaining words are called instances.

These instances are taken as variables. Different types of weights are chosen in order to assign values to each variable corresponding to each document of short news. Many new weights are introduced. Using these data, many researchers had applied supervised learning techniques in order to train the data. Some had applied unsupervised learning techniques. Joachims [4] states that SVM method, which is a supervised learning method, was more accurate than k-NN, Bayes and Rocchio methods. Comparing to training time, SVM is faster than k-NN at classification time [4].

The accuracy in some learning techniques depends on the dimension size of features. ANN is more sensitive to the size of the term vector than the SVM algorithm [2]. However, SVM eliminates the need of feature selection [4]. Basu et al. [2] had carried out Student t-test in order to figure out the performance of SVM and ANN. The results say that SVM performs better than ANN. Furthermore, the researches show that the reduced feature set provides more performance than the full feature set.

Chapter 3

Methodology

3.1 Overview

The previous chapter describes the state of the art and the background literature which are related to this research. There are several theories which were used for this research. This chapter describes the details of the theories which were used throughout the research. The base of the theoretical framework is, Data mining.

3.2 Classification Vs. Clustering

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts for the purpose of using the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known) [21].

Unlike classification and prediction, which analyze class-labeled data objects, clustering techniques analyses data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [21].

3.3 Classification Methods

3.3.1 Support Vector Machine

SVM is a newly introduced classification method which **is** used for binary classification. It was introduced by Vapnik and Colleagues. The researcher had used several techniques to avoid over fitting. The basic idea is to find a hyper plane which separates the dimensional data into 2 classes. For data which is not linearly separable, SVM introduced a notion of a "Kernel introduced feature space". This casts data into high dimensional space where the data is linearly separable [21].

3.3.2 Decision trees

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules.

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basis of several commercial rule induction systems.

C 4.5 (J48)

Among decision tree algorithms, J. Ross Quinlan's ID3 and its successor, C4.5, are probably the most popular in the machine learning community. C4.5 use formulas based on information theory to evaluate the "goodness" of a test; in particular, they choose the test that extracts the maximum amount of information from a set of cases, given the constraint that only one attribute will be tested[22] .

Random Tree

Class for constructing a tree that considers K randomly chosen attributes at each node. Performs no pruning. Also has an option to allow estimation of class probabilities based on a hold-out set (back fitting).

Random Forest

Random forests is an Ensemble method (bagging) of Random Trees. It is a combination of Random tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [17].

3.3.3 Ensemble methods

Ensemble methods are methods that use a combination of models instead of a single model. Bagging and boosting are two such techniques. Each combines a series of k learned models (classifiers or predictors), M1, M2,… Mk, with the aim of creating an improved composite model, M. Both bagging and boosting can be used for classification as well as prediction [21].

Bagging

Given a set, D, of d tuples, bagging works as follows. For iteration i (i = 1, 2,... k), a training set, Di, of d tuples is sampled with replacement from the original set of tuples, D. A classifier model, Mi , is learned for each training set, Di . To classify an unknown tuple, X, each classifier, Mi , returns its class prediction, which counts as one vote. The bagged classifier, M*, counts the votes and assigns the class with the most votes to X [21].

Boosting

In boosting, weights are assigned to each training tuple. A series of k classifiers is iteratively learned. After a classifier M_i is learned, the weights are updated to allow the subsequent classifier, M_{i+1} , to ”pay more attention” to the training tuples that were misclassified by M_i. The final boosted classifier, M, combines the votes of each individual classifier, where the weight of each classifiers vote is a function of its accuracy. The boosting algorithm can be extended for the prediction of continuous values.

3.4 Clustering Methods

3.4.1. EM Clustering

The EM (Expectation-Maximization) algorithm is a popular iterative refinement algorithm that can be used for finding the parameter estimates. It can be viewed as an extension of the k-means paradigm, which assigns an object to the cluster with which it is most similar, based on the cluster mean. Instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of

membership. In other words, there are no strict boundaries between clusters. Therefore, new means are computed based on weighted measures.

EM starts with an initial estimate or guess of the parameters of the mixture model (collectively referred to as the parameter vector). It iteratively rescores the objects against the mixture density produced by the parameter vector. The rescored objects are then used to update the parameter estimates. Each object is assigned a probability that it would possess a certain set of attribute values given that it was a member of a given cluster.[21]

3.4.1 Hierarchical Clustering

A hierarchical clustering method works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. That is, if a particular merge or split decision later turns out to have been a poor choice, the method cannot be backtrack and corrected[21]. In general, there are two types of hierarchical clustering methods: Agglomerative and Divisive Hierarchical Clustering.

Agglomerative Hierarchical Clustering

This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category. They differ only in their definition of inter-cluster similarity[21].

Divisive Hierarchical Clustering

This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the cluster into smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold[21].

3.5 Evaluation methods

3.5.1 Precision

This is the percentage of retrieved documents that are in fact relevant to the query[21]. It is formally defined as

$$Precision = \frac{|Relevent \cap Rtrieved|}{|Retrieved|} \quad (3.1)$$

3.5.2 Recall

This is the percentage of documents that are relevant to the query and were in fact retrieved. It is formally defined as,

$$Rcall = \frac{|Relevent \cap Rtrieved|}{|Relevant|} \quad (3.2)$$

3.5.3 F-Measure

An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used trade-off is the F-score, which is defined as the harmonic mean of recall and precision:

$$Fscore = \frac{recall \times precision}{\frac{(recall+precision)}{2}} \quad (3.3)$$

3.6 Summary

There are several classification methods and clustering methods which are used in this research. It is important to understand the theory of each method so that it could be easy to explain and justify the result of each method. This chapter described the theory of each classification and clustering method which was used for the research. The next chapter describes the data gathering process and the pre-processing process.

Chapter 4

Data gathering and pre-processing

4.1 Overview

The previous chapter briefed out the methodologies which were used for the research. This section will brief out the data gathering and preprocessing of the research. Data gathering and pre-processing holds a large weight of a text classification research. This chapter will describe the main domain of data gathering with the intent to choose a specific domain. It also describes the pre-processing activities of the research.

4.2 Data gathering

The main idea of the research is to extract the hidden useful information from Social Networks. Therefore, the data should be gathered from a Social Network. The major problem was to find a Social Network where the data are publicly available as many Social Networks do not allow sharing of data. Thus, the data was gathered from Twitter microblog. Twitter allows the user to generate a set of keys which allows developing an application based on Twitter data. Using Python-Twitter Wrappers, it is possible to get the data from Twitter.

In order to narrow down the scope, short messages published by news providers were chosen. Five local news providers were selected for data gathering. The researcher had considered about how actively they provides news, when choosing those five news providers. The chosen five news providers are,

- Ada Derana
- Ceylon Today
- ITN
- Lanka Breaking News
- News First

A script was written using python to save the data into a text file. A cron job task was set to automate the data gathering process. For the research purpose, 3600 short messages were taken within 3 months. The Appendix A contains the Python script which was written to extract data. Using this script, the ID, date and time and the text was extracted and saved in database.

The process of classifying text into groups needs training data. Thus, to create training data, we need pre labeled data. Thus, once the data were collected using python wrappers, the groups of short messages were defined manually. This manual classification was done by my research colleague. I took a sample to check the accuracy of manual classification. The short messages were grouped into 12 groups as,

- Accidents
- Development- Government
- Disaster- Climate
- Economy- Business
- Education
- Entertainment
- Health
- International
- Politics
- Society
- Sports
- War- Terrorist- Crime

4.3 Data pre-processing

Once the data is gathered, it should be pre-processed and converted to a form where it can be analyzed easily. Only the ID of the message can be taken as it is. Other entities have to be preprocessed as follows.

The date will be represented as,

Sat Jun 01 06:32:56 +0000 2013

It includes the month, year, date, day and the time. Thus, it should be divided into day, month, date, time and year. By considering it as a String array, String index was used to divide it into segments.

The short messages can be represented as follows.

President pays last respects to Jayalath Jayawardena http://t.co/JIzVlopJJr #lka

The hyperlinks were removed from the message. Then, the sentences were tokenized. These tokens were pooled together and it can be used as the extracted features.

4.4 Summary

This chapter described the data gathering and preprocessing of the research. The data was collected by using Twitter microblog. Only the news was selected as data, in order to narrow down the domain. Five local active news providers were selected to gather the data. The ID, time zone and the text message of each instance was saved in a text file. Once the data was gathered, it was preprocessed in order to extract the features. The time zone was expanded and the date, month and year are saved separately. The hyperlinks of the text message were removed. Then the remaining text message was tokenized. These tokens for each instance were pooled to gather in order to create the set of features. Even though we had created the features, it may contain noise features which do not provide any valuable information regarding the group. Thus, a feature selection procedure had to be conducted. The next chapter will brief out the feature selection process.

Chapter 5

Feature Selection for Classification method

5.1 Overview

The previous chapter briefed out the data gathering and preprocessing of the research. This section will brief out the feature selection methods which were used for the research. The research contains two parts, classifying the news into groups and clustering the news within a group into topics. Thus, feature selection had to be carried out twice, as to fulfill the two objectives separately. In this chapter, the author describes the new feature selection methods which were used for classification in this research.

5.2 Feature Extraction for Classification method

The classification process classifies the short messages into predefined groups. The defined groups were briefed out in Chapter 4. For this classification, the classifier does not need to understand the meaning of the short message. A set of keywords will be enough to classify them into groups. Thus, the short message should be converted into a set of keywords.

Many researchers had tried out different methods for this tokenization process. The ngrams and bag-of-words methods are two popular tokenization methods. In bag-of-words, the sentence is divided into words. In n-grams, the sentence is divided into n number of words. The n-gram method can be described further as follows, according to the number which we consider.

- Unigram - one word at a time
- Bigram - two words at a time
- Trigram - three words at a time

Therefore, the bag-of-words method is similar with the unigram.

However, each method has its own advantages and disadvantages. The n-gram method provides more information than bag-of-words method. However, when using unigrams, bigrams and trigrams, it causes to increase the dimension. This will cause to have more complex model. In classification process, we consider only the keyword of a short message and thereby we do not require the meaning of it. The bag-of-words method can be used to extract the features for classification process.

5.3 Feature Selection for Classification method

Once the features were extracted, the next step is to identify the best feature set from the whole feature set. Usually in text mining, the initial feature set may be high dimensional. This causes to make classifiers more complex and over fitting. Thus, it is essential to use a proper feature selection method.

5.3.1 State of the art for feature selection methods

Many feature selection methods have been developed based on Information Theory. Some of them are Information Gain, Gain Ratio, Inverse Document Frequency, Term frequency etc. There are some statistical feature selection methods such as Forward selection, Backward elimination, Chi square text etc. However, Forward selection and Backward elimination selects only few numbers of features.

Forward Selection

Sequential Forward Selection starts with the empty set and sequentially adds one feature at a time. The main disadvantage of Sequential Forward Selection is that it is unable to remove features that become obsolete after the addition of other features [23].

Sequential Backward Elimination

In Backward feature elimination, it starts with all the features and sequentially eliminates one feature at a time (eliminating the feature that contributes least to the criterion function). A problem with this Sequential Backward Elimination techniques is that when a feature is deleted, it cannot be re-selected [23].

Information Gain

Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A) [21]. That is,

$$Gain_A = Info(D) - Info_A(D) \tag{5.1}$$

Where

$$Info(D) = \sum_{i=1}^{m} p_i log_2(p_i) \quad (5.2)$$

And,

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{D} \times Info(D_j) \quad (5.3)$$

Inverse Document Frequency

Inverse Document Frequency (IDF) [24] represents the scaling factor, or the importance, of a term t. If a term t occurs in many documents, its importance will be scaled down due to its reduced discriminative power [21]. The equation of IDF is,

$$IDF(t) = log \frac{1 + |d|}{|d_t|} \quad (5.4)$$

where d is the document collection, and d_t is the set of documents containing term t. If $|d| \ll |d_j|$ the term t will have a large IDF scaling factor and vice versa.

Term Frequency

The term frequency is the number of occurrences of term t in the document d, that is, $freq_{(d,t)}$. The (weighted) term-frequency matrix $TF_{(d,t)}$ measures the association of a term t with respect to the given document d: it is generally defined as 0 if the document does not contain the term, and nonzero otherwise [21].

$$TF(d, t) = \begin{cases} 0 & \text{if } freq_{(d,t)}=0 \\ 1+log(1+log(freq_{(d,t)})) & \text{otherwise} \end{cases} \quad (5.5)$$

5.3.2 Importance of the new feature selection method

Twitter had restricted the character length into 140 characters per short message. Therefore, the number of words per sentence was restricted automatically. Thus, the main issue which occurs on using these feature selection methods is, they use the frequency of each word to measure the information. If the word occurs in high frequency, it states that it is a common word and if the word occurs in normal frequency, we can state that it is a feature word. However, in Twitter short messages, due to the character length restriction, there is no significant difference in frequencies for common words and useful feature words. There can be some words which occur frequently in one group and others that may not occur frequently due to this character length restriction. Such words provide large amount of information regarding the group but may not get selected as a feature.

Some use dictionary in order to remove common words. The issue when using the dictionary is, there can be some words where it hasn't defined as a dictionary but doesn't give any useful information in order to classify the news. The words of the dataset were needed to reduce as much as possible, because a twit carries a limited number of words.

Using too much features may mislead the classification and clustering. Thus, such words need to be omitted when creating the dataset. Therefore, a new feature selection method was introduced for Twitter short messages.

5.3.3 The new Feature Selection method

For a perfect evaluation of a feature selection, the dataset which was used for feature selection should be independent from the training dataset and testing dataset. Thus, the collected 3600 messages were divided into 3 parts as data for feature selection, data for training the classifier and data for testing the classifier. From the first thousand records (record number 1-1000) was used for feature selection, next 1500 records (1001-2500) was used to train the system and the rest was used to test the system. Before applying the suggested method, the noise words and common words had to be removed. This was done by removing low frequent words and high frequent words. The researcher defined the low frequent words as the words which have a frequent less than 10 and high frequent words were defined as the words which have a frequency more than 350. The value 10 was chosen as there was a significant difference between the number of words which had the frequency 10 and 11 (in 95% confidence interval, with mean = 6.39 and standard deviation = 10.26, the number of words which are greater than 8.29 is significantly high. The number of words which has frequency 10 is 15 and frequency 11 is 6). The value 350 was chosen as the next obtained frequency is significantly high than 350.

Then the researcher applied the feature selection method. The suggested method was based on Information Theory. The main idea of the method was, to eliminate the stop words which were not captured when removing high frequent words and to eliminate other words which do not provide much information to identify the category. A term called Frequency Ratio will be calculated for the selection process.

Assume that our classifier needs to classify Twitter short messages into *n* number of groups. Let $F_{(i,j)}$ be the frequency of the i^{th} word in j^{th} group. Then the term frequency $tf_{(i)}$ of i^{th} word can be calculated by,

$$tf_i = \sum_{i=1}^{n} F_{(i,j)} \tag{5.6}$$

Then, the Frequency Ratio, $FR_{(i,j)}$ for i^{th} word given the group j can be calculated as,

$$FR_{(i,j)} = \frac{F_{(i,j)}}{tf_{(i)}} \tag{5.7}$$

Then, the Maximum Frequency Ratio for given word *i* will be calculated as,

$$MFR_i = max\{FR_{(i,j)}\} \tag{5.8}$$

Now, by providing a threshold value, one can filter the keywords from unrelated words. For the current research, we chose 0.75 by applying trial and error method. The concept lies as follows. Assume that there is a word as “crash” (i^{th} word) and a group as “accident” (j^{th} group) If the “crash”, is a keyword which relates to “accident”, the frequency for i^{th} word in j^{th} group, $F_{(i,j)}$ (frequency of the word “crash” in the group “accident”) is high. For other groups, the frequency

$F_{(i,k)}$ where k ≠ j is low. Thus, the term frequency $tf_{(i)}$ will not be a much larger value compared to $F_{(i,j)}$ and the ratio $FR_{(i,j)}$ will be large for j^{th} group. If the i^{th} word is a stop word or a non-related word, $F_{(i,j)}$ doesn't have much variation among the groups. Therefore, the term frequency $tf_{(i)}$ will not be closer to any $F_{(i,j)}$ value, but approximately equals to $F_{(i,j)} \times j$ and the ratio, $FR_{(i,j)}$ will be small for all groups. In short, the distribution for a keyword is not common to all groups but the distribution for a stop word is almost common for all groups.

5.3.4 Evaluating the new feature selection method

The evaluation was carried out in order to measure the effectiveness of the suggested method. Effectiveness is purely a measure of the ability of the system to satisfy the user in terms of the relevance of short messages retrieved. It is assumed that the more effective the system, the more it will satisfy the user. The effectiveness of the retrieval system was measured using precision, given in equation 5.9, and recall, given in equation 5.10, values [25] (which was explained in 3.1 and 3.2 equations). Precision is the fraction of retrieved short messages that are relevant. Recall is the fraction of relevant short messages that are retrieved [26].

$$Precision = \frac{tp}{tp + fp} \tag{5.9}$$

$$Recall = \frac{tp}{tp + fn} \tag{5.10}$$

The dataset was divided into 3 parts, data for feature selection, for training and for testing. Thus, the feature selection, training and testing process are independent from each other. Therefore, it can ensure that the biasness of the data was removed.

The Ratio method (new method) was compared with 4 popular feature selection methods. They are, Sequential Forward Selection, Sequential Backward Elimination, Information Gain and Chi square. Table 5.1 shows the Precision and the recall values for the group accident and the number of features which was selected by each method. Both Forward Selection and Backward Elimination had chosen 13 features. Information Gain and Chi square methods had chosen 49 features and the Ratio Method had chosen 270 features. These results are based on 3 months results. But to have a successful system which provide best results for long time period, it is important to select best feature set, even the dataset is bit large. Ratio method helps to extract all the possible features, and the actual effect of the bigger dataset can be observe after long time usage.

5.4 Summary

This chapter briefed out the feature extraction and feature selection methods which were used for the classification of the current research. The research was about Twitter news classification for local news providers. Thus, 5 active news providers were chosen to gather the data. The data were Twitter short messages. The hyperlinks were removed from the short messages. For the classification stage the bag-of-words method was used to extract the features

Classifier	Feature selection method	No of features	Precision	Recall	TP rate	FP rate
J48 decision tree	Sequential Forward Selection	13	0.714	0.321	0.321	0.008
	Sequential Backward Elimination	13	0.714	0.321	0.321	0.008
	Information Gain	49	0.844	0.346	0.346	0.004
	Chi square	49	0.844	0.346	0.346	0.004
	Ratio method	**270**	**0.868**	**0.387**	**0.387**	**0.003**
SVM	Sequential Forward Selection	13	0.826	**0.487**	**0.487**	0.006
	Sequential Backward Elimination	13	0.826	**0.487**	**0.487**	0.006
	Information Gain	49	0.969	0.397	0.397	0.001
	Chi square	49	0.969	0.397	0.397	0.001
	Ratio method	**270**	**0.977**	0.361	0.361	**0.000**
Random Forest	Sequential Forward Selection	13	0.792	0.487	0.487	0.008
	Sequential Backward Elimination	13	0.792	0.487	0.487	0.008
	Information Gain	49	**0.939**	0.59	0.59	**0.002**
	Chi square	49	**0.939**	0.59	0.59	**0.002**
	Ratio method	**270**	0.935	**0.689**	**0.698**	**0.002**
Random Trees	Sequential Forward Selection	13	0.97	0.410	0.410	0.001
	Sequential Backward Elimination	13	0.97	0.410	0.410	0.001
	Information Gain	49	0.885	0.59	0.59	0.005
	Chi square	49	0.885	0.59	0.59	0.885
	Ratio method	**270**	**0.988**	**0.689**	**0.689**	**0.000**
Simple CART	Sequential Forward Selection	13	0.792	0.487	0.487	0.008
	Sequential Backward Elimination	13	0.792	0.487	0.487	0.008
	Information Gain	49	**0.897**	0.449	0.449	**0.003**
	Chi square	49	**0.897**	0.449	0.449	**0.003**
	Ratio method	**270**	0.822	**0.504**	**0.504**	0.006

Table 5.1: Evaluation methods for Feature selection for the group accident

from the short messages. The selected feature dimension was too large and therefore, it was likely to cause a complex classifier. Thus, a feature selection technique was used. The available techniques do not perform well in Twitter short messages as Twitter restricts the number of words per message. Thus, a new feature selection method was introduced. The classification was done using these new features. The next chapter will brief out the classification technique which was used for Twitter news classification.

Chapter 6

Classification Process

6.1 Overview

Previous chapter had described the feature selection process for the classification process. Once the features were selected, the next step was to train the data using a suitable classifier. There are several classification techniques which can be used for this text classification. Researcher had tested the following classifiers in order to obtain the best classifier. This chapter describes and compares the efficiency of each classification.

6.2 Creating the dataset

The collected data set contains 3600 messages. This set was divided into three sets. Feature selection data, training data and testing data. Record number 1-1000 was chosen for feature selection. Record number 1001-2500 was taken for training and 2501-3600 was taken for testing. There were 12 pre-defined class labels. All training and testing data would be pre-labeled into their classes manually in order to evaluate the classifiers. Thus, using the training data, the classifier was built and the performance can be tested using test data. The dataset was manually classified by research colleagues and researcher verified samples.

6.2.1 Training using Naive Bayes Algorithm

Naive Bayes algorithm is a text classification technique which was based on Bayes Theorem. It was a simple and fast classifier. As we are dealing with word frequencies, multinomial model of the Naive Bayes was chosen to classify the text [15]. However, Naive Bayes algorithm follows a set of assumptions and the result will be totally dependent on those assumptions. The main assumption is that all features are needed to be independent. Rennie et al. [14] shows that the words are not independent from each other. Thus, it will cause to generate poor results from Naive Bayes classifier.

6.2.2 Training using SVM

SVM is a newly developed classifier which can be used for text classification. It was introduced by Vapnik and Colleagues. There are many Kernel functions which can be used to convert the linearly inseparable dataset into a linearly separable dataset [21]. For the current research, we had used the Radial Basis Function as the Kernel where the gamma value is 0.1. SVM has its own ability to deal with high dimensional data. It is capable of reducing the complexity and therefore, avoids the over fitting of the classifier.

6.2.3 Training using Random Tree

Random tree is a tree which was constructed using K number of attributes for each node. For the current research, we had used $log_2(numberOfAttributes) + 1$ as the K value. Random Tree has an option to allow estimation of class probabilities based on a hold-out set (back fitting). However, for the current research, the researcher has not used back fitting data as the data are not time related.

6.2.4 Training using Random Forest

Random Forest is the ensemble version (bagging) of Random trees. For the current research, 20 Random Trees were used where each tree consider *K* number of attributes. Lesser number of trees gets low accuracy and the accuracy gets saturated when the number of trees is equal to 20. The *K* value can be calculated as $log_2(numberOfAttributes)+1$.

6.3 Evaluation

The best training method was detected by identifying the most efficient classifier. The efficiency can be measured by calculating the *Precision* and *Recall* values. However, it is not easy to compare a set of classifiers using two calculated values. Thus, it is essential to calculate a single value using precision and recall. Thus, *harmonic mean* can be used as the single value to evaluate the classifiers.

$$Precision = \frac{|Rlevnt \cap Retrievd|}{Rtrivd} \tag{6.1}$$

$$Precision = \frac{|Rlevnt \cap Retrievd|}{Rtrivd} \tag{6.2}$$

$$Fscore = \frac{recall \times precision}{\frac{(recall+precision)}{2}} \tag{6.3}$$

Table 6.2 shows the results of the evaluation. According to the results, it is clear that *Random Forest* classifier performs better than other classifications. The result can be discussed as follows.

Group	Classifier	Precision	Recall	F-Measure
Accident	Naive Bayes	0.939	0.286	0.438
	SVM	0.567	0.733	0.640
	Random Tree	0.717	0.646	0.680
	Random Forest	0.759	0.646	0.698
Development and Government	Naive Bayes	1.000	0.068	0.127
	SVM	0.400	0.359	0.378
	Random Tree	0.510	0.295	0.374
	Random Forest	0.524	0.299	0.381
Disaster and Climate	Naive Bayes	0.000	0.000	0.000
	SVM	0.208	0.333	0.256
	Random Tree	0.500	0.091	0.154
	Random Forest	0.556	0.152	0.238
Education	Naive Bayes	1.000	0.309	0.472
	SVM	0.743	0.642	0.689
	Random Tree	0.788	0.710	0.747
	Random Forest	0.852	0.710	0.774
Entertainment	Naive Bayes	0.000	0.000	0.000
	SVM	0.514	0.603	0.555
	Random Tree	1.000	0.571	0.727
	Random Forest	0.889	0.635	0.741
Health	Naive Bayes	1.000	0.118	0.211
	SVM	0.439	0.691	0.537
	Random Tree	0.979	0.676	0.800
	Random Forest	0.978	0.662	0.789
International	Naive Bayes	0.960	0.053	0.100
	SVM	0.552	0.574	0.563
	Random Tree	0.930	0.585	0.718
	Random Forest	0.884	0.604	0.719
Politics	Naive Bayes	1.000	0.186	0.313
	SVM	0.627	0.713	0.668
	Random Tree	0.974	0.724	0.831
	Random Forest	0.949	0.719	0.818
Society	Naive Bayes	0.750	0.010	0.020
	SVM	0.453	0.509	0.479
	Random Tree	0.949	0.505	0.659
	Random Forest	0.859	0.519	0.647
Sports	Naive Bayes	0.955	0.592	0.731
	SVM	0.865	0.869	0.867
	Random Tree	0.977	0.871	0.921
	Random Forest	0.969	0.869	0.916
War Terrorism and Crime	Naive Bayes	0.866	0.424	0.569
	SVM	0.715	0.717	0.716
	Random Tree	0.973	0.740	0.841
	Random Forest	0.942	0.744	0.831

Table 6.1: Evaluation of the classification methods

6.3.1 Comparing Naive Bayes Algorithm with SVM

As the result of Table 6.2, it is clear that SVM performs better than Naive Bayes Algorithm. Dilrukshi et al. [27] had given the theoretical and practical reasons as to why SVM performs better than Naive Bayes method. The reason behind this can be described as follows. Naive Bayes Algorithm follows Bayes Theorem. However, Naive Bayes Theorem follows several assumptions. The main assumption is, the assumption of independence among variables. However, we cannot predict that words of a sentence are independent from one another. Thus, this assumption may get violated.

SVM is a newly developed algorithm which is capable of dealing with high dimensional data. Even though we choose bag-of-words method instead of n-gram method, still it results in a large dimension dataset. The feature selection method can reduce the number of attributes up to certain level. However, a text classification research still contains a large dimension. Thus, SVM can produce accurate results.

6.3.2 Comparing SVM with Random Tree

Even though SVM can deal with high dimensional data, the results show that Random Trees (Weka) provide better results than SVM. The technique behind SVM is, they choose marginal data points to draw the hyper plane between two datasets. The technique behind Random Trees is, it chooses random number of attributes per time and build the decision tree without pruning. This will let all features to contribute to the tree. This proves that that the technique used for Random Trees was more powerful than the technique which was used for SVM.

6.3.3 Comparing Random Trees with Random Forest

Random Forest is the bagging version of the Random Tree method. Therefore, the maximum vote will be taken when assigning it to the group. Thus, it is provable that using Random Forest can provide better results than using Random Trees. The results show that Random Forest provides the best results for classifying the news into groups.

6.4 Further analyzing with Random Forest

Even though Random Forest was proven as the best classifier, the parameters of Random Forest can make a large impact to the result. The parameters are as follows.

- The maximum depth of the trees.
- The random number seed to be used.
- The number of execution slots (threads) to use for constructing the ensemble.
- The number of trees to be generated.

Classifier 1	Accident: others	161:3177
Classifier 2	Development:others	251:3087
Classifier 3	Disaster:others	33:3305
Classifier 4	Economy:others	230:3108
Classifier 5	Health:others	68:3270
Classifier 6	International:others	455:2883

Table 6.2: Ratio between classes for each classifier

Number of trees	Accident	Development	Disaster	Economy	Health	International
10	0.676	0.384	0.279	0.521	0.561	0.463
20	0.698	0.404	0.293	0.525	0.589	0.463
30	0.686	0.401	0.333	0.517	0.591	0.454
40	0.682	0.397	0.333	0.513	0.596	0.454
50	0.661	0.386	0.333	0.522	0.595	0.456

Table 6.3: Changing the number of attributes per tree for groups

- The number of attributes to be used in random selection.

We never restricted the maximum depth of a tree. Restricting the depth will help to increase the speed by reducing the complexity but it will decrease the accuracy. Since we pay more attention to the accuracy, we kept the depth of a tree as unlimited.

The number of seeds will be used by Baggings seeded generator for resampling/randomizing the train set. The default value is 1. We tested this with several other values and figured out that increasing the value will cause to increase the speed but reduce the accuracy. Thus we used the default value.

The number of execution slots (threads) can increase the speed of the classification process. The default value is 1. However, depending on the performance of the computer, there is a limit for increasing the number of execution slots. The large numbers will crash the process. Therefore, we used 2 slots which performed well for the computer we used.

Table 6.2 shows the ratio between classes for each classifier. According to the table, it is clear that ”Disaster” classifier and ”Health” classifier had small ratios compared to other classifiers. According to Table 6.3, it is clear that ”Disaster” classifier and ”Health” classifier requires a greater number of trees in order to get high accuracy. For other classifiers, 20 trees results in high accuracy. Thus, the researcher used more number of trees for the classifiers which has less number of ratios and less number of trees for the classifiers which has large number of ratios.

The researcher had tested the random number of attributes. Breiman [17] suggests the default value as $log_2(numberOfAttributes) + 1$. For current situation, the default value will be 9. Table 6.4 shows the F measures which were generated for different number of attributes. The accuracy was tested for selected groups. According to Table 6.4, it is clear that the number of attributes for the highest accuracy were distributed around 9. Thus, the default number of attributes was chosen.

Number of Attributes	Accident	Development	Disaster	Economy	Health	International
2	0.671	0.402	0.200	**0.560**	0.544	**0.460**
4	0.660	0.403	0.238	0.535	0.574	0.452
6	0.662	0.393	0.190	0.538	0.584	0.470
8	0.678	0.390	0.233	0.521	0.591	0.465
9 (Default Value)	0.676	0.384	**0.279**	0.521	0.561	0.460
10	0.682	0.403	0.238	0.519	0.614	0.451
12	**0.693**	0.385	0.195	0.525	0.591	0.455
14	0.669	**0.404**	0.233	0.529	**0.626**	0.461
16	0.671	0.402	0.273	0.518	0.591	0.451
18	0.687	0.394	0.233	0.518	0.559	0.453
20	0.691	0.404	0.238	0.512	0.581	0.448
22	0.682	0.397	0.267	0.517	0.605	0.442
24	0.678	0.396	0.233	0.512	0.600	0.451
26	0.686	0.393	0.238	0.509	0.569	0.446
28	0.680	0.391	0.267	0.505	0.569	0.446

Table 6.4: Changing the number of attributes per tree for groups

6.5 Manage a new category

The current system uses predefined groups to categorize the news and these groups were defined based on the current situation of the world. There were 12 predefined groups which were used in this system. These groups would not change regularly so using the same data set and same features for about 10 years would not make much of an issue. However, when using these groups for 50 years or 100 years, we cannot assume that the world has not change and that there will be no more interesting new news groups. Thus, we suggest a concept to identify a new news group and to manage news with that group. This suggested concept will be actively run for every 10 years to identify whether there is any new category of news.

6.5.1 Identify a new category

In order to manage a new category, the first step is that the system should be able to identify that there was a new category. As we are doing a binary categorization, there will be some short messages which do not belong to any group. This will be the first signal that there will be a new group. The suggesting concept is to analyze these unclassified short messages and extract new keywords from those short messages. The system uses the same feature extraction method, feature selection method and stemming process that were used previously. Using these keywords, the system will cluster the short messages (the clustering process will be same as in chapter 7) and if one group gets more than 75% of short messages, the system assumes that there is a new category.

6.5.2 Naming the new category

Once the system identifies that there is a new group created, the next step is to name the new group. The system uses the selected keywords in order to create the new name. The most frequent word will be used as the name of the category.

6.5.3 Creating the new Training Data sets

The system had already created training data sets for the pre-defined 12 groups. However, the newly recognized features had to be merged to the feature list. Therefore, new datasets for the pre-defined groups will be created using the new feature list. In order to create the data set for the new recognized group, the system should identify the training categories for the new news group. As the manual data categorization was impossible at this stage, the system will use the clustered results to create the new dataset. The short messages which belonged to the new cluster will be tagged as the short messages that belong to the new group and other short messages will be classified as Others.

6.5.4 Classification the data further

With the newly recognized group, there will be 13 news groups. The newly extracted short messages will be classified into these 13 groups, as it classified into 12 groups. The new feature list will be used as the features of the classification method.

6.6 Summary

This chapter describes the classification process of the research. The data was divided into 3 parts: for feature selection, to build the classifier and to test the classifier. The classifiers, Naive Bayes, SVM, Random Trees and Random Forest were selected to classify the situation. These classifiers were evaluated using F measure. The results show that SVM performs well than Naive Bayes, and Random Trees performs well than SVM. Random Forest performs well than Random Trees. As Random Forest is the bagging version of Random Trees, this performance can be predetermined. Thus, we had to use Random Forest to classify the system. There are set of parameters which we need to set while using Random Forest. For most cases, the default value can be used. However, when classifying into 2 classes, if the ratio among the classes has a large difference, using more trees will increase the accuracy. A concept was suggested to identify and analyze new news groups, which can be introduced later, with time.

By classifying the data, one can measure the number of news per group. However, for further analyzing, one may need to get more popular news topics rather than a popular news group. A clustering technique can be used to get the news topics from a set of news. Thus, the next chapter will describe the clustering techniques which were used for the research.

Chapter 7

Clustering Process

7.1 Overview

Previous chapter described the classification process for news headlines. It resulted in classifying of news into 12 groups. For further analyzing, one may need to view the popular news headlines. These clustering techniques will be used to cluster the news into news topics.

7.2 Feature Selection

Unlike the classification method, it is not easy to use a pre-defined feature for clustering. The classification method classifies the text into pre-defined 12 groups. The clustering method needs to cluster the news into their topics. This means, it is a further more explanation than classification. Thus, an information loss can harm the results terribly. Therefore we generate**d** the features real time and used it for clustering.

The feature extraction was done using bag-of-words method. We did not use n-gram method in order to avoid the high dimension issue. Using these extracted features, we had to select the best features. Thus, we removed the stop words manually. Then the next step was to remove the noise words and remaining common words. To remove the noise words, we ordered the words according to their frequencies and considered the frequency change from each word to word. We figured out that there is a statistically significant difference from the number of words which has a frequency 4 and the number of words which has a frequency 5. Thus, for this sample, we selected the lower frequency cut off value as 5. Table 7.1 shows the highest frequencies which were obtained. It is clear that there is a significant difference of the frequencies from the word CHOGM. Thus, we decided the upper cut off value as 65 for this sample, and to remove the words which have a frequency greater than 65. This resulted in 269 features. We had used these features for the clustering. Because we are doing clustering for each sample, these cutoff values need to be set for each sample. Therefore, the significant cutoff value for each sample will be defied statistically.

Word	Frequency	Difference
.	.	.
.	.	.
.	.	.
.	.	.
India	50	2
video	53	3
lankan	56	3
colombo	58	2
arrest	65	7
chogm	**65**	0
On	**108**	**43**
Itn	122	14
New	168	46
.	.	.
.	.	.
.	.	.

Table 7.1: highest frequencies

7.3 Clustering process

The data was clustered using 2 clustering methods. Expectation Maximization (EM Clustering) clustering and Hierarchical Clustering.

7.3.1 Clustering using EM Clustering

The data was clustered using EM Clustering. Weka library was used for the implementation. The number of maximum iterations was set to 10000. The minimum standard deviation was set as 1e-12. In order to increase the accuracy, if a cluster contain more than 10 short messages, we performed another clustering for that particular cluster. The clustering technique clusters it further if possible, or keeps it in one cluster.

7.3.2 Clustering using Hierarchical Clustering

We had to perform Hierarchical clustering for the dataset. Weka Hierarchical clustering library was used for the implementation. We set the Euclidean Distance as the distance function. The Single Linkage was used as the linkage function because, when clustering 2 topics, there would be only a few connections between the two topics. The linkage function would be able to detect such small links. Simple Linkage Function can detect small links as it considers the smallest distance between 2 clusters. As mentioned in EM Clustering, we performed re clustering if the cluster size was larger than 10.

7.4 Evaluation

The evaluation was a challenging point of this clustering. Since we did not know the cluster, we could not use precision or recall. Thus, we used a manual process for the evaluation. The process is as follows. We set an evaluation panel to evaluate the clusters. The Researchers who work at the

Gas Leak Injures Over 70: More than 70 persons were injured and hospitalized due to a gas leak 72 hospitalized following gas leak in Piliyandala
Factory temporarily shut down following gas leak

Figure 7.2: News regarding a gas leak-Hierarchical Clustering result

Gas Leak Injures Over 70: More than 70 persons were injured and hospitalized due to a gas leak **Child Molester Arrested And Fined: A child molester arrested by the Slave Island Police** 72 hospitalized following gas leak in Piliyandala
Four In Serious Condition After Burn Injuries: A man had set fire to his mistress
Bulgarian convicted in suffocation deaths of 18 Lankans arrested
Factory temporarily shut down following gas leak

Figure 7.3: News regarding a gas leak-EM Clustering Result

laboratory were selected as the panel of evaluation. The evaluation panel read the news clusters and identified the mismatch twits manually. The percentage of corrected twits were considered as the accuracy of the clustering method.

According to their evaluations, EM clustering performs with 68.52% accuracy and Hierarchical clustering performs with 89.93%. Figure 7.2 is an example of a result of Hierarchical clustering. The same situation which was result by EM Clustering was given in Figure 7.3. It is clear that EM clustering tries to merge the situations while Hierarchical clustering provides an optimal result.

However, Hierarchical clustering do have several issues. According to Table7.4, it is clear that due to the ambiguity of the word fire, "No fire zone: Commonwealth rehabilitating Sri Lankan regime - Channel 4 News (blog)" was clustered as an accident fire.

The advantage of using Hierarchical clustering with single linkage is, in most of the cases, it can recognize the small connection between 2 clusters which need not be merged. According to Table7.5, both words CHOGM and summit was used to refer the same situation, Commonwealth Summit. Hierarchical Cluster has been able to identify the relationship between 2 sets and had merged together.

7.5 Summary

This chapter describes the clustering part of the research. It clustered the news into news topics. The bag-of-word approach was used for this clustering. The stop words were removed from the feature set. Then the remaining common words were removed by re- moving high frequent words. The noise words were removed by removing low frequent

Man Dies After Setting Fire To Forest: An individual who had set fire to a forest in the Horowpathana Jaffna-bound private bus gutted in fire
Fire At Abandoned FTZ Building: A fire erupted at an abandoned building
No fire zone: Commonwealth rehabilitating Sri Lankan regime - Channel 4 News (blog)
80-year-old dies in Avissawella house fire
Man Killed After House Catches Fire In Avissawella: A fire has erupted in a house covered in polythene sheets and

Figure 7.4: News regarding a Fire

Colombo gets a facelift ahead of high-profile summit - Khabar South Asia Despite opposition, PM likely to go to Sri Lanka for CHOGM summit - IBNLive Bangladeshi PM may attend Commonwealth summit
CHOGM summit in Sri Lanka faces heat from International rights group - IBNLive
CHOGM summit in Sri Lanka faces heat from international rights groups - IBNLive Sri Lanka Commonwealth summit defended - BBC News
Commonwealth SG defends holding summit in Sri Lanka

Figure 7.5: News regarding Commonwealth summit

words, resulting in 269 features. Two clustering techniques, EM clustering and Hierarchical clustering, were used to select the best clustering. The results prove that Hierarchical clustering provides best results with the accuracy of 89.93%. The Euclidean Distance was used to get the distance between 2 data points and Simple Linkage was used as the linkage function which will define as to how to use the distance. The next chapter will brief out the discussion of the research.

Chapter 8

General Discussion

8.1 Overview

This chapter briefs out the general discussion of the research. The research contains mainly two parts: classifying the news into predefined groups and clustering the news into news topics. Thus, if one needs to identify the most popular news area in a given time, he/she can use the classified groups and if one needs to identify the most popular news topic, he/she can use the clustered news. The results, implementation issues, findings and conclusion of the 2 parts will be briefed out in this chapter.

8.2 Discussion

With the development of technology, now-a-days, many people tend to collaborate with the internet and World Wide Web (WWW). Thus, many organizations tend to share their news and useful information in blogs and social networks. News providers are one such common and useful organization types and Twitter is one such common social network, which is also known as a common microblog. By considering the amount of news shared in Twitter, it can be a source for an ”information generator”. Thus, the aim of this research is to develop a tool which is capable of organizing this news in a useful manner.

The researcher used Twitter microblog for data gathering, because it allows gathering and accessing the short messages, which are commonly named as Tweets, which are shared publicly. News were selected using five news providers: Ada Derana, Ceylon Today, ITN, Lanka Breaking News and News First. Twitter has a feature - the restriction of character length, which can be considered as an advantage and a disadvantage. The advantage is, it is easy to extract the main idea of the news. The disadvantage is, it is hard to remove the common words from the feature set. In this research, the researcher had used the advantage of the character restriction properly and had provided a solution for disadvantage of character restriction.

While planning to organize the news, the most important point was to identify the usefulness of the information. We had identified that there would be two major ways to

display the information. One was to give a count of the amount of news which belonged to a given category. The other was to Cluster the news according to their topics and display the clusters using keywords. Thus, the researcher used a classification method, in order to classify them into known groups and the researcher used a clustering method in order to cluster them into news topics.

8.2.1 Classify the news into pre-defined groups

News classification for pre-defined groups would be useful for one who was interested in the number of news reported towards each topic, such as accidents, education, health etc. For classification, some training data and testing data were needed to be created. Thus, the dataset was tagged into groups manually.

The most important factor here was the feature extraction and the feature selection. The feature extraction was done using bag-of-words methods. There were other alternative methods such as n-gram method. However, the researcher used bag-of-words method because, n- gram method leads to the increase of dimension of feature set.

Even though we used bag-of-word method, still the feature set was high dimensional. Thus, a feature selection method was required to be conducted. There are several feature selection methods such as forward selection method, backward elimination method, term frequency etc. Those methods can be useful for situations such as document classification. The difference between twitter news classification and document classification is that, twitter has a character length restriction. Due to this restriction, it is hard to detect the difference between common words and keywords.

Thus, a new method was needed to create for feature selection, where the new method considers how proportionally a given word belongs to a given group with respect to other groups. In that case, if a word is a common word, where it does not take significantly high frequency to a particular group, it may occur in the message with high frequently or low frequently, the ratio of the word frequency towards each group was nearly same.

In other cases, if a word is a keyword of a given group, the ratio of the word frequency toward the group is not similar. There will be one group where the given word occurs frequently and other groups may have low frequency. In that case, that word is a feature of the given group. Therefore, the features can be easily identified. This new method was named as Ratio Method and for the given situation, it had chosen 270 features.

Once the researcher selected the features, the next step was to classify the news. The researcher had tested the situation using four classifiers: SVM, Naive Bayes, Random Trees and Random Forest. Random Forest, which is the ensemble method of Random Trees, provides the best result. SVM may perform better than Naive Bayes because of the independent assumptions which Naive Bayes consider. Random trees may perform better than SVM because; Random Trees use all features and all instances to classify the situation. SVM uses only marginal data points to classify the situation. This may cause to have a low accuracy when using SVM. When comparing Random Trees and Random Forest, it is obvious that Random Forest is the ensemble

method of the Random Tree. Thus, it tends to give more accurate value.

8.2.2 Clustering the news into news topics

There could be some situations where the users do not pay much interest on the count of the news which was included in a given group, but is interested on the popular news headlines. For such situation, the clustering technique can be used.

We had used bag-of-words method to extract the features. For the clustering method, we cannot pre define a set of features. The features should generate in real time. Thus, the features which were used for classification were unable to use. Therefore, we defined a new feature set which can be generated real-time.

In order to generate the new feature set, all the stop words were removed from the extracted word list. Then the noise words were removed from the word list by removing the low frequent words. The common words were removed by removing the most frequent words. The low cut off word frequency was defined as 4 and the upper cut off word frequency was defined as 65. The result is 269 features.

There are several clustering techniques available and EM clustering and Hierarchical clustering were popular among them for text clustering. For the current research, we had used EM clustering and Hierarchical clustering to test and obtain the best clustering method.

The results show that Hierarchical clustering performs well than EM clustering. The accuracy of hierarchical clustering is 89.93% and the accuracy of EM clustering is 68.52%. Even though EM clustering performs well, Hierarchical clustering has the ability to detect small connections between two clusters. When using hierarchical clusters with simple linkage, it considers the smallest distance between the clusters. Thus, it is capable of detecting small connections between 2 clusters.

8.3 S2Net Tool

S2Net is an online tool which was built with these findings. This system allows user to detect the most popular news group in a given time frame. The techniques which were explained in chapter 6 were used to obtain this part of the system. The results of the classification method will be displayed in a pie chart. As there are 12 number of groups, the most popular 5 groups will be shown in the pie chart separately. The count of others will display as ”Others”.

Figure8.1 shows the web site of the created online tool. We have to click on ”Analyze” tab to start the analyzing. As in Figure8.2, we have to set the time duration which we require to analyze. Once you submit the time, you will get a link as shown in Figure8.3. If the selected date is correct, click next. Then the percentage for all 12 groups will be displayed as Figure8.4 and the pie chart will display as in Figure8.5.

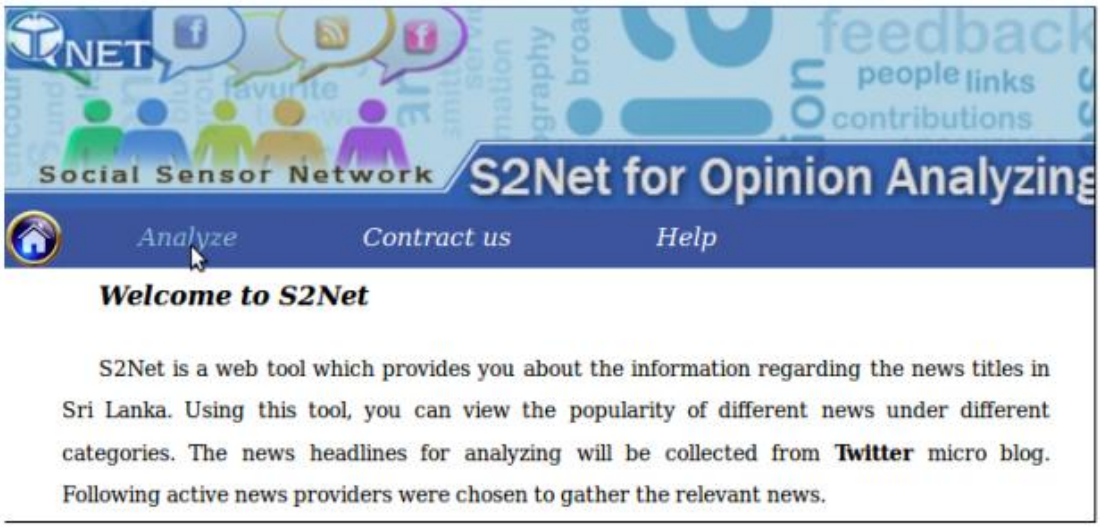

Figure 8.1: S2Net Tool

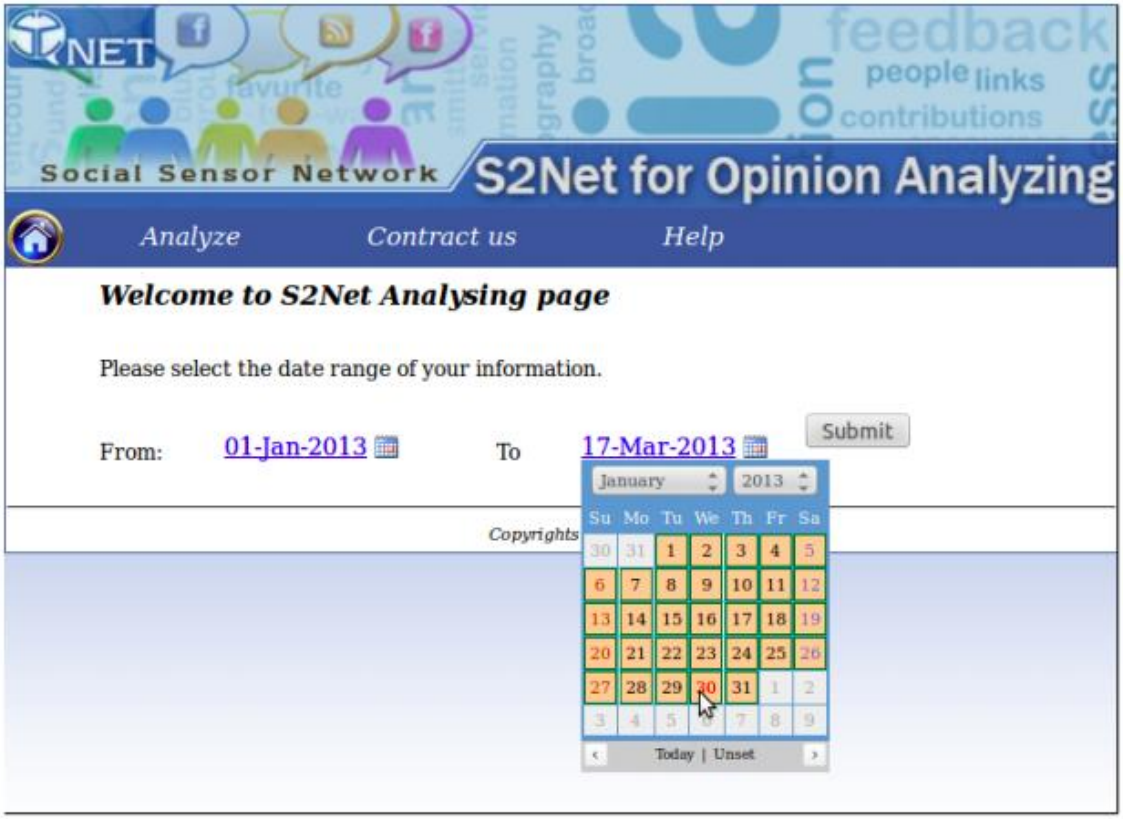

Figure 8.2: Select time range

If a user requires details about news topics, the system allows user to display the news keywords. The clustering techniques which were explain in chapter 7 were used to obtain this part of the tool. The cluster name was created using the key words. Once the clusters had created, the frequency for each keyword within the cluster was calculated. The highest frequent keyword was

used as the cluster name. The font size will be proportional to the popularity of a cluster, which

Figure 8.3: Check time range

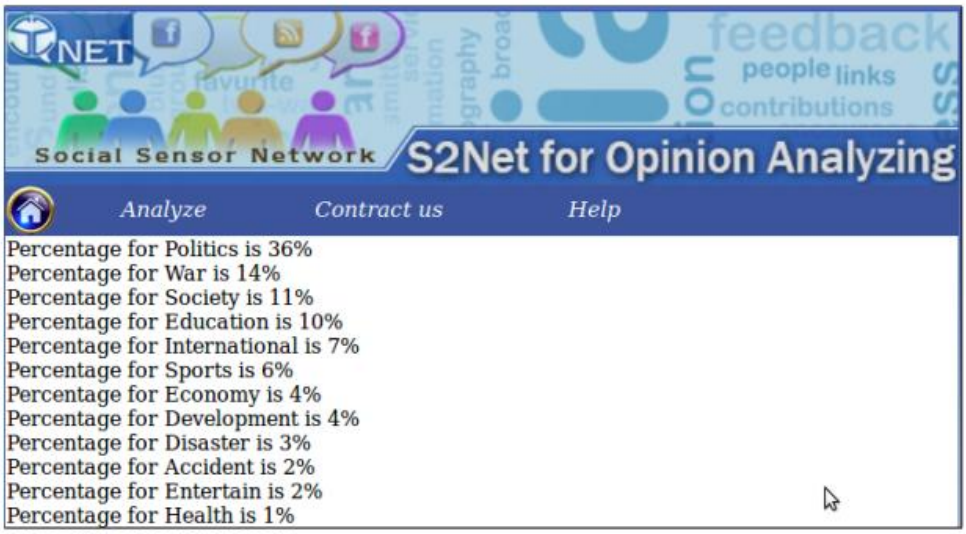

Figure 8.4: Clusters with percentage

means, the cluster size. The keywords of high cluster size will show in a large font size and keywords of low cluster size will show in a small font size. In order to do that, the user will have to click the link which is at the bottom of the pie chart as given in Figure 8.6. Then, it will direct you to the further analyzing page.

You have to set the date for further analyzing. Once you submit the date, the date will appear. Once you click the next button, you can see the cluster numbers with different font size as given in Figure 8.7. The font size will represent how many news include in the given cluster. Once you click a cluster, the content will display as in Figure 8.8.

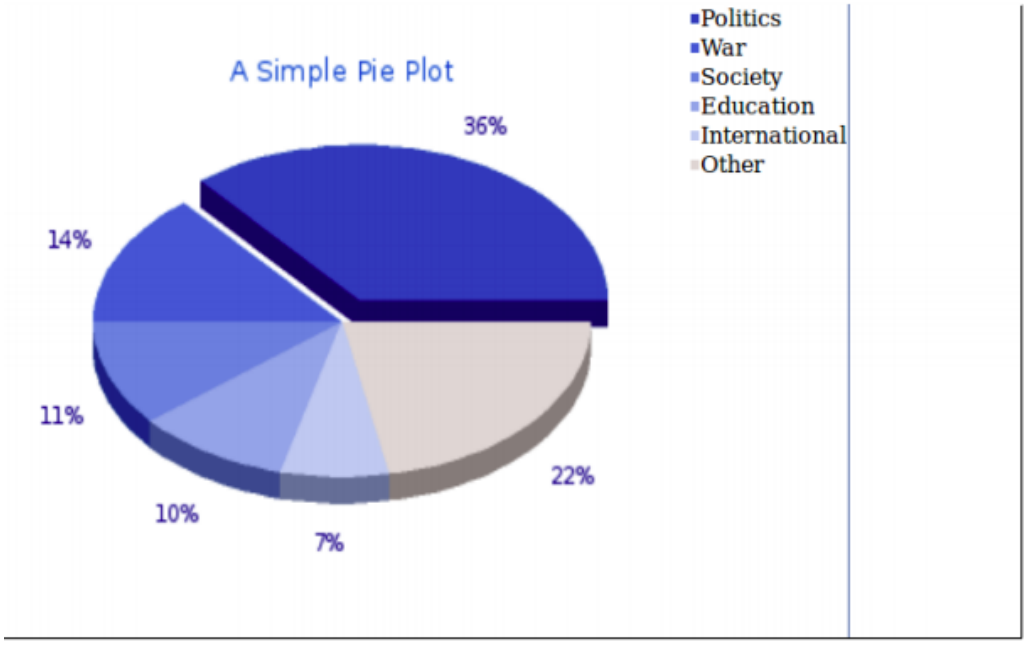

Figure 8.5: Cluster pie chart

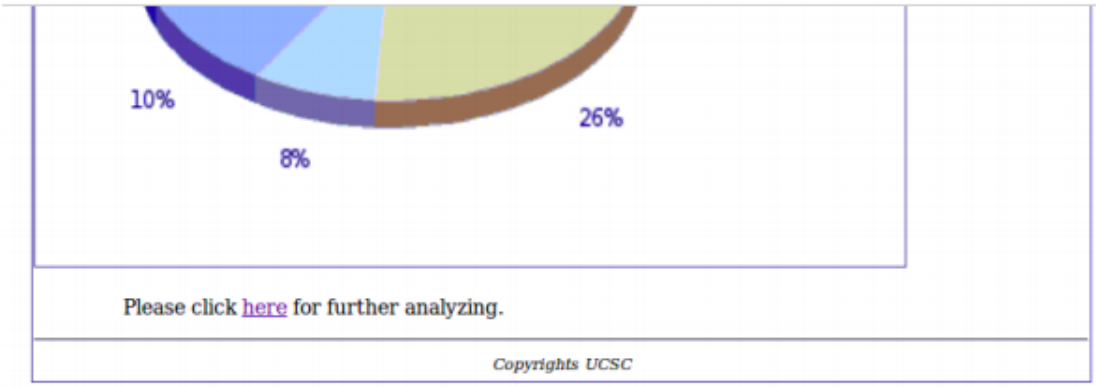

Figure 8.6: Link for further analyzing

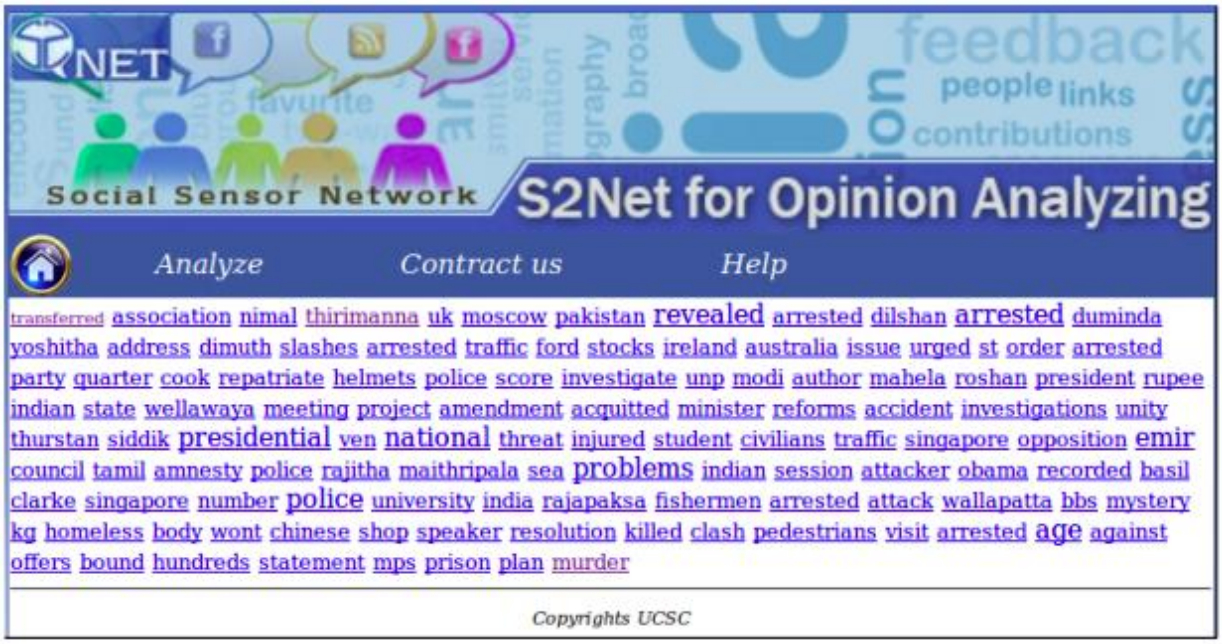

Figure 8.7: Result of Clustering

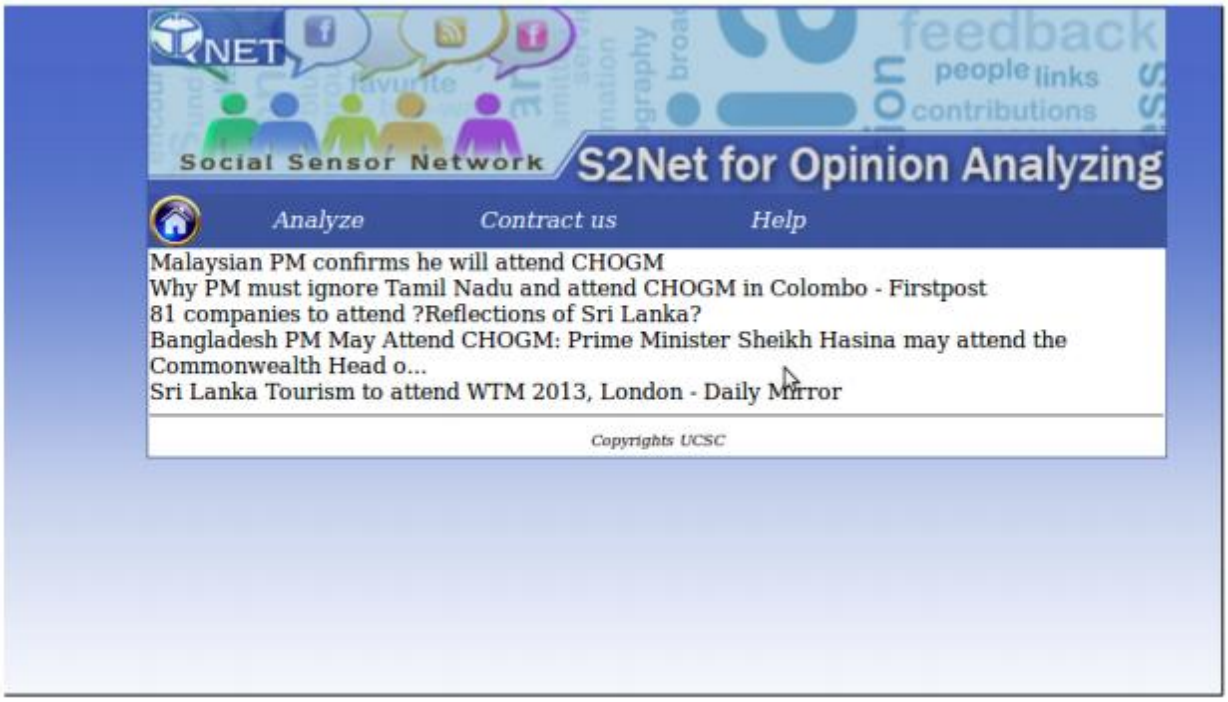

Figure 8.8: Cluster details

8.4 Conclusion

The conclusion and the findings of the research is as follows:

- When extracting features from a document, bag-of-words method carry less amount of sufficient features than n-gram method. Thus, using bag-of-word method will help to avoid complex models and therefore, avoids over fitting.

- In most of the situations, the low frequent words are noise words and high frequent words are common words. Thus, removing the low frequent words and high frequent words will cause to reduce the dimension

- When classifying the news, the keywords are constant. Thus, we can define a set of keywords. For this, we introduce**d** a new feature selection method because; existing methods do not work well for Twitter short messages, due to the character length restriction.

- The new method is called as Ratio Method and it considers the importance of a word towards a given group. However, still the result will be a sparse matrix. Thus, we need to use a classifier which is capable of handling sparse matrix.

- When clustering the news, the keywords are subject to change. Thus, it is impossible to pre define a feature set. Therefore, we had provided a list of stop words which are needed to remove from the feature set.

- For classification, 4 classifiers: Naive Bayes Algorithm, SVM, Random Trees and Random Forest, were used. The results show that Random Forest, which is the embedded version of Random Trees, performs better than other classifiers.

- We used harmonic mean (F-measure), to calculate the efficiency of classifiers.

- In order to cluster the news into topics, we tested EM clustering and Hierarchical clustering. The results show that Hierarchical clustering performs well than EM clustering

- In Hierarchical clustering, the linkage function is the most important feature to select. The results show that simple linkage function performs better than other linkage functions. The reason is because, there will be a very small connection between news and the clustering technique should be able to detect such small connections. Simple Linkage is capable to recognize such small connection.

8.5 Further suggestions

The S2Net analyses the news and briefs out in a more descriptive manner. However, to get the sentimental idea of the news, we have to read the news. A further modification can be done as follows. By analyzing the comments of each news, one can get the sentimental idea as, whether it is good news or bad news. Thus, for a group like education, even though it become more popular in a given time period, we can refer to the analyzed results of the comments and without reading the news, we can get an idea about whether it is good news of education or bad news of education.

Bibliography

[1] I. Dilrukshi, K. De Soysa, and A. Caldera, "Twitter news classification using SVM," in *Computer Science & Education (ICCSE), 2013 8th International Conference on*. IEEE, 2013, pp. 287–291.

[2] A. Basu, C. Watters, and M. Shepherd, "Support vector machines for text categorization," in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 4 - Volume 4*, ser. HICSS '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 103–3.

[3] I. Witten, "Text mining," *Practical handbook of Internet computing. CRC Press, Boca Raton, FL*, 2004.

[4] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, ser. Lecture Notes in Computer Science, C. N. dellec and C. l. Rouveirol, Eds. Springer Berlin / Heidelberg, 1998, vol. 1398, pp. 137–142.

[5] M. Alexa and C. Zuell, "Text analysis software: Commonalities, differences and limitations: The results of a review," *Quality & Quantity*, vol. 34, pp. 299–321, 2000.

[6] D. Giorgetti, I. D. Linguistica, and F. Sebastiani, "Automating survey coding by multiclass text categorization techniques," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 1269–1277, 2003.

[7] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of LREC*, vol. 2010, 2010.

[8] B. Yu, "An evaluation of text classification methods for literary study," *Literary and Linguistic Computing*, vol. 23, no. 3, pp. 327–343, 2008.

[9] M. F. Porter, "An algorithm for suffix stripping," *Program: electronic library and information systems*, vol. 14, no. 3, pp. 130–137, 1980.

[10] H. Bacan, I. S. P, and D. Gulija, "Automated news item categorization," in *Proceedings of the 19th Annual Conference of The Japanese Society for Artificial Intelligence*. Springer-Verlag, 2005, pp. 251–256.

[11] S. H. Lin and J. M. Ho, "Discovering informative content blocks from web documents," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 588–593.

[12] D. Boswell, "Introduction to support vector machines," 2002.

[13] J. Kivinen, M. K. Warmuth, and P. Auer, "The perceptron algorithm versus winnow: linear versus logarithmic mistake bounds when few input variables are relevant," *Artificial Intelligence*, vol. 97, no. 1, pp. 325–343, 1997.

[14] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of naive Bayes text classifiers," in *ICML*, vol. 3. Washington DC), 2003, pp. 616–623.

[15] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive Bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

[16] G. Rios and H. Zha, "Exploring support vector machines and random forests for spam detection." in *CEAS*, 2004.

[17] L. Breiman, "Random forests," *Machine learning*, pp. 1–33, 2001.

[18] N. Bhan, "COMPARATIVE STUDY OF EM AND K-MEANS CLUSTERING TECHNIQUES IN WEKA INTER- Introduction to Weka Evolution of Weka," *International Journal of Advanced Technology & Engineering Research*, vol. 3, no. 4, pp. 40–44, 2013.

[19] Y. Zhao, G. Karypis, and U. Fayyad, "Hierarchical clustering algorithms for document datasets," *Data Mining and Knowledge Discovery*, vol. 10, no. 2, pp. 141–168, 2005.

[20] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, no. 1. Boston, 2000, pp. 525–526.

[21] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 2nd ed., 2006.

[22] J. Ross, Q. Morgan, and K. Publishers, "Book Review : C4 . 5 : Programs for Machine Learning," vol. 240, pp. 235–240, 1994.

[23] H. Liu and H. Motoda, "Computational methods of feature selection."

[24] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[25] C. van Rijsbergen, *Information Retrieval. 1979*. Butterworth, 1979.

[26] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 1.

[27] I. Dilrukshi and K. De Zoysa, "Twitter news classification: Theoretical and practical comparison of svm against naive Bayes algorithms," in *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*. IEEE, 2013, pp. 278–278.

Appendix A

Code of Python script for data extraction

Listing A.1: data.py

```
import twitter
api = twitter.Api(consumer key='uM3bxxxxxxxxxxxxxPrAQ',
consumer secret='fuTaxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxhnxU',
access token key='4738xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxagwy',acc
ess token secret='rwyXxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxQmV8')

statuses = api.GetUserTimeline(176337215,count=200)
f = open('/home/dell/Documents/data/adaderana.txt', 'a')
for s in statuses:
 f.write(str(s.id) + '\t' + s.created at encode('utf8') + '\t' +s.t
 ext.encode('utf8') + '\n\n')

f.close()

statuses = api.GetUserTimeline(336444791,count=200)
f = open('/home/dell/Documents/data/Ceylontoday.txt', 'a')
for s in statuses:
 f.write(str(s.id) + '\t' + s.created at encode('utf8') + '\t' +s.t
 ext.encode('utf8') + '\n\n')

f.close()

statuses = api.GetUserTimeline(190521777,count=200) f =
open('/home/dell/Documents/data/ITN.txt', 'a') for s in
statuses:
 f.write(str(s.id) + '\t' + s.created at encode('utf8') + '\t' +s.t
 ext.encode('utf8') + '\n\n')

f.close()

statuses = api.GetUserTimeline(87921110,count=200)
```

```
f = open('/home/dell/Documents/data/lankabreaking.txt', 'a')
for s in statuses:
 f.write(str(s.id) + '\t' + s.created at.encode('utf8') + '\t' +s.t
 ext.encode('utf8') + '\n\n')

f.close()

statuses = api.GetUserTimeline(339564751,count=200)
f = open('/home/dell/Documents/data/news1st.txt', 'a')
for s in statuses:
 f.write(str(s.id) + '\t' + s.created at.encode('utf8') + '\t' +s.t
 ext.encode('utf8') + '\n\n')

f.close()
```