A STUDY OF FEATURE EXRACTION TECHNIQUES FOR CLASSIFYING TOPICS AND SENTIMENTS FROM NEWS POSTS

Wafa Zubair Abdullah Al-Dyani

814383

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

UNIVERSITI UTARA MALAYSIA

2014

Permission to Use

I'm presenting this thesis in fulfilment of the requirement for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to Universiti Utara Malaysia and to me for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

ABSTRAK,

Banyak saluran berita mempunyai laman Facebook sendiri yang jawatan berita telah dikeluarkan di harian. Oleh yang demikian, posting berita ini mengandungi pendapat duniawi tentang peristiwa-peristiwa sosial yang mungkin berubah dari masa ke masa disebabkan oleh faktor-faktor luaran serta boleh menggunakan monitor untuk peristiwa-peristiwa penting berlaku seluruh dunia. Hasilnya, banyak teks perlombongan penyelidikan telah dijalankan dalam bidang analisis sentiment sebagaimana satu tugas yang paling mencabar adalah untuk mengesan dan mengeluarkan ciri-ciri utama dari siaran berita yang tiba secara berterusan lebih masa termuka dalam menghasilkan dataset tidak seimbang. Walau bagaimanapun, mengekstrak ciri-ciri ini adalah satu tugas yang mencabar kerana sifat-sifat yang kompleks di post, juga posting tentang topik tertentu mungkin berkembang atau hilang kerja lebih masa. Oleh itu, kajian ini telah membangunkan satu analisis perbandingan mengenai ciri-ciri kaedah pengekstrakan yang mempunyai pelbagai ciri-ciri pengekstrakan teknik (TF-IDF, TF, b, IG, chisquare) dengan tiga ciri n-gram berbeza (Unigram, Bigram, Trigram), dan menggunakan SVM sebagai Pengelas. Tujuan kajian ini adalah untuk mencari yang optimum ciri pengekstrakan teknik (FET) yang dapat mencapai hasil ketepatan optimum untuk topik dan sentimen klasifikasi. Sehubungan dengan itu, analisis ini adalah dijalankan ke atas tiga saluran berita datasets. Keputusan eksperimen bagi topik klasifikasi telah menunjukkan bahawa chisquare dengan unigram telah terbukti menjadi FET yang terbaik berbanding kaedah lain. Selain itu, untuk mengatasi masalah tidak seimbang data, kajian ini telah digabungkan FET ini dengan teknologi OverSampling. Keputusan penilaian telah menunjukkan peningkatan dalam prestasi di Pengelas dan telah mencapai ketepatan yang lebih tinggi pada 93.37%, 92.89% dan 91.92 BBC, Al-Arabiya dan Al-Jazeera, masing-masing, berbanding dengan apa yang telah diperolehi pada dataset asal. Begitu juga, gabungan yang sama telah digunakan untuk pengkelasan sentimen dan memperolehi ketepatan perakaman pada kadar 81.87%, 70.01%, 77.36%. Walau bagaimanapun, ujian yang diiktiraf optimum TFT jawatan dipilih secara rawak berita tersembunyi telah menunjukkan ketepatan perakaman yang agak rendah bagi kedua-dua topik dan sentimen klasifikasi akibat dari beberapa perubahan topik dan sentimen dari masa ke masa.

Kata kunci: Teks perlombongan, klasifikasi teks, analisis sentimen duniawi, teknik pengekstrakan ciri, saluran berita, acara sosial, data yang tidak seimbang.

ABSTRACT

Recently, many news channels have their own Facebook pages in which news posts have been released in a daily basis. Consequently, these news posts contain temporal opinions about social events that may change over time due to external factors as well as may use as a monitor to the significant events happened around the world. As a result, many text mining researches have been conducted in the area of Temporal Sentiment Analysis, which one of its most challenging tasks is to detect and extract the key features from news posts that arrive continuously overtime. However, extracting these features is a challenging task due to post's complex properties, also posts about a specific topic may grow or vanish overtime leading in producing imbalanced datasets. Thus, this study has developed a comparative analysis on feature extraction Techniques which has examined various feature extraction techniques (TF-IDF, TF, BTO, IG, Chi-square) with three different n-gram features (Unigram, Bigram, Trigram), and using SVM as a classifier. The aim of this study is to discover the optimal Feature Extraction Technique (FET) that could achieve optimum accuracy results for both topic and sentiment classification. Accordingly, this analysis is conducted on three news channels' datasets. The experimental results for topic classification have shown that Chi-square with unigram have proven to be the best FET compared to other techniques. Furthermore, to overcome the problem of imbalanced data, this study has combined the best FET with OverSampling technology. The evaluation results have shown an improvement in classifier's performance and has achieved a higher accuracy at 93.37%, 92.89%, and 91.92 for BBC, Al-Arabiya, and Al-Jazeera, respectively, compared to what have been obtained on original datasets. Similarly, same combination (Chi-square+Unigram) has been used for sentiment classification and obtained accuracies at rates of 81.87%, 70.01%, 77.36%. However, testing the recognized optimal FET on unseen randomly selected news posts has shown a relatively very low accuracies for both topic and sentiment classification due to the changes of topics and sentiments over time.

Keywords: Text mining, Text classification, Temporal Sentiment analysis, Feature extraction techniques, News channels, Social events, Imbalanced data.

Acknowledgement

All praise is to Allah, who by His grace and blessings I have completed my thesis. Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my advisor and supervisor, Dr. Farzana Kabir Ahmad. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. Dr. Farzana taught me how to question thoughts and express ideas. His patience and support helped me overcome many crisis situations and finish this dissertation. I hope that one day I would become as good an advisor to my students as she has been to me.

Many friends have helped me stay sane through these difficult years. Their support and care helped me overcome setbacks and stay focused on my graduate study. I greatly value their friendship and I deeply appreciate their belief in me.

Most importantly, none of this would have been possible without the love and patience of my family. My family to whom this dissertation is dedicated to, has been a constant source of love, concern, support and strength all these years. I would like to express my heart-felt gratitude to my family especially to my most important person in my life, my beloved mother who has been a mother and father to me throughout my life and whose without her prays I might not be able to gain what I have achieved until now, so all thanks and gratitude to my dear mother.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

A	Accuracy
API	Application Programming Interface
BOW	Bag Of Words
ENN	Edited Nearest Neighbour
ETD	Emerging Trend Detection
FE	Feature Extraction
GIBC	General Inquire Based Classifier
HMM	Hidden Markov Model
HTTP	Hyper Text Transfer Protocol
IG	Information Gain
IMDB	Internet Movie DataBase
KDD	Knowledge Data Discovery
KNN	K Nearest Neighbourhood
LDA	Latent Divichlet Allocation
ME	Maximum Entropy
MI	Mutual Information
ML	Machine Learning
MLT	Machine Learning Techniques
NB	Navie Bayes
NFS	New Feature Selection
NLP	Natural Language Processing
OM	Opinion Mining
PMI	Pointwise Mutual Information
POS	Part Of Speech
RBC	Rule Based Classifier
RMSE	Root Mean Square Error
ROS	Random Over Sampling
SA	Sentiment Analysis
SBC	Static Based Classifier
SC	Sentiment Classification
SMOTE	Synthetic Minority Oversampling Technique
SNs	Social Network sites
SVM	Support Vector Machine
SWN	SentiWordNet
TD	Trend Detection
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TH	ThresHold
TOM	Temporal Opinino Mining
UGC	User Generated Content
WMP	Weighted Mean Precision
WMR	Weighted Mean Recall

CHAPTER ONE

INTRODUCTION

1.1 Overview of Study

Recent years have gained a great attention in the text mining and temporal sentiment analysis research field due to the large amount of opinion data generated in Social Networks sites (SNs) such as Facebook and Twitter. Facebook is the most famous and common SNs among Internet users for expressing their feelings, opinions, emotions and thoughts. Furthermore, Facebook has shown a tremendous increase in usage as it offers a valuable source for real time news and act as an opinions platform [1,2]. Hence, large number of news channels committee have created their own pages on Facebook, to allow news reader to post their opinion and thought on daily news items. The key idea at this point is to gain deep insight about what news readers think and feel towards various events.

Generally, news posts can be used as a monitor mechanism to detect the significant events which have been happening around the world. Furthermore, some events may grow up or vanish over time due to external factors such as change of time, evolution of recent events, or emergence of new events. As a result, such events may affect the overall opinions and consequently change correlated sentiments. Hence, in order to analyze these changes, a new field of sentiment analysis has been emerged in this area which is called Temporal Opinion Mining (TOM). TOM is defined as "a process of detecting and monitoring possible changes to particular opinions and their correlated sentiments over a given period of time and can be seen as a continuation of opinion mining" [3]. The main idea of TOM is to find the opinions average on a specific topic at different times. This analysis leads to

a complete timeline, which can be represented as a graph based on sentiment's values [4], or opinions average. However, one of the most challenging tasks in TOM is to detect and extract meaningful data or key features from documents that arrive continuously overtime.

Key features are a set of major words in a post that provide a high-level description of its contents to reader. Furthermore, these features are very important to be recognized as it will contribute to the correct detection of post's topic and sentiment category. Although many studies have been conducted in this area, detecting relevant key features on a specific event from news text is very difficult task [5]. This is due to the fact that large number of news posts have been released continuously over a period of time [6] , and consequently leads to a major problem which is the high dimensionality of the feature space as well as the complex properties of news post have further complicated the analysis process.

Besides that, news post usually includes a large subject domain. In other words, it may contain different features that belong to various categories, also it may contain complex event descriptions, various group of people. Thus, many studies [5,15,16] have addressed this problem by using news documents corpus as input [9], then topic key features extraction process is executed. Mainly, feature extraction technique aim to select suitable set of keywords based on the whole corpus of news documents, and assign weights to those key words. Various feature extraction techniques usually treat the document as a group or Bag of Words (BOW). In addition, various feature extraction techniques like Chi-squared [10], Pointwise Mutual Information (PMI) [11], Information Gain (IG) [12], Term Frequency and Term Frequency-Inverse Document Frequency (TF-IDF) [27, 36–38]have been used

to extract the main key features in studies such as temporal sentiment analysis, information retrieval, text classification, text categorization, document summarization and topic detection.

As a result, different computational works have been conducted, which mainly used text mining as an approach to analyze the unstructured data. Such studies have used Twitter as it attracts the attention of scholars and professionals [7–9]. In contrast, relatively a small number of studies have been noticed using Facebook [10–12], to either identify trends, topic classification, or detect posts' sentiments [1, 5,22–24]. This may happen due to Facebook privacy policies, which restrict and confine the data from being collected.

Furthermore, many studies have been carried out on Topic Detection and Tracking (TDT) over social media, but mainly focus on analyzing long text discussions from blog posts [25], and Google trends [26]. Majority of work in TDT field has followed the content based approach, in which content features of a document were selected using unigram or n-grams, then the document will represented as a vector of features in a text stream. Consequently, the new arrival document will be classified based on the similarity with the existing pool of events. If the similarity exceeds a specified threshold, the document will classified into the nearest events and update the new document. Otherwise, a new event will be formed [27].

In addition, news posts usually contain indirect sentiment words such as "kill", "hit" rather than direct sentiment words like "angry", "happy". Thus, such scenario makes the analysis of sentiment classification task become more difficult. Besides that, different news websites may have similar or opposite sentiment orientation

toward a contentious news topics such as (terrorism, disease, and conflict) based on their own stand of points or interests. This orientation is not restricted to positive-negative sentiment only, but may include other types of sentiments such as anger, joy, fear and so on. Studies on news text sentiment analysis primarily concentrate on the writer's perspective, which try to identify writer's sentiment orientation toward a specific target. In contrary, in many cases the reader's sentiments triggered by the news document do not usually agree with that of author's. Previous studies [21–23] have focused generally on the single label classification techniques which assign only one sentiment to the document from a set of multiple sentiment categories, while other research focus on multi label classification considering that a reader may have combination of more than one base sentiment toward a particular entity [31].

Anyway, for all previous studies the labelling process have been done using either supervised or unsupervised approach. Supervised approach such as machine learning technique in which a classifier will be trained on a pre-defined corpus of documents, so it would later contribute in labelling the other documents [22,25]. While unsupervised approaches such as clustering [33] and symbolic techniques. Where, the main idea of symbolic technique is to assign each term a sentiment score manually [13,22], or through utilizing one of the free online lexica resources [34]. However, due to the low coverage issue of the symbolic way some studies have gone to create their own emotion lexicons through utilizing annotated documents with emotions [8,35] .

Moreover, news posts about a specific topic or its associated sentiments may grow or fade in intensity for some periods of time [6] and consequently lead to the emergence of imbalanced dataset problem for text classification, is the imbalanced

data issue in which imbalances data some classes may have higher number of samples compared to other classes and consequently affects the classifier performance. Additionally, due to the problem of imbalanced data which usually occurs in text classification and temporal analysis, some researchers have gone with applying resampling technologies at either algorithmic [36] or data levels [10,37,38].

As a result, this study aims to identify the appropriate feature extraction technique that could lead to the determination of significant key features from large number of news posts on Facebook and consequently improve the topic and sentiment classification performance. Hence, this research has proposed a topic and sentiment classification models for news posts on Facebook on three different news channels datasets such as al-Arabiya, Aljazeera, and BBC. In this study, various number of feature extraction techniques namely Term Frequency-Inverse Document Frequency (TF-IDF), Term Frequency (TF), Binary Term Occurrence (BTO), Chi-squared, and Information Gain (IG) have examined on different n-gram features (unigram, bigram, and trigram) in order to identify which combination of feature extraction technique and n-gram that could offer better classification accuracy. Moreover, this study has applied a resampling technology called Random Over Sampling (ROS) in order to overcome the imbalanced problem and consequently improve classification performance. Furthermore, this analysis is performed on prominent classifier named Support Vector Machine (SVM).

The rest of this chapter is organized as following. Section 1.2 introduces the definition of the problem statement related to this study's title. In Section 1.3 research questions are presented followed by research objectives in Section 1.4.

Sections 1.5, 1.6, 1.7 explain the scope, significant and the organization of the report, respectively. Finally, Section 1.8 introduces the chapter summary.

1.2 Problem Statement

Recent years have witnessed a great attention to the text mining researchers due to the availability of large amount of opinion data that have been generated in Social Networks sites (SNs) such as Facebook and Twitter. Facebook is the most famous and common SNs among internet users with more than 800 million active users. Most of the times, Facebook users used this medium to express their feelings, opinions, emotions and thoughts. Consequently, many news channels have taken this opportunity to create their own Facebook pages in which news posts have been releasing in a daily basis.

These news posts on SNs reflect author's opinions, emotions towards different events, and which usually written with the intention to "provoke" emotions and attract the reader's attentions. Typically, such news posts can be used as a medium to detect significant events which have happened all around the world such as airplane crash, terrorism, conflict as well as follow up the changes of opinion's polarity associated with each event over a period of time and cause the behavioural convergence of the expressions of shared sentiments [27].

In addition, news posts contain temporal opinions about social events that may have variation over time or be outdated after sometime due to external factors like change in time, evolution or emergence of recently events. Such events may affect the overall opinions at any given time, resulting in dynamically changing sentiments. This kind of opinions is known as Temporal Opinion Analysis (TOM). However, one

of the most challenging tasks in TOM is to detect and extract meaningful data or key features from news posts that arrive continuously overtime [39]. Subsequently, extracting significant features for an event is a difficult process and extremely time consuming, especially when it is done manually [5]. As a result, various computational works have been done in this area, which mainly have used text mining as an approach to analyze the unstructured data.

Thus, the main problem for this study is how to detect and extract the features for a specific event from news posts on Facebook, which have released continuously over a period of time [3,5], and consequently leads to a major problem which is the high dimensionality of the feature space. Subsequently, this study also aims to investigate the opinion's polarity changes as events have emerged. The extracted features are important to be recognized as they would contribute in detecting the correct post's topic category as well as to classify the opinion's polarity. However, detecting changes within opinions and detecting relevant key features for a specific event from news posts on Facebook overtime seems to be very difficult task mainly due to the complex properties of new posts. News posts may include a large subject domain which means contain different features belonging to various categories as well as include complex event descriptions or various group of people [40].

Besides that, news posts usually consists of sentences with mixed views or sentiments and journalists at most of the time attempt to exclude themselves from expressing directly positive or negative opinions especially for sensitive issues. Therefore, they may express their opinions using various ways such as using a complex discourse or argument way, or quoting others sentences in order to express their own thoughts [40]. Furthermore, that news posts often contain indirect

sentiment words or informative words to describe the event [40] such as “injure”, “explosion” rather than direct sentiment words like “sad”, “glad”. Hence, sentiment classification for news post depend highly on the words with strong class information rather than sentiment words and that would make the task of sentiment classification for news posts become more difficult [35].

Moreover, news posts with various opinions about a specific topic may grow or vanish in intensity for some periods of time [3]. This leads to difficulty in extracting features because there are large number of features that required to be extracted and analysed as there is an increase in the news posts. In contrast, the decreasing number of news posts would result in a lack of features to be extracted and therefore these features may not be able to represent the new upcoming posts of the same event. Consequently, this leads to the emergence of another problem related to text classification which is imbalanced data issue [10] that usually appear in text classification tasks due to the overwhelmed of majority classes which affects the classification performance and consequently decrease the classification performance accuracy of the minority classes [37].

Finally, data collected from social media usually contain noisy words such as (stop words, URLs). These noisy words may affect the performance of the classifier in the classification phase. Hence, a proper pre-processing techniques has applied in order to remove all noisy words such as tokens filter and English stopwords removal techniques.

As a result of all previous problems, feature extraction technique has become more important than classification techniques, especially for highly imbalanced datasets [35,42] as the situation in this study. Therefore, this study has implemented

various features extraction techniques such as Chi-squared [10], PMI [11], IG [12], TF-IDF [27,40–42] for extracting features.

Hence, this study aims to identify the appropriate feature extraction technique that could lead to the determination of significant key features from large number of news posts and consequently improve the topic categorization and sentiment classification performance. Furthermore, this research has proposed a topic and sentiment classification models for news posts on Facebook for three different news channels datasets such as al-Arabiya, Aljazeera, and BBC.

Additionally, the various feature extraction techniques have been inspected on different n-gram features (unigram, bigram, and trigram) in order to identify which combination of feature extraction technique and n-gram that could offer better classification accuracy. Subsequently, this analysis has been tested on most commonly used classifier SVM. Besides that, this study has implemented ROS resampling technique in order to solve the imbalanced data problem and has shown its effectiveness on classification performance. In the following sections research questions and objectives are presented.

1.3 Research Questions

The following questions are formed to address research problem:

1) What kind of feature extraction technique could lead to the determination of temporal robust features from large number of news posts?
2) How to classify news post into its corresponding topic and sentiment categories for the social events?
3) How to evaluate the performance of the proposed feature extraction technique?

1.4 Research Objectives

The main objective of this study is to develop a comparative analysis of feature extraction techniques in order to recognize topics and sentiment classes for social event detection. Additionally, to obtain the main research objective the following sub-objectives need to be addressed:

1) To analyse feature extraction technique that could lead to the determination of temporal robust features from large number of news posts.
2) To classify the news post into its corresponding topic and sentiment categories for the social events.
3) To evaluate the performance of the proposed feature extraction technique.

1.5 Scope of Study

This research focused on developing a comparative analysis on feature extraction techniques in order to identify the feature extraction technique that could lead to the determination of temporal robust features from large number of news posts on Facebook. These features are essential to be recognized as it will contribute in identification of the correct topic and sentiment categories for the news post.

Thus, the study has implemented various feature extraction technique, namely TF-IDF, TF, BTO, IG, and Chi-square which have been examined on different n-gram features such as unigram, bigram, and trigram in order to determine which combination of feature extraction technique and n-gram that could lead to a higher classification accuracy for both topic and sentiment classification of the news posts.

In addition, these feature extraction techniques have been evaluated using the same validation technique for either original datasets, resampled datasets, or unseen news posts for topic categorization, while for sentiment classification it has been evaluated on already resampled datasets and unseen news posts. The utilized validation technique is an automatic 5-fold cross validation for the SVM classifier. Each evaluating process for each news channel has been done using various performance metrics such as Accuracy (A), Weighted Mean Recall (WMR), Weighted Mean Precision (WMP), Root Mean Squared Error (RMSE), and Confidence score. Additionally, this research has applied random over sampling technique called OverSampling Bootstrapping in order to improve the classification performance of the minority classes for both topic and sentiment classification.

Furthermore, this study has focused on English news posts which have been released in the year 2014 by three popular news channels on Facebook, namely Aljazeera, BBC, and Al-Arabiya. These three channels are selected as they have recognized as the popular and most viewed news channels with reasonable credibility measures, where BBC has ranked as the best second news channel on Facebook [43]. Also, it has gained the name of the best international news channel at the association for international broadcasting award in 2006 [44,45]. Meanwhile, Al-Arabiya has achieved the second most frequently viewed channel after Al-Jazeera in Middle East especially among Arab world according to survey published by Northwestern University in Qatar [46]. Moreover, a study done by Thomas [47] has compared the degree of accuracy, fairness, believability, trustworthiness, and expertise for four Arab news channels including Al-Jazeera and Al-Arabiya, where the results of this study has shown, that a higher degree measurement for the five credibility have been obtained for Al-Jazeera followed by Al-Arabiya then the other two news channels.

Additionally, this study focused on the following five categories, namely conflict, terrorism, airplane crash, disease, natural disaster as these events are among the most frequent happened events during the last year 2014 around the world [48–50]. These events were the most notified talked-about global topics on Facebook [51,52] and Twitter [53,54]. Moreover, this study aims to classify opinion's polarity about the five former events to eight sentiments (sad, afraid, amused, inspired, happy, don't care, annoyed, or angry) in order to figure out the sentiment orientation of those posts toward the five topics. These specific sentiment categories have been used as they have provided by the DepecheMood, the tool which has been used by this study to label each news post into its suitable emotion category from the eight available emotions. DepecheMood is created by Jacopo and Marco in [35], where their study's experiment results have shown high precision and high coverage of DepecheMood compared to other sentiment lexicons like (ANEW, Warriner, SWN-prior) as well as other emotion lexicons such as WNAffect, also indicate significant improvements over unsupervised methods

1.6 Significance of Study

This study has identified some areas of significance. Study's results can be beneficial for anyone interested in a particular category/topic of news channels posts as well as, investigating the changes of associated opinions. The beneficiaries of this study may include news channels owner, journalists, government, stakeholder or analyst, and for research area.

i. News channel owner and journalists

News channel owner and journalists can benefit to do some analysis about how many posts have been published on a particular topic category and what

sentiments associated with these topics which subsequently may help journalists to obtain significant insights and inspire ideas from those posts in order to write the new news posts about the same event in the future.

ii. Stakeholder and analysts

This study is beneficial for the stakeholder and analysts who are concerned about future economic or analysing certain event trends. They could extract useful information from the graphical representation of the news posts to obtain deeper understanding before taken any action. For example, investigate the possible effects of news about social events on various domains such as in stock market. In general, bad news means bad news for stocks and currency prices in financial sector, thus affecting the economic.

iii. Government

Government may benefit from the knowledge of current events and try to take an initial steps to solve the existing problems or prevent the occurrence of such events in the future as well helps in making the right decisions.

iv. Research area

Although, various feature extraction techniques have been implemented and examined with different n-grams models, but there is still needs to discover which combination of feature extraction technique and n-gram that would give better performance results. Hence, this present study proposes a temporal feature extraction technique which could lead to the determination of temporal robust features from large number of news posts whereby these features would contribute in classifying the news posts on Facebook into their appropriate topic and sentiment categories. In

addition, and based on our literature review there is no evidence shows that any research has done a comparative analysis study of feature extraction techniques for topic categorization and sentiment analysis of news posts on Facebook. Thus, this study is consider to be the first one.

1.7 Report Organization

This report is organized into six chapters. The description of each chapter is given as follows:

In Chapter One, a brief overview of study is introduced followed by thorough discussions on the main problem encountered in this research field. Research question and objectives were also presented along with the scope, and significant.

In Chapter Two, a previous works on temporal sentiment analysis and other related domain area are meticulously reviewed. The key idea is to discover problems have encountered in related works, and identify methods and techniques that have been used in order to overcome the obstacles. Furthermore, the main concepts of this research field are also given in this chapter to assist reader to obtain better understanding of the concept used throughout the thesis.

In Chapter Three, the methodology for this research is proposed based on the problems mentioned in Chapter 1 in order to address and achieve the research objectives.

In Chapter Four, the results and discussion of the topic categorization experiment are presented.

In Chapter Five, the results and discussion of the sentiment classification experiment are described.

In Chapter Six, the conclusions of the both topic categorization and sentiment classification experiments are presented followed by the contribution and limitation of study and finally future works for this research are described.

1.8 Chapter Summary

This chapter presents a brief overview of this study followed by the meticulous discussion on the problem statement. Consequently, research questions and objectives are laid down in order to state problems. Scope and significance of this study are also introduced in this chapter.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The emergence of the Web 2.0 provides a huge virtual area to share information and express opinions. This changed the way of generating the web's contents, as internet users become active participants rather than being passive ones. Billions of internet users today utilizing Social Network sites (SNs) for various purposes such as keep in touch with their friends or families, share information and their view of points on different topics such as political issues, movies, technology, products, services, religious issues as well as different social events happening all over the world [16]. Various manners of sharing have been using by SNs users like status, comments, videos, pictures or links to other pages.

Consequently, this large data have offered an opportunity as, 30% of internet users prefer to evaluate services and products online rather than depending on company marketing via looking at the reviews wrote by other consumers about prior experience of products and services [6]. Hence, SNs affect consumers' opinions by shaping their behaviours and attitudes. In addition, analyzing and tracking SNs activities may provide a deeper insight about sentiments towards brands or products, measure consumer's loyalty, as well as predict the fail or success of a candidate or election campaign and follow-up the changes of social events in the world over periods of time. Nowadays SNs has been used by people as a communication tool. This indicates that they already moved from traditional medium like email and blog to micro-blogging. SNs have been produced a huge volume of the dataset, which can

be useful for any specific entity or organization in order to do marketing and social studies [12,38]. Thus, this mixed mood approach eliminates the survey, as no need to conduct survey person to person or home, where opinions can be collected and analysed from these SNs and conclude what people like/dislike by analyzing them.

Recently, many SNs like Facebook, Twitter, Google, LinkedIn, and Myspace have shown a sharp increase in usage by internet users [22]. Currently, Facebook is the largest SNs with more than 800 million active users [43,44], followed by Twitter users with 280 million users [45,46]. A statistic study done by [25] showed that Facebook user get over 1000 posts every week by nearly 130 friends, and internet's user spend about 55 minutes every day on the site as well as Facebook become an effective factor on people's lives and activities as it uses as communication tool to share feelings, thoughts, as well as inform others about conditions in their countries and cities through sharing short posts, pictures and videos in order to push them to interact and care about what they think [1].

Generally Facebook provide five disparate types of posts like status, links, videos, music, and photos. Facebook's posts and comments consider to be very useful since they can be used to extract their authors' opinions towards various entities or events. Hence, analysing and detecting User Generated Contents (UGC) on SNs like Facebook can be used as indicator to obtain deeper insight into the significant topics which have attracted the interest of large number of internet users such as sociologists [2], news reporters mainly those who are monitoring the fast evolving of news stories [59], scholars that are interested in elaborating a phenomena in SNs, and marketing professionals [60].

Trend of monitoring users' opinions has been studied lately in Twitter as a subject attracts the attention of scholars and professionals. In contrast, Facebook has relatively a small number of studies [10–12], and there has been little commercial or research efforts done on either trends detection, classification of posts or detection of the posts' sentiments [12–14,38]. This situation may occur due to limitation in volume and type of available data to analysis due to Facebook privacy policies, which restrict the data to be collected. Although, the limitation in collecting Facebook data, some papers have chosen Facebook as a source of data because of its main features such as: 1) Facebook is dominating SNs by its large number of internet users. 2) Facebook does not belong to any particular party, its contents is general and not bias for any survey. 3) Facebook has various users from diverse countries, and categories [61].

As a result, the large amount of UGC on SNs has become a valuable sources of real time news and opinions [24,25]. Hence, many news channels currently use SNs to improve the interactivity with newsreaders by enabling them to comment on news articles [64]. A study by Kilner and Hoadley [65], has shown that users usually comments on news in order to add information, ask and answer, share their experience, support or criticize the news and express their sentiments. Recently, many online news channels have their own pages on Facebook such as BBC, Aljazeera and Al-Arabiya, etc, on which daily news posts are released. These posts normally contain their authors' opinions, thoughts, and sentiments, which could be used as an indicator to analyze various social events' changes over a period of time. Beside that news posts are usually written by creative authors with the intention to "provoke" emotions as well attract the reader's attention. In addition, to understand news' trends, detecting the topic is an important key as many articles in a single

newspaper or a set of pages as well a set of Facebook posts may describe the same topic [66].

Therefore, an analysis process is required for those posts to keep track on the social events happening all over the world. This can be done using state of art text mining and sentiment analysis. Which has been used in this kind of researches to detect topics and monitor possible changes of opinions towards an event occurring over a period of time. In the next section, text mining and the role of sentiment analysis is described along with its common used approaches and significant challenges within it.

2.2 Text Mining

Text mining has become an important and popular field in the research of data mining due to the rapid increase of electronic documents in the web. Text mining is different from data mining. On one hand, text mining is defined as the process of extracting useful information or knowledge from unstructured text documents like the ones on SNs [24]. On the other hand, data mining focus on structured textual databases [67]. Text mining includes various tasks such as sentiment analysis, opinion mining, document summarization, text clustering, information extraction, topic tracking, information visualization and text categorization and so on.

Text mining utilizes techniques of information extraction, information retrieval and Natural Language Processing (NLP) and connect them with methods and algorithms of Knowledge Data Discovery (KDD), data mining, machine learning and statistics. In addition, textual information is divided into two types: facts which

are defined as “objective expressions about events, entities and their features”. The other type is opinions which are defined as “subjective expressions that describe people’s sentiments towards entities, events and their features” [68]. Consequently, text mining has been used to perform sentiment analysis on public blogs, forums and SNs such as Facebook and Twitter.

2.3 Sentiment Analysis

In the past, the process of obtaining data sources was a big issue, however nowadays this issue is no longer existed due to the availability of large volumes of web data that are produced on a daily basis. SNs like Facebook and Twitter are one of these available sources that provide User Generated Content (UGC). UGC consists of real time opinions that may bring marketing up and down. Thus, companies and individuals are heading towards using mechanisms to search and collect these data in order to mine and convert them into useful information which could help them in making an appropriate decision [28, 29].

Recently, this area of study has been attracting the attention of many researchers in order to extract, analyze and summarize the information that is merged within unstructured data on SNs such as text and speech. Information can be in many forms, for instance opinions, feelings, sentiments, emotions, or thoughts which are embedded in various reviews like product reviews, blogs, restaurant and movie reviews as well as news articles and posts [29,30]. This large amount of data has led to the emergence of a new field in Natural language Processing (NLP) and Computational linguistics which is called Opinion Mining (OM), Sentiment Analysis (SA), sentiment detection, review mining, or polarity classification.

SA and OM share the same meaning of studying people's sentiments, opinions, attitudes, and emotions towards various entities such as events, topics or individuals. Furthermore, SA and OM both attempt to identify subjectivity of text from objectivity ones, then classify subjective text into its polarity. On the other hand, objective texts contain factual information about a particular entity [72]. However, OM and SA have a slight difference, where SA searches opinions, detects sentiments about opinions, and then classifies them into either (positive, negative, or neutral). Meanwhile, OM focuses on extracting and analyzing people's opinions about any entity [73].

SA also can be defined as classification process where the majority of studies has classified the sentiments within reviews into different classes. There are various kinds of classification which include good versus bad [74], like versus dislike [75], support versus opposition [76], happy, sad, or neutral [22], or a set of various emotions like anxiety, anger, grief, shock, complaint, fatigue, and suffering [4], and so on. Sentiment analysis can be utilized in different ways for instant, it can be used in identifying which service or product gain more popularity, helping to determine the success or failure of an ad campaign or launch of new products as well as specifying which particular features have been liked or disliked by demographics.

Recently, there are three main predominate fields in SA, namely sentiment classification, feature based sentiment classification and opinion summarization. Sentiment classification classifies the whole document depending on the opinions it has towards specific entity, while feature based sentiment classification find and classify the opinions about entity's features. Opinion summarization on other hand summarizes opinions expressed by customers on entity's features [77].

In general, there are three levels of classification in SA, namely document, sentence, and attribute or aspect classification levels. Document level considers the whole document as a single unit describing about one topic and having opinions on one target [78]. However, this assumption is true in some cases such as product, restaurant, movie reviews [71], and not applicable for other reviews such as news articles and news posts. This is mainly due to general targets in news post or articles, and it may consist a group of people, organizations, or complex event descriptions.

In addition, journalists may exclude themselves from expressing direct opinions, especially when describing sensitive issues. For example, they may quote others phrases in order to express their own opinions or express feelings using complex arguments [79]. Hence, the challenging task for SA in document level is to extract the features on which opinions have been expressed. Another challenging task is to determine the polarity orientation of a document, since the document may contain several opinions about one topic, for instance positive and negative opinions are mixed in one review. Due to these challenges, news articles have attracted much attention compared to other types of reviews [40].

Another classification level called sentence level, which has two main steps. First, subjectivity classification to check whether the sentence is subjective or objective. Second, sentiment classification where this phase use to classify the sentence to either negative, positive, or neutral. The challenging task for this level is to determine which feature or target has been mentioned in the sentence which will be later used to determine the polarity [71].

Nevertheless, there is no big difference between document and sentence level classification since sentence level classification considers to be as a short document

[69]. Additionally, both of them do not require to go down into aspects level of an entity which is necessary in some kind of reviews such as product reviews where summary that contains evaluations for product's features believe to be more useful than an overall opinions summary. The challenging task for aspect level is to extract the aspect like (target, feature, or topic). However, an automatic extraction of features has been done by using NLP techniques [71].

In general SA has various approaches for features selection and sentiment classification: supervised or unsupervised approaches. Each approach has various methods in order to detect and classify sentiments from texts such as combination of NLP techniques and linguistic rules, Lexicon based method, or Machine Learning techniques. In addition to that, some studies attempt to use hybrid approaches to obtain good results [80]. For lexicon based approach, a list of pre-populated seed words (positive and negative) are expanded by finding their synonyms and antonyms using a well-known corpora like WordNet [81], or thesaurus [82], then use them to detect and predict the sentiments.

On the other hand, NLP and linguistic rules are used to identify opinions and predict their sentiments. It usually finds adjective noun sequence through applying POS tagging and determines its polarity by examining context rules. In Machine Learning based approach the dataset will be usually divided into training and testing sets. A training set will usually classified manually by polarity values, then trains a selected classifier the various characteristics of documents, meanwhile for testing set, it will be used to evaluate the performance of the trained classifier.

A number of machine learning techniques such as Support Vector Machine (SVM) or Naïve Bayes (NB) [71], Maximum Entropy (ME) has been used to classify

reviews and have achieved major success in text categorization [77]. Other examples of well-known machine learning techniques in NLP are C5, ID3, K-Nearest Neighbourhood, window classifier, centroid classifier, and N-gram model [77]. Most machine learning techniques has been used for classification tasks, therefore called "supervised learning" since the target is pre-defined. Additionally, the biggest issue within supervised learning is that, it needs a large volume of dataset with a good quality for training because learning process may fail if data is biased or insufficient [71]. Anyway, there are also several studies that have used unsupervised approaches in analyzing sentiments which uses sentiment's words and phrases as the main indicators for sentiment detection and classification [71].

Feature extraction techniques deal with the documents either as a single string that maintains the sequence of words or as Bag Of Words (BOW). The later one has been used more frequently because of its simplicity for classification phase. Moreover, the most common and easiest technique used by many researchers is called the Term Frequency – Inverse Document Frequency (TF-IDF) for detecting and extracting the robust features [1,27,36,38]. Other methods also have been widely used such as lexicon based and statistical methods for instance Point-wise Mutual Information (PMI), Latent Dirichlet Allocation (LDA), Information Gain (IG) and Chi-square.

Furthermore, different types of features have been used in sentiment classification most common ones are: 1) N-grams features with their frequency value or state of their presence or absence in the document [71]. 2) Part Of Speech (POS): where each word presents by its position in the sentence (noun, adjective, or adverb[83]). 3) Negation words: words that change polarity of the sentence. Various

techniques for feature selections and classification have been used by many studies depending on their objectives. Some of these studies will be introduced and discussed in sections (2.5.4 and 2.5.5).

Regardless, the popularity of sentiment mining research in recent years, analyzing sentiments of SNs data is a challenging topic, wh,ere extracting several features have denoted to be the challenging task in analyzing large SNs datasets which are listed in [3,72] as following: public's opinions can represent in complicated ways where a review may include mixed opinions about certain topics and these opinions may change over time or be outdated after some time. Besides that, a document may contain opinions or sentiments about multiple topics.

Sentiments can be expressed in different ways like using indirect expressions or complex structure which needs a common sense to be recognized. In addition, misleading opinions because of using sarcasm, irony, negation words, or having different styles in writing. Moreover, document level classification failed in detecting sentiments about topic's individual aspects. As well as tracking sentiments toward a specific topic found to be a difficult task, for this reason most opinions extraction algorithms perform badly in this area. Thus, the documents usually labelled manually or the opinion would assign to a topic term co-existing in the same context. In the next section, an introduction of a new recently emerged field of SA called Temporal Sentiment Analysis and its applications will be presented.

2.4 Temporal Sentiment Analysis

News posts on SNs reflect the opinions, thoughts, and emotions of their authors towards various social events and which are written in order to provoke the

emotions of their readers. Hence, these posts can be used as a mechanism to monitor the significant events happening around the world and discover the reasons behind the changes in events' sentiments over time. Some events may grow up or vanish over time due to external factors such as change of time, evolution of recent events, or emergence of new events. As a result, such events may affect the overall opinions and consequently change correlated sentiments. Hence, in order to analyze these changes, recently a new field of sentiment analysis has been emerged in this area which is called Temporal Opinion Mining (TOM). TOM is defined as "a process of detecting and monitoring possible changes to particular opinions and their correlated sentiments over a given period of time and can be seen as a continuation of opinion mining" [3].

The main idea of TOM is to find the opinions average on a specific topic at different times. This analysis leads to a complete timeline, which can be represented as a graph based on sentiment's values [4], or opinions average. Therefore, the most important advantages of TOM over OM are that, changes in opinions can be identified, and the causes of these changes can be detected [85].

However, identifying changes within opinions for a specific topic\event overtime is found to be very hard due to the different styles in writing the documents as well as relevant features are unknown and required a huge amount of data. In contrary, detecting known factors or features were much easier. Hence, one of the most challenging tasks in TOM is to detect and extract meaningful data or features from documents that arrive continuously overtime.

Furthermore, documents with various opinions about certain topic\event may fade or grow in intensity for a specific period of time [3]. For example, as time

passes by, consumer's concerns about certain product may dramatically diminish, or when the new advanced product is released by a competitor, then consumer opinions towards old one might plummet. In addition, in order to investigate and analyze why such changes have occurred in opinions, SA techniques are used, which consist of the phases such as feature extraction, burst detection, and visualization techniques. Feature extraction is a technique to identify and extract entity's features, burst detection is used to detect and analysis abnormal changes, and visualization means to display information graphically.

2.4.1 Related works of Temporal Sentiment Analysis in Various Domain Areas

Related studies in the area of TOM is the study [4], in which the authors work on news and blog articles to produce two types of graphs. First one is a topic graph which draws out all topics associated with a certain sentiment. This helps to figure out which event had the most impact on the given sentiment for a specific time. The second graph was emotion graph, which showed a set of sentiments overtime regarding to certain topics. This approach considers to be the most used ways in elaborating temporal opinion results, although it needs both sentiment and topic to be known in advance to analyse. Hence, it does not try to find or detect recently unknown events which may impact on sentiments. Similar approach will be implemented in our study. By having specific event and show how sentiments towards this event may change and fluctuate over time.

In the same track, another study done by Dipankar Das et al [86], who created a prototype system called TempEval-2007 along with the time ML framework . This system produces a graph to visualize opinions overtime and track changes, but concentrate on temporal relations between events associated with emotions.

Therefore, it differs from [4] and this study since they do not try to identify unknown events based on changes in emotions. A logically similar study to TempEval-2007 done by Mishne et al [87], who designed a system called MoodViews3. The system downloads updates repeatedly from a corpus of blogs in order to follow up, analyses moods and predicts what moods might turn into in an attempt to know why such changes in mood have occurred. The mood predictor uses both NLP and ML techniques to predict moods based on text contents. As well as identifying which terms have occurred more or less during specific peaks in mood using language statistics.

Quite different from those previous studies, the researchers in [3], apply OM techniques to analyse changes in hotel reviews extracted from travel sites. The authors showed how the results of the temporal sentiment analysis can be visualized graphically using Google Maps. Moreover, they provided a robust temporal feature extraction technique to explain the reasons behind changes in opinions as well as display "good", and "bad", geographical areas based on hotel reviews. On the other hand, Preethi el at.[39] present a prediction model based on temporal sentiment analysis of tweet to determine the causal relation between the events which consequently can be used to predict the event sentiment and identify the possible time period between predicted events.

2.4.2 News Channels on Facebook

Recently years have witnessed a great rise in news channels pages on SNs like Facebook and Twitter, where many online news channels nowadays have been heading towards raising the interactivity of their readerships via various methods. For instance, by allowing the readers to use any of their SNs accounts to comment on

news items located on either the news channel's main website or on the channel's Facebook or Twitter pages.

International news channels usually offer latest news on worldwide events through cable, satellite, or internet. Some examples of these international news channel are Al-Jazeera, Al-Arabiya, and BBC. BBC has launched in 1991 and has nearly an estimated of 76 million viewers weekly in 2014. In addition, it has almost 12 million followers on twitter and 26 million fans on Facebook as well as has ranked as the best second news channel on Facebook [43]. Additionally, BBC has been nominated as the best international news channel at The Association for International Broadcasting award in 2006 [44,45]. Due to its credibility, data from BBC has been widely used in many previous studies [88,89].

Al-Jazeera (English version) on the other hand, has been established in 2006 as the world's first English language news channel located at the Middle East. This channel has provided both regional and global news and its broadcast in 2009 has reached and viewed by nearly 130 million homes over 100 countries via cable and satellite, and still extending its broadcast to reach North America. In addition, Al-Jazeera has been ranked as the fourth most popular Middle East news channel across the world [90]. Furthermore, Al-Jazeera English news channel has obtained 45 prizes of medals and awards, and consequently has become a good competitor to other international news channels such as BBC,CNN and Al-Arabiya [91].

Al-Arabiya English news channel has launched in 2007 from United Arab Emirates to be a directed competitor of the Qatar based Al-Jazeera English news channel [92]. In addition, Al-Arabiya has ranked as the eighth place in the most viewed top ten Middle East You Tube channel list [90]. Additionally, Al-Arabiya has

achieved the second most frequently watched channel after Al-Jazeera in Middle East especially among Arab world according to survey published by Northwestern University in Qatar [46]. This survey has included more than 10,000 people from Egypt, Lebanon, Tunisia, Qatar, Bahrain, Saudi Arabia, Jordan and the UAE. The results of this survey has recognized Al-Jazeera and Al-Arabiya as two news sources that wide-ranging viewed in these countries. Furthermore, Al-Jazeera was recognized as the top news source by 26% of respondents throughout the region, while Al-Arabiya was preferred by 16% [93].

Another study has done by Thomas [47] which examines how Arab viewers judge the credibility of four news channels such as Al-Jazeera, Al-Arabiya, Al-Hurra, and Local Arab Station. This study has compared the degree of accuracy, fairness, believability, trustworthiness, and expertise for the four former news channels. The results have shown a higher degree of credibility measurements for Al-Jazeera followed by Al-Arabiya then the other news channels. As a result of such reasons, this study has used the news posts released on Facebook by the three news channels (Al-Jazeera, Al-Arabiya, and BBC) as the datasets for this research since these news channels have recognized as popular and most viewed news channels with reasonable credibility measures.

In addition, commentary on news items gives a deep knowledge about reader's feelings, opinions, and thoughts [64]. Similarly, news posts and news articles represent their writer's view of points, thoughts, and emotions towards a particular event, and it is written by creative authors in a way to "provoke" emotions and attract their readers' attention. Typically, news posts consist of few words and contain a high load of emotions of their writers as they describe major worldwide

events. These characteristics make news posts become an appropriate material for this study's goal in detecting the significant events happening all around the world as well as investigating and following up the changes in sentiments associated with the opinions about the social events.

Opinions within news posts on SNs have a strong impact on our lives. Authors of the research [94], have stated that financial news has a big effect on financial markets and subsequently on economic. Furthermore, analyzing sentiments of news posts may help to predict stocks [95], assist government in making the right decisions about sensitive issues, as well predict the results of an election campaign. According to [1], SNs like Facebook has become an important communication medium in the significant period, which known as " Arab Spring", whereby most of Facebook users at that time were Arab people who had used it as a tool to share their feelings and opinions via sharing pictures, short posts, and videos to show the crimes and inequality between regions. Thus, it has been used to inform the world about what is going on in their countries in order to push people to think, care and interact with them.

Regardless, the important information of news posts on Facebook, analyzing this large amount of data is not a simple task. Data on SNs can be used to find temporal related information [96], and which could be used to explain the reasons behind the changes of sentiments for an event over a period of time either growing up or fading. The Majority of studies have been conducted on sentiment analysis of subjective text which contains a direct message to their readers, has a specific target, and the writers express their opinions freely so the readers can easily distinguish its

polarity orientation. In contrast, targets in news posts and articles are general; it may consist a group of people, complex events description or many entities.

Additionally, news posts usually contain indirect sentiment words which depend on the text contents like (kill, injure) rather than explicit sentiment words like happy, sad, and angry. Which consequently makes it more difficult to determine news post's polarity. Besides that, journalists may exclude themselves from directly express their opinions or sentiments. They might use other methods such as using others sentences (quoting), or by using complex discussion and argument structure, especially when writing about sensitive events [79]. Moreover, the writing style of news items differs from that in blogs since the writing of the former is formal, more objective and have opinions written by famous people while the latter one has a casual writing style, more subjective and the opinions may come from an unknown name [15].

2.4.3 Social Events and Associated Emotions

Social events involve a public performance towards an event that happens at a specific place and a particular time. These events take many forms such as natural disasters, conflicts, diseases, airplane crash, terrorism as well as sports, contests and entertainment events. Firstly, a natural disaster is a prime adverse event caused by natural processes of the Earth like earthquakes, tsunami, volcanic eruptions, floods and other geologic processes [97]. A natural disaster may cause property damages, loss of lives and leaves some kind of economic damages [1]. Secondly, disease is a special abnormal state, unrest of structure or function which affects a portion or whole an organism [98]. This can be caused by external factors like infections disease, or it may occur by internal dysfunctions such as autoimmune diseases.

Diseases usually affect people not just physically, but also emotionally as it can change one's perspective on life and one's personality. Thirdly, terrorism is usually defined as violent behaviour (or the threat of violent acts) that attempts to create some kind of fear (terror), committed for political, religious or ideological goals and which intentionally target or neglect the safety of civilians [99]. This kind of attack may affect the stock markets [100], damage government or private properties and cause losing lives as well as the deployment of panic, fear and instability in the attacked areas.

Fourthly, conflicts indicate to some kind of disagreements, friction, or discord arising among the members belonging to the same group or between members of two or more groups. Conflict is usually occurring due to the differences in opinions, disagreements, shortage of resources, or seize of power and governance [101], and its effects may include losing lives from either civilians or soldiers, economic damages, and generate some kind of instability, insecurity and unrest state in conflict areas. Finally, airplane crash accidents that defined as an accident in which an airplane may hit land or water and consequently crashed or destroyed [102]. This event may cause huge economic damage to Airline Company to which the airplane belongs as well as loss lives.

This study has focused only on these five specific social events (conflict, terrorism, airplane crash, disease, natural disaster) as these events are among the most frequent happening events during the last year 2014 around the world [48,49,50]. These events were among the most talked-about global topics on Facebook [51,52] and twitter [53,54]. In addition, the reason behind focusing on these specific topics/events is, that these events are most influential to other news

domains whereby analysing and knowing the sentiments associated with these events could help in obtaining information on other domains. For instance, the stock market prices could be affected by the small number of negative posts.

Moreover, each event usually associate with a set of opinions and emotions generated by the people who have lived these events. These human emotions contribute in everyday life activities as people share what they feel toward certain event, products, brands, people, situation, things, thoughts, dreams, sense, and memories. Thus, sentiment detection from the text allows us to identify the emotions of the writers towards those entities, and consequently improve human computer interaction system. For example, help consumers to take the right decision related to products, marketers to gather feedback about their products, and government to take initial steps to solve the current existing issues or to prevent them from occurring again in the future. Sentiments can be expressed via text using either words directly referring to emotional states such as “fear, happy”, or through indirect words that depend on the context, which means words imply to possible emotion like “killed”. The former is called direct affective words while the latter one is known as indirect affective words [103].

Sentiments/emotions have been categorized into various classes like joy, anger, fear and so on, where these classes have no clear boundary between them. However, many psychologists basically divided human emotions into six main types. Which are fear, anger, sadness, happiness, disgust, and surprise. They stated also that all other emotions are just varieties of the basic ones. For instance, depression, grief are varieties of sadness, while pleasure is variety of happiness, and so on. Robert Plutchik in his study [104] mentioned that the secondary emotions are actually

formed by merging various degrees of basic emotions such as surprise and sadness produce disappointment, and multiple emotions may form a single emotions like anger, love, and fear can produce jealousy.

Physiological and behavioural qualities characterize each emotion through specifying people's movements, posture, voice, facial expression, and pulse rate fluctuation. For example, sadness are characterized by tightening the throat and relaxes the limbs. The definition of the six basic emotions according to [105] are as follows: 1) Anger: is "the emotion that expresses dislike or opposition toward a person or thing that is considered the cause of aversion". 2) Disgust: is "the emotion that expresses a reaction to things that are considered dirty, revolting, contagious, contaminated, and inedible. 3) Happiness: is "the emotion that expresses various degrees of positive feelings ranging from satisfaction to extreme joy". 4) Surprise: is "the emotion that arises when an individual comes across an unanticipated situation". 5) Fear: is "the emotional reaction to an actual and a specific source of danger". 6) Sadness: is "the emotion that expresses a state of loss and difficulty".

Table 2.1.
Main Six Basic Emotions and Their Secondary Emotions

Basic Emotions	Secondary Emotions
Fear	Nervousness, Anxiety, Distress, Dread, Tenseness, Worry, Horror, Panic, Shock, Fright, Hysteria.
Sadness	Depression, Despair, Suffering, Hurt, Disappointment, Guilt, Shame, Grief, Neglect, Insecurity, Embarrassment, and Humiliation.
Joy	Cheerfulness, Delight, Happiness, Amusement, Thrill, Excitement, Relief, Optimism, Pride.
Anger	Rage, Hate, Frustration, Irritation, Disgust, Contempt.
Surprise	Shake, Amazement, Wonder, Shock, Disclosure, and Disbelief.
Disgust	Sickness, Loathing, Revulsion, Hatred.

Plutchik also mentioned that humans experienced not only primary emotions, but also secondary emotions. These basic emotions along with some subsequent secondary emotions are mentioned in the Table 2.1. In contrary, Glasgow University has stated recently, that human emotions are only four types which are happiness, sadness, disgust/anger, and fear/surprise, because they have found that fear and surprise looked very similar to the observers in the early stages then developed later due to social reasons, the same goes for anger and disgust. For example, both surprise and fear share raised eyebrows. However, different researches have been used other models and emotions classes for labelling the corpus data in order to do sentiment classification, more detailed about this will be introduced in Section 2.5.2.

In order to do sentiment analysis process of social events and their correlated opinions, data about those events should be collected. There have been many kinds of resources for social events such as traditional sources (interviews, questionnaires, magazines and newspapers) or modern sources (news channels, news websites, electronic magazines, newspapers and articles on the web as well as news posts on SNs like Facebook and Twitter). News channels posts on SNs such as Facebook are one of the main data resources that allow follow-up social events that occur around the world. However, before analyzing the opinions and sentiments embedded within those posts, there is a fundamental step that has to be done, namely, determine or detect the topic of these posts in order to follow the developments of the social events happening all around the world.

2.4.4 Topic /Event Detection

The wide spread of social media and as users become more free to express their opinions, this changes the way users generate and consume the news. Hence,

publishers, editors and newsreaders face some kind of difficulty in gathering and following up the world's events. An event is defined as "something that happens at a certain place and time", [106]. Users these days are involved in every stage of news making and consuming for example, as a commentator who is commenting on news items, or as a journalist that writes and publishes news through traditional or digital social media.

Recently, individuals, organizations, and news reporters have become interested in detecting and analysing topic's trends by tracking UGC on SNs, where topic concept and time factor are very important to understand the news trends, and their correlated sentiments [66]. Therefore, a new field has been emerged which is called as Trend Detection (TD) or a recent concept named Emerging Trend Detection (ETD), in which it is defined as " topic area that is growing in interest and utility overtime" [107].

Topic trend usually uses document corpus as input [9], then topic extraction process is executed. This process is executed on time-stamped corpus of document. Various computational works have been conducted in this area, which mainly used text mining as an approach to analyze the unstructured data over social media such as blog posts [23], and Google trends [72]. Trend of monitoring users' opinions has been studied lately in Twitter as this subject has attracted the attention of scholars and professionals [7–9]. In contrast, relatively a small number of studies have been noticed using Facebook [10–12] , to either identify trends, determine relevant topic classification, or detect posts' sentiments [1,5,12–14]. This situation may occur due to limitation of data volume because of privacy policies of the Facebook, which restrict the data to be collected.

In addition, topic classification has major problem which is the high dimensionality of the features which can decrease the classifier performance due to redundant and irrelevant features. Therefore an efficient method to solve this problem is by reducing the feature space dimensionality [35,42]. As a result, many previous studies [5,15,16] have addressed the this problem usually by using news corpus as input [9], then the document will represent as a vector of features in a text stream. After that topic key feature extraction process is executed where these feature extraction techniques, mainly aim to select a suitable set of keywords based on the whole corpus of news documents, and assign weights to those key words. Consequently, the new arrival document will be classified based on the similarity with the existing pool of events. If the similarity exceeds a specified threshold, the document will classify into the nearest events and update the new document. Otherwise, a new event will be formed [64]. Feature extraction methods usually treat the document as a group or Bag of Words (BOW). Different feature extraction techniques have been used for either text classification, information retrieval, or topic/event detection to extract the main key features. More detail about these techniques presented in the Section 2.5.4.

2.4.5 Sentiment Analysis for News Websites

As the number of online news websites has increased recently, this makes the process of browsing and retrieving news articles become more much easier than before. However, various news websites may have similar or opposite sentiment orientation to a contentious news topic such as (economic, political) based on their own views of points or interests. This orientation is not restricted to positive-negative sentiment only, but may include other types of sentiments. For instant, “Iraq war”,

basically reflect the sentiment of “disgust” or “acceptance”, while “sport games” usually associate with “happy” or “sad”. Besides that, sentiment orientation may vary overtime, and a single news website may always persist in consistent sentiment. The study [30] had introduced a system that extract news articles’ sentiments, find the sentiment variation within a single news website, and finally compare the sentiment orientation between different news websites.

Moreover, studies on news sentiment analysis primarily concentrate on the writer’s perspective, which try to identify the writer’s sentiment orientation toward a specific target. In contrary, in many cases the reader’s sentiments triggered by the news document do not usually agree with that of author’s. In addition, previous studies [21–23] have focused generally on the single label classification techniques which assign only one sentiment to the document from a set of multiple categories, while other research focus on multi label classification considering that a reader may have a combination of more than one base sentiment toward a particular entity [24,29]. Anyway, since this is not our area of study the author of this research will apply single label classification to news posts on Facebook.

2.5 Analytics/Data Mining Tools

There are various types of tools have been used for performing data analysis and data mining tasks. These tools could be programming languages such as R, Python, Java, and MATHLAB [108], or could be data analytics applications. These applications have been developed by various researchers like Weka, Orange, and RapidMiner.

2.5.1 Rapidminer Text Mining Software

Rapidminer is a software platform developed starting by Ingo Mierswa, Ralf Klinkenberg, and Simon Fischer at the Artificial Intelligence Unit of the Technical University of Dortmund [109]. This software provides an integrated environment for data mining, text mining, business analytics, and machine learning. It contains all steps of data mining processes with results visualization, validation, and optimization. It is used for research, training, education as well as business and industrial applications [110]. In 2013 and 2014 annual software poll KDnuggets [111] has ranked RapidMiner as the most common and popular data analytics tool with the majority of the poll respondents. In addition, RapidMiner achieved the most satisfaction ratings in the 2011 based on Rexer analytics data mining survey [112]. Moreover, it has gain over 3 million downloads and has over 200,000 users including eBay, Intel, and PepsiCo. Furthermore, RapidMiner has recognized as the market leader in the software for predictive data analytics against other competitors such as IBM, SQL server, Revolution analytics[113]. About 50 developers from all over the world participate in the development of the open source RapidMiner with majority contribution belong to the employees of RapidMiner.

Additionally, this software has more and better features for analysing than other data mining tools such Weka, R, and Orange, where it provides learning schemes, models, and algorithms from the other former data mining tools through the extensions. As a result, RapidMiner has become the leader of data mining tools market [113]. Hence, this study has used this software to do text mining and in order to build the topic and sentiment classification models for the news channels' posts on Facebook.

2.6 Temporal Sentiment Analysis Methodology

As a new merged field of sentiment analysis, temporal sentiment analysis has utilizes the same methods and techniques used by sentiment analysis as follows:

2.6.1 Data Collection Techniques

Today different types of business are heading in taking new directions, using the web to collect a large amount of data. Many online SNs are targeted by researchers in order to collect and mined data for better recognition, study attitudes and behaviours of users. Currently various datasets are available on the web such as blogs, reviews sites for product and services, movie reviews, and microblogging sites for expressing opinions about different topics such as Facebook and Twitter. These data usually are collected in various ways, such as providing surveys to ask users about their opinions in real time, reviewing information posted through SNs, where private and personal information may merge together, or by developing some kind of applications which allow to extract data. An example of these applications, is Application Programming Interface which is provided by most online SNs like "Facebook API and Twitter API". Another application is an automated script program that collects and explores the site by applying "HTTP Request & Responses"[114].

Twitter released its API streaming on April 2009 allowing high access to real time public data as well protect data. In contrast, Facebook provided similar API graph in April 2010 [16], but with a limitation in accessing its data in terms of how many times data can be accessed, which type of data can be collected, and what is the allowable size of collecting data. Conditions are varied significantly between

microblogging sites' APIs. For example, in contrast to Twitter, Facebook is quite restrictive due to its privacy policies which affect the functionality of API in four ways [115]: first, the actual status of a Facebook user will define which data are available to be collected. For instant, any detailed user's data can only be collected from accounts that a user is friend with or has to be a member of a group. Second, extracting data from a user account depend on the account's privacy settings either allowed or blocked. Finally, any application tries to access any data will proceed a request in order to gain access, where this request will appear as the user uses the application for the first time.

Facebook API allows access to various objects (for instance, people, pages, or posts). For this study the posts type is of interest. Every post consists of the following information: 1) Post's details like name, caption, description, and message, 2) User who posted the message, 3) Type of the message (for instance, photo, status, video, or link), 4) Time of post, 5) Application used for posting. Additionally, Facebook API returns a limit number of posts for each query (default 25 minimum, 500 maximum) [16].

Several studies have been implemented Facebook API using different development tools to extract data. In [16], researchers collect public posts by running a simple algorithm that contains a loop of twenty six search queries. Similar data type and tool were used also in [1], nearly 260 status posted by Tunisians users were collected during the revolution (January- December, 2011) in order to analyze their behaviours. Others researches developed tools such as Facebook Restfb JavaAPIs in [22], to collect 2000 posts from various users, and Facebook Harvester tool for collecting Facebook comments related to a particular topic [116]. While the authors

in [117], developed a tool called Status Puller for extracting random posts from either Facebook or Twitter.

Recently, several data extractors' tools have been developed targeting Facebook and utilizing its API for gathering data. These tools in general export data in common format (for instance, xml, csv, or xls), and focus on specific sections of the platform namely on pages' contents due to page showcase feature. An example of these tools are Netvizz, Nodexl, Facepager, Infoextractor, and Screen Scraper. This study will use Netvizz as a tool to collect Facebook posts.

2.6.1.1 Data Collection Tool (Netvizz)

The Netvizz application was developed by the author of [118] in 2009 as a practical experiment to study Facebook's API as a new media entity. Netvizz is a simple application on Facebook written in PHP and runs on a server provided by Digital Method Initiative. It can be found by typing its name in the search box on the platform like any other Facebook's applications, however, it requires the user to be logged in, using an existing Facebook account.

Additionally, Netvizz currently support three types of modules for extracting data [115]: 1) Group data: produces networks and tabular files for both interaction within groups and friendships, 2) Page like network: generates a network of pages connected via the likes between them, 3) Page data: produces a network and tabular files for users' activity around posts on pages. The last module gets last posts either specific number such as 50 as minimum or 999 as a maximum, or gets posts which are between two specific dates. Furthermore, it supports various choices to get posts like by page only, page and users, or posts together with the network and comments.

Hence, due to all previous features of Netvizz, this research has selected this tool as a medium to collect data.

2.6.2 Labelling Techniques

Lately, the number of internet online users grows rapidly and many of them are willing to take part in the social interactivity. Thus, many websites nowadays like Yahoo, Amazon, MBDI, and Sina society news websites are heading toward allowing the users to express their sentiments to the various entities (blogs, products, movies, news articles and posts) using various novel services such as stars rating [33], voting systems to select a sentiment from a set of emotions provided by the website [21,24]. Furthermore, and before starting with next phase of Pre-Processing data, some studies would perform labelling process on the data. In general, there are two main approaches for labelling: supervised and unsupervised approaches. A supervised approach such as machine learning techniques in which a classifier will be trained on a pre-defined corpus of documents, so it would later contribute in labelling the other documents.

In [83] authors extracted dataset of movie reviews from the Internet Movie Database (IMDB) for sentiment classification whereby only reviews with rating stars are collected, then ratings are extracted and converted into one of categories, positive, negative, on which then a classifier was trained. While in [22,25] machine learning techniques have applied to automatically label news headlines using annotated training dataset called ISEAR.

In contrary, unsupervised approaches such as symbolic and clustering techniques. The main idea of symbolic technique is to assign each term (feature) a

sentiment score, which consequently will determine the polarity of the term (positive or negative). Once the score is calculated for each term, then the score of the whole document can be computed through applying (sum or average) functions. However, the crucial step for this technique is to specify the method used for scoring the terms.

Various types of scoring methods have been introduced such as human scoring via asking a number of people to score the terms [119], this called manually hand labelling. Categorization process has done on twitter data in order to categorize tweets into one of 18 general classes [13], where tweets for training dataset have labelled manually after reading the definition of topic's trend as well some tweets. Also, researchers in [22] labelled 2000 Facebook posts into either (sad, happy, or neutral) by using lexical online dictionaries like WordNet in which the distance between the term and the reference/root word will be calculated [34].

In recent years, there have been a lot of online lexica resource available for analyzing proposed of either sentiments or emotion analysis. For example, Senti-WordNet (SWN) is a lexica for single sentiment labelling to either (positive or negative) with scores ranging from 0 to 1. On the other hand, ANEW is a source for multi-class label (very-negative to very positive. However, each lexica resource differs from the other in its coverage space for the terms as well as the sentiment scoring values. Compared to sentiment lexica emotion lexica has been generated also but with far less coverage than sentiment lexica. One of the popular used resource is WordNetAffect, which consist of manually assigned labels to WordNet synsets (fear, joy, surprise, etc.). It offers 900 annotated synsets and 1.6k words, while AffectNet contains 10k words and extends WordNetAffect labels for concepts like "have

breakfast". Other emotion lexicons are fuzzyAffectLexicon , Emolex and Affect database which is the only lexicon providing scores for each emotion category [35].

Additionally, due to the low coverage issue for online lexica resources, some researchers have gone on creating their own emotion lexicons through utilizing annotated documents with emotions. This idea is not new, but has the limitation of offering a single emotion label rather than score for each emotion, and moreover, the annotation was usually done by the author. Various studies have created their own emotions lexicons such as authors in [8] produced Plutchik's four dimensional model: "joy <-> sadness", "acceptance <-> disgust">, "anticipation <-> surprise", "fear<-> anger", to classify the sentiments of news articles, while a three dimensional model of emotions: "happy <-> sad", "glad<-> angry", and "peaceful <-> strained", was generated since authors found those emotions more suitable for news articles [60].

In addition, another type of unsupervised approach for labelling is the clustering based method. This method aims to discover natural groupings, provide an overview of the classes in a collection of documents, does not require a pre-define class of a document, and does not need a training process. Thus, it saves time and it is free from human participation [120]. One of the most commonly used technique in clustering is basic k-means. However, all those former ways have some kind of drawbacks such as symbolic and clustering techniques obtain relatively very low rates of accuracy, while supervised learning approach, although it had achieved good results, but it has considered to be very costly and time consuming.

2.6.2.1 Depechemood an Emotion Lexicon

As mentioned before many studies have created their own emotions lexicons, however, some of them have made their models available online for free on the web for research proposes such as TheySay PreCeive API Demo [121], sentiment analysis with Python NLTK text classification [122], DepecheMood [123].

DepecheMood is a high coverage and high precision lexicon consists of roughly 37 thousand terms annotated with scores for eight emotions (afraid, sad, amused, happy, don't care, angry, inspired, annoyed). This lexicon has created by Jacopo and Marco in [35]. Their approach has exploited in an original method "crowed-sources", which were annotated in affective way "by readers" of news articles from rappler.com news website. The researchers have built their lexicon by building term-by-emotion matrix.

In order to produce this matrix the researchers conducted several steps. First, they obtained document by emotion matrix, where each document was annotated with the score for each eight emotions provided by the website. This score has obtained through returning the percentage of votes for each emotion labelled by the readers of a given story. Second, word or term by document has constructed after applying lemmatization and POS tags on all news documents harvested from rappler.com and kept only terms of type adj, nouns, verbs Adv as well as those which present in WordNetAffect, WordNet, and SWN-prioritize resources to which they had aligned with them. Then, the researchers computed term-by-document by using three weighting techniques, namely TF-IDF, raw frequencies, and normalized frequencies in order to determine which of the three weights is better. Finally, they do multiplication between document-by-emotion matrix and term-by-document

matrix to produce the word-by-emotion matrix and consequently generate their emotion lexicon.

In addition, to evaluate the performance of their lexicon, Jacopo and Marco have used the public dataset provided for SemEval2007 "Affective Text" [124]. Their experiment results have shown high precision and high coverage of the lexicon compared to other sentiment lexicons like (ANEW, Warriner, SWN-prior) as well as other emotion lexicons such as WNAffect, also indicate significant improvements over unsupervised methods. As a result, this study has used DepecheMood emotion lexicon as a tool to label each news post into its suitable emotion category from the eight emotions classes provided by the same lexicon. Consequently, each news post has been assigned to a single emotion category with the highest score.

Additionally, this study has chosen this specific free lexicon version among other free tools because this lexicon has built based on news crowed-source articles, which is quite similar to the datasets used for this study news post on Facebook. On the other hand, the other tools have built either on movie or product reviews that are not appropriate for the goal of this study, which is, to classify the news text into its corresponding emotion category. In the next Section an introduction to Pre-Processing stage and its commonly used techniques will be presented.

2.6.3 Pre-Processing Techniques

The huge variety of real time data available on the SNs like Facebook and Twitter usually include plenty of noisy terms along with some unclear parts such as "HTML tags, scripts, and advertisements", [125]. Thus, pre-processing phase is required to reduce or eliminate those noisy parts in order to enhance and increase the

performance of the classification process. Pre-processing is the process of preparing data for classification phase. This phase includes several steps as mentioned in [126]: "online text cleaning, stemming, expanding abbreviations, white space removal, stop word removal, negation handling and finally feature selection called filtering".

Stop words are words that have no significant meaning or sentiment in the dataset (i.e., conjunction, prepositions, pronouns, and, the, or). Consequently, removing them will reduce the size of the dataset for training and testing process and further simplifies the targeted analysis. Stemming is the process of removing affix and derive the word into its stem or root form. Affix is a verbal element that attaches at either end (suffix), beginning (prefix), or middle (infix) of the word. This process contains a series of steps, where at each step a specific type of affix is removed [1], [79]. The porter algorithm is the most used algorithms for this kind of task invented by Martin F.Porter in 1980 [127]. Additionally, other stemming algorithms have been used such as "R stem, Snowball, Wordstem [126], Reverse porter, and Back-Forward algorithm [79].

On the other hand, different linguistic techniques have been also used for Pre-Processing process like "Tokenization, Entity Extraction, Part Of Speech tagging (POS), and Relation Extraction", [128]. A word tokenization algorithm is a simple process that has been applied to identify the occurrence of a single word in a document. Means, representing each word along with the number of its occurrences [79]. POS tagging is a very important monitor for the position of the term within a sentence either (noun, verb, adverb, or adjective), also can be used as an indicator of sentiment expressions since same word may have various meanings depending on its usage. Negation words consider to be very important as to evaluate sentence's

polarity since they convert the sentiment orientation. For example, “the sentence, I don’t like this mobile”, has negative polarity.

A wide variety of sentiment analysis has applied different combination of pre-processing techniques. Some apply only stemming and stop words removal [1], [66], while others merge them with other steps such as POS tagging and tokenization [22], or URL removal as in [72]. However, it has found that most common two steps in pre-processing are stemming and stop words removal, especially for studies which are applying Term Frequency- Inverse Document Frequency (TF-IDF) technique for features extraction [54,67,102,103], along with URL removal [16], and tokenization [13]. In the next section, feature extraction phase will be explained along with its various techniques which have been used in different studies.

2.6.4 Feature Extraction (FE) Techniques

The increasing amount of opinions and reviews on the web makes sentiment analysis become a hot topic for analyzing data. Therefore, selecting the key feature has become a crucial step in determining the performance of later classification step. FE phase is very important to select the significant features which hold useful meaning, and eliminate those features that do not help in distinguishing between the documents. Consequently, FE techniques discard the features that either appear very often or very few times in the whole corpus.

Furthermore, FE phase is considered to be more important than classification phase in highly imbalanced situations [131]. Where choosing features with strong class information can enhance classifier performance: for instance, the word “goal” often appears in the class “sports”, [132]. Furthermore, due to the increasing volume

of published news posts and articles, it is becoming almost impossible to extract the key features manually as it is very difficult to do and also consider to be extremely time consuming. Thus, for obtaining the key features an automated feature extraction techniques have been introduced

FE techniques usually treat the document as a group or Bag of Words (BOW) which is frequently used or as a sequence of words within the document. In [29] authors set up an experiment in order to determine the appropriate robust features for emotion classification for ISEAR dataset. They applied BOW, POS tagging, length of sentence and lexical emotions attributes. Their experiment results showed that BOW features obtain better accuracy as Boolean representation compared to term frequency.

Diverse feature extraction methods have been introduced such as lexicon based methods, machine learning techniques, and statistical methods. Lexicon based methods usually begin with a small set of "seed" words. Then, enlarge the dataset by finding the synonyms from various online resources. Taboada et al [133] present this method by utilizing negative and positive words associated with their polarities values from a specific dictionary for classification tasks, then calculate their semantic orientation. This approach has shown 59.6% to 76.4% accuracy on documents from movie dataset. Machine learning techniques also have been used for FE in [134] where SVM has been trained to identify expressed opinions which are related to specific target in a dataset of the car and camera reviews in which both target and opinions were already annotated. The main statistical feature extraction techniques are as follows:

2.6.4.1 Term Frequency-Inverse Document Frequency (TF-IDF)

One of the most common extraction techniques which have been used to detect and identify the topic of a document is Term Frequency- Inverse Document Frequency TF-IDF in order to extract the most frequency terms in the document which consequently will contribute in detecting the document's topic. TF-IDF is a popular way to compute terms' weights using the following formula:

$$TF\text{- }IDF = TF * \log N/n \qquad (2.1)$$

Where N is the total text of all categories, n is the number of texts which contain term t. The main idea of IDF is that whenever n is smaller, IDF is larger, and term t would have the better category differential ability. Consequently, this term should be appearing frequently in the text but oppositely in other texts.

In [49] researchers try to select only the relevant opinionated sentences for opinion summarization by determining if the sentence focus on a specific topic. Hence, search for a term which can be used to form the major topic based on the idea of Fukumoto and Suzuki [102] that considers a term to be representative of a document's topic if it appears frequently in each document or across the set of documents. Therefore, scholars compute TF*IDF scores for each term at both document and paragraph levels with a bit modification in the original formula for paragraphs. Other formulas were used also to denote how frequently the term appears in each paragraph and document. Additionally, specific ThresHold (TH) was used in order to control the number of representative terms in a relevant corpus. Where the big TH is, the more representative terms will be included. The study's results showed good performance in extracting relevant sentences.

Similarly, researchers in [42] convert each token into its TF-IDF weight to find out the most common terms for each 18 labelled categories. Where these terms will be used to build the training set for machine learning in order to predict the topic of tweets. Identical to this study, but using Facebook posts authors in [1] used modified TF-IDF to predict Facebook post's topic because they found that the original formula is not applicable for Facebook posts due to the limited length of the post which would decrease the value of term frequency. Besides that, if frequency over all corpus is calculated, this would end the whole corpus with a single post only, hence losing the IDF component. Therefore, in order to overcome this problem the researchers modify the formula so the sum of all term occurrences over all posts exchange with term occurrences per post only.

Same problem of TF*IDF faced by the authors of [41], but with a bit different in terms of categories. Where the researchers used an improved TF-IDF to solve the problem of wide range of TF instead of IDF as it was in [1]. Because TF neglects the frequency in texts for specific category as well as can not represent the distribution of features in a certain category, degree and distinguish between categories. Thus, the authors added new weight to the original TF-IDF, which considers the frequency of the feature in a particular category for the whole text corpus. Hence, the new formula become as following:

$$TF \times IDF \times D = TF \times \log N/n \times (p-q+1)/n \quad (2.2)$$

Where D is "the value between the maximum and second maximum number of texts containing term t in a certain category". Moreover, TF-IDF used in [59,60] to extract the relevant sub topics key words of news articles and based on these key

words other news articles had retrieved in order to analyze them and present the visual sentiment tendencies of various news websites.

2.6.4.2 Term Frequency (TF)

TF weight is the standard method used in information retrieved since it indicates the most significant relative features which can be used to represent the document. TF has various weighting schemes such as Binary Term Occurrences (BTO) which uses the presence of a term (value 1) or not (value 0). On the other hand, Term Occurrences (TO) uses the presence of a term as TRUE or not as FALSE. Although some studies have shown that binary weighting is better compared to a frequency scheme for polarity classification task, the feature's frequency is much more applicable for topic categorization. Perhaps due to this fact that topic categorization highly depends on content features that seem to be repeated. However, this finding is still an open area of research. Pang et al. [83] achieved better performance using a binary values feature. Similarly, Akaichi et al. [1] apply the binary technique for Facebook status classification through using seven various combinations of n-gram features with two different machine learning techniques (SVM, NB). The results showed that SVM outperforms NB, where SVM best performance was by using unigram as n model with 72.78% accuracy. In contrast, NB optimal performance was by using bigram with 69.42% of accuracy.

2.6.4.3 Chi-Square (X^2)

Chi Squared Statistic technique calculates the weight of attributes with respect to the class attribute. The higher the weight of an attribute, the more relevant it is considered. The chi-square statistic is a nonparametric statistical technique,

which is used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies. Chi-square statistics use nominal data, thus instead of using means and variances, this test uses frequencies. The value of the chi-square statistic is given by:

$$X^2 = \text{Sigma } [(O\text{-}E)\ 2/\ E] \qquad (2.3)$$

Where X^2 is the chi-square statistic, O is the observed frequency and E is the expected frequency. It is proved to be better in performance than PMI since it is a normalized value and its value is more comparable across terms in the same category [99]. Chi-square is used for FE by Fan and change [100] to specify blogger's immediate interests from real ads and actual blog pages in order to improve online contextual advertising. Their outcomes showed effectiveness in identifying those ads which are positively correlated with blogger's personal interests.

2.6.4.4 Information Gain (IG)

Information Gain computes the weight of attributes regarding to the class attribute by using the information gain. The higher the weight of an attribute, the more related it is considered. Let Attr be the set of all attributes and Ex the set of all training examples, value(x,a) with x € Ex defines the value of a specific example x for attribute a € Attr, H specifies the entropy. The values (a) function denotes set of all possible values of attribute a € Attr. The information gain for an attribute a € Attr is defined as follows:

$$IG(Ex,a) = H(Ex) - \sum_{v \in values(a)} \left(\frac{|\{x \in Ex | value(x,a) = v\}|}{|Ex|} \cdot H(\{x \in Ex | value(x,a) = v\}) \right) \qquad (2.4)$$

The main problem in information gain is that whenever it applies to attributes that include a large number of different values. For instance, assume some data that defines the customers of a business. When information gain is employed to conclude which of the attributes are the most significant, the customer's credit card number may have a high information gain. This attribute has a high information gain, because it uniquely classifies each customer, but we may not want to assign higher weights to such attributes.

In addition, some studies were conducted to compare between the performance of classification process using various data sets as well as different FE techniques and feature types. One of these studies done by Yelena and Padmini [130] the authors include different kind of features in their experiment n-grams (unigram, bigram, trigram, and phrase), also test various pre-processing techniques (stemming, negation) on three diverse sizes of datasets (movie reviews, product reviews, and multi-domain). Their outcomes have shown no any significant changes in terms of using either term frequency or the presence of the term for all three datasets. In addition, 1-gram and 2-gram has achieved best performance for smallest dataset while for the other two datasets the best performance achieved for 1-2-3 grams.

In contrary, using the phrase as feature showed decrease in accuracy for all datasets. Another similar study done by Bo Pang et al. [83] who try to classify movie reviews using also different kind of features based on frequency or presence as well using three various machine learning algorithms (SVM, NB, ME). The results showed better performance than in [1] for both SVM and NB at 82.9% and 81.5% respectively. Thus, as a result of the two latter studies, it is found that n-gram features should be chosen accordingly to the size of the dataset.

Fengxiang et al. [37] handle the problem of imbalanced data in text classification through comparing their new feature extraction technique named NFS with other FE techniques such as Chi-square and Mutual Information (MI). Where NFS technique depends on selecting the words that hold class information rather than selecting the terms with high document frequency. Furthermore, to overcome the problem of imbalanced data the researchers combine their FE technique with oversampling resampling technology. Their experimental results showed that NFS performs better than statistical Chi-square and MI when the features are more than 1000. In contrast, Chi-square was better when the features are from 100 to 500. Also the results have shown an improvement in performance of the classifier after applying the resampling technique.

Moreover, some studies have applied burst detection technique in order to detect robust features. Burst is defined in [135] as "an unusual growth in intensity of an observation of interest". A very simple approach for detecting robust features, is to find the relative frequency of the features and use threshold to control the number of detected features. Balog et al. [136] applied thresholding to detect peak times of moods and explain the abnormal changes specified by events due to overuse of some features in the interval.

However, this method has a problem which is, how to determine the appropriate threshold. Another method has introduced that depends on learning the hidden states generating the observations of time series like Hidden Markov Model (HMM) [137] which is identical as Kleinberg [135] who states "the higher the state is, the higher the rate of generating". Where Kleinberg has noticed that specific topic

in his email corpus were easily characterized by a sudden effectiveness of message sending, rather than using text features of the message.

A little different from [136] the authors in [3] applied burst detection algorithm, but with different threshold in order to detect and extract robust features to specify the abnormal changes in hotel reviews and try to analyze why such changes have occurred. Their algorithm based on a simple unigram frequency with specified thresholds. The first thing their burst detection does is to check if the monthly average of the sentiment score has changed. If it is changed, then there is a possibility of a change in sentiment. However, they found out that this phenomenon may be a one-time event. Therefore, they try to check if the sentiment score for the next month is stable to eliminate possible fluctuation. Further, in order to discover why those changes have occurred, the authors applied simple feature extraction by grouping two words together such as adverb with verbs in one pair, and adjective with nouns as another pair.

In another study [83], researchers use the same idea of term frequency for sentiment classification, but by applying various types of features (unigram, bigram, trigram, POS, and position) with Machine Learning Techniques (MLT) such as SVM, NB, and ME. The results showed that using unigram presence as a feature for SVM has achieved the best performance. Also they have found an increasing in performance when select top 2633 unigram features. Similarly, in [10], the authors use n- grams features along with entity word and dependencies as features for MLT, and in order to reduce the number of features for training set they select top 250 features with the most correlation with the class by ranking the features using Chi-

squared measure. The next section will introduce classification phase and its famous methods and techniques used.

2.6.5 Classification Techniques

A classification process in sentiment analysis contains two types of classification, i.e., Binary classification (e.g., negative, positive) or Multi-class classification (e.g., strong positive, positive, neutral, strong negative, negative). Majority of prior studies has focused on the former type [84]. There are various techniques for Sentiment Classification (SC): Lexicon based approach, Machine Learning (ML) approach and Hybrid approach [138].

Lexical approach is based on identifying the opinions lexicon which is utilized for analyzing the text. This approach has two methods: 1) Dictionary based method which relies on specifying opinion seed words, and then find their synonyms and antonyms from online source dictionaries. 2) Corpus based method which begins with a seed list of opinion words, then search for other opinions words from the large corpus in order to help in finding the opinion words within context using semantic or statistic methods. Previous works that applied lexicon based approach like the one done by Ohana and Tierney [139], who used SentiWordNet for classifying 1000 movie reviews. Their results showed less accuracy around 65% compared to [3] which achieved roughly 70% to 80% accuracy for hotel reviews where it is considered to be very good.

The ML approach depends on using the famous ML techniques such as (SVM, NB, ME) which makes use of syntactic and/or linguistic features. Subsequently, this approach is divided into two methods: 1) Supervised learning

which is split into training and testing datasets. The training set is usually labelled into one or more classes using manual hand or automatic labelling techniques. Furthermore, this dataset should be varied and rich to support the algorithm with various types of text as well help the classifier in the future. Testing dataset is used to evaluate the performance of the trained classifier to see if the classifier is able to assign a new text to its correct class [1]. 2) Unsupervised methods are applied when it is hard to find a labelled training dataset.

On the other hand, Hybrid based approach uses a combination of methods from the two previous approaches where it is very commonly used with sentiment lexicon. Rudy Prabowo in [140] apply a Hybrid approach in which the author combines supervised ML technique SVM with rule based classifiers like Rule Based Classifier (RBC), Static Based Classifier (SBC), and General Inquire Based Classifier (GIBC). The main idea of this approach was if one based rule classifiers could not classify the document, then it will pass the document to the next classifier in line until the document is classified or no more classifier exists. In case no one of the previous classifiers classifies the document, then it will be given to SVM to classify it.

2.6.5.1 Machine learning Classifier (SVM)

Two of the most widely used ML techniques are Support Vector Machine (SVM) and Naïve Bayes (NB). NB classifier is defined as “a probabilistic classifier based on probability models that incorporate strong independence assumptions among the features”, [1]. Meanwhile, SVM classifier [141] is “a discriminative classifier which seeks a decision surface to separate the training data points into two

classes and make decisions based on the support vectors that are selected as the only effective elements in the training set", [77].

The basic SVM proceeds a set of input data and predicts. Each given input data that of the two potential classes consist of the input, making the SVM a non-probabilistic binary linear classifier. In training set, each example is marked as belonging to one of two classes. Consequently, SVM training algorithm shapes a model that assigns new examples into one class or the other. An SVM model represents the examples as points in space, mapped so that the examples of the distinct categories are divided by a clear gap which is as wide as possible. Subsequently, new examples are then mapped into that same space and predicted to belong to a class based on which side of the gap they fall on.

In other words, a support vector machine creates a hyperplane or set of hyperplanes in a high- or infinite- dimensional space that can be used for classification, regression, or other tasks. In addition, a good separation is obtained by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin). Generally, the larger the margin the minor the generalization error of the classifier. Thus, the original problem may be stated in a finite dimensional space, it frequently happens that the sets to distinguish are not linearly independent in that space. Hence, it was suggested that the original finite-dimensional space be plotted into a much higher-dimensional space, apparently making the separation easier in that space. Additionally, to keep the computational load sensible, the mapping used by the SVM schemes are intended to ensure that dot products may be figured easily in terms of the variables in the original space, by defining them in terms of a kernel function K(x,y) designated to fix the problem. The

hyperplanes in the higher dimensional space are “defined as the set of points whose inner product with a vector in that space is constant”.

However, many studies have developed SVM so it can be used in sentiment classification for Multi-classes. Authors in [1] tend to use both SVM and NB as classifiers for Facebook statuses due to their good performance in previous works such as [71,115] for NB as well as [12,71,115–118] for SVM as the best performance for classification task compared to the other ML techniques. Cui et al [146], state that SVM is more appropriate for sentiment classification in a large data set because it can differentiate between mixed sentiments exist in the same review. As well as, performs better than generate models. In contrast, NB proved to be a better classifier in small dataset.

In [10] the researchers discuss the main issue involved in classifying the real media data such as newspapers and magazine articles in electronic forms into positive and negative favourability. These data consider to be imbalance due to the different distribution of the documents arriving over time. Hence, affects the performance of ML model. Therefore, in order to overcome this issue the authors applied various features with ML classifiers namely (SVM, NB, J48, RBFNet, JRip) as well as balance the training dataset by using under sampling method and use geometric mean for evaluation. Their outcomes showed an improvement in performance for all classifiers after balancing, but adversely affected for NB.

Furthermore, many studies have been conducted to compare the performance of ML classifiers with different set of features and the corpus of data used. In addition, most of these studies have found that SVM outperforms other ML algorithms in sentiment classification. Similarly, Songho Tan in [147] conducts an

experiment study for sentiment classification on Chinese documents. The researcher compares four various feature extraction techniques (MI, IG, CHI, and DF) along with five ML techniques (centroid classifier, K nearest neighbour, window classifier, NB, and SVM). The study reported that IG is the best selection technique while SVM is the best classifier for sentiment categorization. Same result for SVM was achieved in [148] for distinguishing reviews.

Another comparison study has done in [22] in which authors applied different classification algorithms (NB, SVM, J48, BayesNet) in order to classify Facebook news posts into life or entertainment classes. Where the best accuracy has obtained by the SVM classifier followed by BayesNet at 94.75% and 94.69 respectively. Furthermore, the same study classifies the life posts into sad, happy and neutral using SentiWordNet dictionary after apply POS tag for each term in the post. In the next section, an evaluation phase will be introduced with its various metrics definitions.

On the other hand, some recent studies have been done in analyzing sentiment and classification of news article contents or news headlines. In [32], [149] authors apply machine learning technique SVM for sentiment classification of headline news into several types of emotions and compare their proposed method with other models. Their proposed models using SVM showed better performance than the others.

Other researchers in [16,23] investigated the similarity and opposite of sentimental tendencies between news websites. Where both studies had used TF-IDF technique for extraction features that used to represent the topics of retrieved news articles, and then utilize manually constructed sentiment dictionary of four different sentiment dimensional for classifying the sentiments of the news articles. However,

all former studies focused on classifying the sentiments of news items from a writer's perspective like this research, while studies, like [21,24] concentrate on classifying sentiments based on the reader's perspective into either single or multi sentiment classification.

2.6.6 Evaluation Techniques

In order to evaluate the performance of feature extraction techniques for topic categorization and sentiment classification, various metrics have been used by different studies. These metrics such as Accuracy (A), Precision (P), Recall (R) and Root Mean Square Error (RMSE). Subsequently, the common method for calculating these metrics is depend on the confusion matrix as shown in Table 2.2.

Table 2.2
Confusion Matrix

	Predicted positives	**Predicted negatives**
Actual positive instances	# of True Positive instances (TP)	# of False Negative instances (FN)
Actual negative instances	# of False Positive instances (FP)	# of True Negative instances (TN)

The definition of these metrics as follows:

i. Accuracy (A): "the portion of all true predicted instances against all predicted instances".

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.5)$$

ii. Precision (P): "the portion of true positive predicted instances against all positive predicted instances".

$$Precision = \frac{TP}{TP + FP} \quad (2.6)$$

iii. Weighted Mean Precision (WMP): "The weighted mean of all per class precision measurements, and it is calculated through class precisions for individual classes".

$$WMP = \frac{\sum_{i=1}^{c} TPi}{\sum_{i=1}^{c} TP\,i + \sum_{i=1}^{c} FPi} \quad (2.7)$$

iv. Recall (R): "the portion of true positive predicted instances against all actual positive instances"

$$Recall = \frac{TP}{TP + FN} \quad (2.8)$$

v. Weighted Mean Recall (WMR): "The weighted mean of all per class recall measurements, and it is calculated through class recalls for individual classes".

$$WMR = \frac{\sum_{i=1}^{c} TPi}{\sum_{i=1}^{c} TP\,i + \sum_{i=1}^{c} FNi} \quad (2.9)$$

Where C is the total number of classes.

vi. Root Mean Squared Error (RMSE): "a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed",

In order to get (RMSE), first the residuals should be determined. Residuals are "the difference between the actual values and the predicted values", represented by :

$$y_i\hat{} - y_i$$

where y_i is the detected value for the ith observation and $y_i\hat{}$ is the predicted value. Then, getting the square of the residuals, calculating the average of the squares, and taking the square root generate the RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i\hat{} - y_i)^2}{n}} \qquad (2.10)$$

Additionally, for further evaluation of the best feature extraction technique on unseen datasets for both topic categorization and sentiment classification addition metric has been used besides the previous metrics called confidence score. This metric is a measure of the accuracy and repeatability of a statistical test often used to decide how confident the prediction of each category either topic or sentiment category.

2.7 Imbalanced Data and Resampling Techniques

The problem of imbalanced data is often occurred in temporal sentiment analysis as well as in text classification usually because of differences in the distribution of class samples. This generally affects the classifier performance, especially on minor classes where the performance would be poor due to the overwhelmed of large classes [37]. Previously, there is a number of solutions proposed for class-imbalance problem at both algorithmic and data levels. Algorithmic level solutions focus on adjusting or optimizing the existing algorithms or design new algorithms in order to improve the performance of classifiers.

While for data level solutions various forms of re-sampling would apply such as directed oversampling (in which the choice of samples to replace is informed instead of randomly chosen also no new samples are created), random oversampling

with replacement, random undersampling, directed undersampling (similar choice of replacement is informed), oversampling with informed creating of new samples, and combination of all previous techniques [150].

Although, undersampling and oversampling are the most commonly used techniques to overcome the imbalanced data problem, each of them has its own disadvantages. For instance, the undersampling can remove certain important samples, while oversampling can add more computational task if the data set is already large but not balanced. Thus, how much undersample or oversample is usually detected empirically. Anyway, oversampling method has become more famous compared to undersampling technique. That is because it is simple to be implemented [151], and using it would avoid loss of necessary information [36].

2.7.1 Oversampling Technique (Bootstrapping)

In oversampling technique the dataset is divided into two sub-sets, one for minority classes while the other for majority classes. Consequently, the bootstrapping algorithm is applied on the minority classes' dataset that generates a multiple randomly resampled subsets which have identical size as the original minority subset [151]. Then, the dominated classes' dataset is combined with resampled sub-sets to construct the resampled training dataset that is ready to use in subsequent training procedures.

In addition, widely held studies in feature selection for imbalanced data sets have dedicated on text classification or web categorization [125,126]. Batista et al. [154] have presented a comparison study of different sampling techniques. They concluded that combining over and undersampling techniques like SMOTE+Tomek

or SMOTE+ENN is suitable when there are a few examples of the minority class or when data sets are highly imbalanced. Fengxiang et al. [37] applied oversampling technique that showed an improvement in the classifier performance for minority class. In contrast, Daoud et al. [10] used undersampling technique in which random samples were taken from the majority class in order to balance the minority class. Besides that they used geometric-mean as the evaluation measure for the SVM classifier. While in [38] the researchers tackle the issue of learning from a number of imbalanced Rail dataset through applying new iterative SVM algorithm with bootstrapping based undersampling and oversampling resampling techniques. Their results have shown that SVM is a suitable algorithm for resampling rail data classification problem. Furthermore, the output results have shown that undersampling technique surpass the oversampling technique with a significant performance and yields to reduce the complexity of memory and training time.

2.8 Chapter Summary

In this chapter, literature review has presented for the previous works, which have been done in the same area of this study. What kind of problems have encountered, how they have solved them, what kind of methods and techniques have been used in order to overcome all obstacles they had. Furthermore, explaining the main concepts within the title of this study.

LITERATURE REVIEW TABLES

Study	Dataset	Data Collection Tool	Label Method	Pre-Processing	Types Of Features	Feature Selection Method	Classification Technique	Evaluation Metrics	Comments
Feature Selection Method To Handle Imbalanced Data In Text Classification (2015)	Subset Of The Top 10 Categories From Reuters		Already Labelled Data To 10 Categories			NFS, CHI, MI	SVM	F1micro, F1maccro	Minority Class Is Oversampled To Improve The Classification Performance NFS Is Better When The Features Above 1000 Otherwise CHI Is Better
Temporal Sentiment Analysis And Causal Rules Extraction From Tweets For Event Prediction (2015)	Twitter Data				Unigram	Alchemy API	SVM	Precision, Recall, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)	The Study Identifies The Causal Rule Between Aspect Key Words In Order To Predict The Events
Depechemood: A Lexicon For Emotion Analysis From Crowd-Annotated News (2014)	News Articles From Rappler.Com		Already Labelled	Lemmatized, POS Tagging	Unigram	Frequencies, Normalized Frequencies, TF-IDF	Wordnet, SWN-Prior, Wordnetaffect Resources	F1measure	This Study Aims To Build A High Precision Lexicon Of 37000 Terms Annotated With Emotion Scores
Support Vector Machines For Class Imbalance Rail Data Classification With Bootstrapping Based Over-Sampling And Under-Sampling (2014)	Rail Data		Already Labelled				SVM	Accuracy, Specificity, Sensitivity	This Study Has Showed That Bootstrapping Undersampling Leads To Better Accuracy Performance That Bootstrapping Oversampling

LITERATURE REVIEW TABLES

Study	Dataset	Data Collection Tool	Label Method	Pre-Processing	Types Of Features	Feature Selection Method	Classification Technique	Evaluation Metrics	Comments
Exploration Of Robust Features For Multi-Class Emotion Classification (2014)	ISEAR Dataset		Already Labelled Data To 7 Emotion Classes	Stop Words, Remove Useless Punctuation, Replace Short Term, Stemming, Tokenization	BOW, POS Tagging, Length Of Sentence, Lexical Emotion Words	TF-IDF, MI	MNB	Accuracy	BOW Better At 79.73 %
Opinion Mining And Classification Of User Reviews In Social Media (2014)	User Reviews On Twitter Website	Twitter API	Manual Labelling Into Positive Or Negative Based On Knowledge	URL Removal			Noble NB	Accuracy	Accuracy At 85%
Classification Of Facebook News Feeds And Sentiment Analysis (2014)	Facebook Posts	Facebook Restfb Java APIs	Automatic Labelling Based On Manual Label For Classes Of Posts Sentiwordnet Dictionary For Sentiment Labelling	Stemming, POS Tagging, Tokenization			NB, SVM, J48, Binary Logistic, Bayes Net	F1, Precision, Recall, Accuracy	SVM Better At 97-99%
Predicting Reader's Emotion On Chinese Web News Articles (2013)	News Articles From Sina Society News		Already Labelled Using Voting System On The Site	Filtering, Sampling Strategies, HTML Tags Removal, Tokenization, Stop Words		TF-IDF, Chi-Squared	SVM, NB, Emotion Dictionary	F1, Precision, Recall	Classifiers With Emotion Dictionary Better In Accuracy Than Either Dictionary Alone Or MLT Alone.

LITERATURE REVIEW TABLES

Study	Dataset	Data Collection Tool	Label Method	Pre-Processing	Types Of Features	Feature Selection Method	Classification Technique	Evaluation Metrics	Comments
Text Mining Facebook Status Updates For Sentiment Analysis (2013)	Facebook Posts Posted By Tunisian Users		Manual Labelling To Positive And Negative	Stop Words,, Stemming	Combinations Of Unigram, Bigram, Trigram And POS	TF-IDF Presence	SVM, NB	Accuracy	SVM Better In General Especially Using Unigram At 72%
Sentiment Classification Of Malay Newspaper Using Immune Network (SCIN) (2013)	Malay News Papers			Stop Word, Stemming (Porter, Reverse Porter, And Backward-Forward Algorithms), Tokenization		Artificial Immune Network (AIN) , TF-IDF	K Nearest Neighbour	Accuracy	AIN With Reverse Porter Algorithm Is Better At 96%
Domain Keyword Extraction Technique: A New Weighting Method Based On Frequency Analysis (2013)	News Articles Extracted From The HTML Pages On Internet Portal Site		Already Labelled	Removal Stop Word, Word Stemming		TF-IDF with the addition of a weight called common word possession rate			The weight added to overcome the problem of TF-IDF in extracting the key features for each news domain
Extracting Similar And Opposite News Websites Based On Sentiment Analysis (2012)	News Articles		Sentiment Dictionary 3 Dimensional Model	POS Tag	Noun, Verbs, Adjectives	TF-IDF	Sentiment Dictionary	Accuracy	
Emotion Prediction Of News Articles From Reader's Perspective Based On Mutli-Classification (2012)	News Articles From Sina Society News		Already Labelled Using Voting System On The Site	Tokenization, Stop Words,		Chi-Squared, TF-IDF, POS Tagging	SMO, NB, MKNN, BR, RAKEL	Accuracy, HL, OE, COV, RL AVP	SMO+Chi Better For Single Label Classification RAKEL+ Chi ∩ DF Better For Multi Label Classification

LITERATURE REVIEW TABLES

Study	Dataset	Data Collection Tool	Label Method	Pre-Processing	Types Of Features	Feature Selection Method	Classification Technique	Evaluation Metrics	Comments
Emotion Tagging For Comments Of Online News By Meta Classification With Heterogeneous Information Sources (2012)	News Articles From Sina ,QQ News		Content Based Model CM Manually, User Emotion Tagging , Meta Classification Model MCM	Tokenization				Accuracy, MRR	CM Better Than UM MCM Better Than CM And UM
Opinion Extraction And Classification Of Real Time Alfacebook Status (2012)	Facebook Comments	Developed Tool Called Facebook Twitter Puller					Developed Classifier LIBSVM	Accuracy	Accuracy At 70%
Emotion Classification Of News Headlines Using SVM (2012)	News Headlines From (Google News, CNN, 10 News Papers		1000 Labelled Headlines From Semeval2007 Affective Task				SVM	Accuracy	Proposed System Better Than Systems In Semeval 2007
Use Of Porter Stemming Algorithm And SVM For Emotion Extraction From News Headlines (2011)	News Headlines From ISEAR Dataset		Already Labelled By 7 Emotions	Tokenization, Stemming, Pruning, Stop Words, Filtering	Emotion Words	TF-IDF, Wordnet Affect Dictionary	SVM, NB, VSM	F1, Precision, Recall, Accuracy	SVM + Stemming Porter Algorithm Is Better Than VSM, NB At 71% Accuracy

LITERATURE REVIEW TABLES

Study	Dataset	Data Collection Tool	Label Method	Pre-Processing	Types Of Features	Feature Selection Method	Classification Technique	Evaluation Metrics	Comments
Twitter Trending Topic Classification (2011)	Public Tweets	Twitter API	Manually Labelled Into 18 Categories	Stop Words, Tokenization		BOW TF-IDF	NBM, Decision Tree C5.0	Accuracy	Better Decision Tree At 70%
Monitoring Trends On Facebook (2011)	Facebook Public Posts	Facebook Graph API		Stop Words, URL Removal		TF-IDF	Clustering Algorithm By Distribution And Co-Occurrence	F1, Recall, Precision	
Sentiment Bias Detection In Support Of News Credibility Judgment (2011)	News Articles From 11 Japanese News Websites		Constructed Dictionary Of 4 Dimension Sentiments	HTML Removal, Post Tagging	Nouns, Verbs, Adjectives	TF-IDF	Sentiment Dictionary	Accuracy	Performance Is Good
Developing Robust Models For Favourability Analysis (2011)	3 Datasets From Different Media Analysis Company		Manual Labelling To Favourability And Non-Favourability With A Non-Negative Score		Unigram, Bigram, Trigram, Entity Word, Dependencies	Weka Tokenizer, Open NLP, Standford Dependency, Chi-Square	LIBSVM, KNN, RBF, J48, JRIP, Zeror+Random	Accuracy	NB Better For Sentiment Classification Various Classifiers Are Better For Sentiment Classification

CHAPTER 3

RESEARCH METHODOLOGY

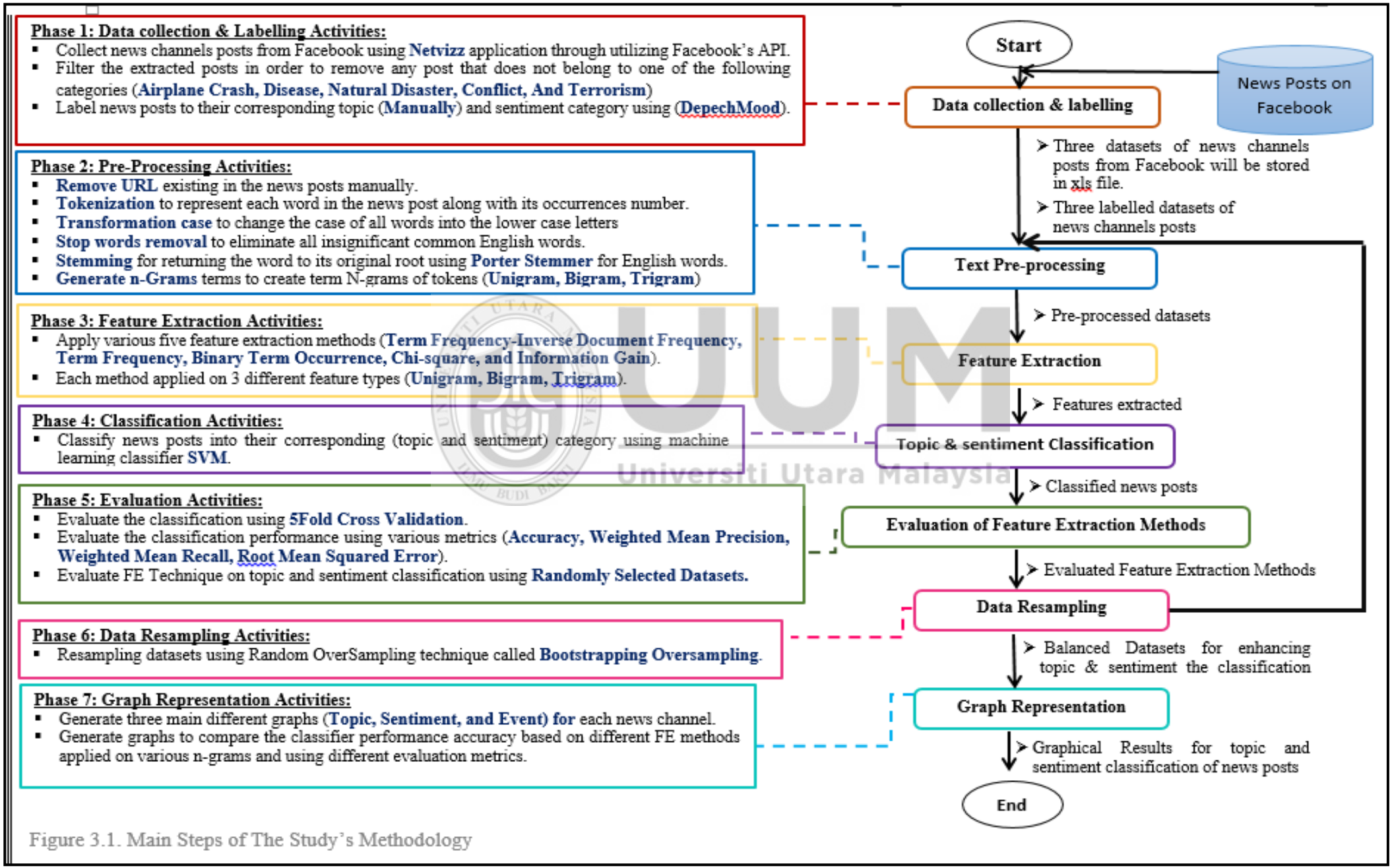

Figure 3.1. Main Steps of The Study's Methodology

3.1 Introduction

This chapter shows the proposed methodology which used to accomplish the objectives of this research. The proposed research methodology consists of seven main phases as shown in Figure 3.1.

3.2 Data Collection & Labelling Phase

News channels posts on Facebook have been used in this study as a trained data. Around 52,876 posts released in 2014 have been extracted from three different news channels, namely Aljazeera, BBC, and Al-Arabiya, more detailed about these three news channels can be found in Chapter 2, Section 2.4.2. Thus, in order to collect all these news posts, this study has used an application called Netvizz that utilizes the Facebook API for gathering purposes. Subsequently, this study has applied the third model supported by Netvizz which focuses on page contents, more specifically on posts released by the page owner only, which is in this case the news channel itself. For more information about this application refer to Chapter 2, Section 2.6.1.1. Consequently, a large number of news posts have been collected from January to December 2014 for each news channel. In general, there are four types of news posts which include links, stats, video and photo. All previous types has been used in this research.

As shown in the Table 3.1, the majority of the posts are of link type, followed by photo, video and status. All previous types have been used for this study. Each extracted post contains several details such as type of post, post message, link of post, domain link, post published date and time, number of comments and likes. All this information has been exported and saved in .xls format for the next phase. Prior

to pre-processing step, a filtering step has been performed on the extracted posts in order to remove any post that does not belong to one of the following categories (airplane crash, disease, natural disaster, conflict, and terrorism). Thus, the actual number of news posts used in this study is about 3,985, 3,872, and 1090 news posts from the three news channel Aljazeera, al-Arabiya, and BBC, respectively.

Table 3.1.
Statistical Analysis on Three News Channels

News channel	Number of Posts	Video Posts	Link Posts	Status Posts	Photos posts
Aljazeera	16,629	2,685	11,987	858	1,099
BBC	6,251	1,230	883	2	4,136
Al-Arabiya	12,978	0	12,308	149	521
Total	**35,858**	**3,915**	**25,178**	**1,009**	**5,756**

This study has focused on the following particular five categories (conflict, terrorism, airplane crash, disease, natural disaster) as these events are among the most frequent happening events during the last year 2014 around the world [1,2,3]. Also, these events were between the most talked-about global topics on Facebook [4,5] and twitter [6,7]. In addition, the reason behind focusing this study on these negative topics/events is, that these events are most influential to other news domains such as financial sector. In other words, analysing and knowing the sentiments associated with these events could help in getting knowledge about their impacts on the other markets like financial market. As well know, that stock prices responds to negative news more quickly than it would react to a positive news as well as good news locally could be dominated by the negative news across the globe.

Furthermore, the slightest bit of irritating negative news is enough to decline the prices of stocks. The opposite is also true. Moreover, news effects basically

depend on the associated sentiments rather than the actual significance of the news. Hence, this study has been implemented on these five categories with their associated sentiments in order to achieve the second sub-objective of this study which is, to classify the news posts on Facebook into their corresponding topic and sentiment categories.

Later, a topic labelling process has been conducted on the filtered posts, so each post will be categorized into its suitable category. This labelling step is done manually after reading the topic's definition and post's message contents. Similar, method has been done in previous studies [10,13,22]. On the other hand, to annotate a specific sentiment such as afraid, happy, angry, sad, amused, inspired, annoyed, and don't care to each news post for sentiment classification purpose, a free online demo for research propose calling DepecheMood created by Jacopo and Marco in [35] was utilized. This demo uses an emotion lexicon produced by the same authors, where the lexicon was generated based on the crowed annotated news articles collected from rappler.com. More detail about this lexicon can be found in Chapter 2, Section 2.6.2.1

Additionally, this study has chosen this specific free lexicon version among other free tools because this lexicon has built based on news crowed-source articles, which is quite similar to the datasets used for this study news post on Facebook. On the other hand, the other tools have built either on movie or product reviews that are not appropriate for the goal of this study, which is, to classify the news text into its corresponding emotion category.

3.3 Pre-Processing Phase

In order to enhance the process of detection and extraction features as well as to increase the performance of topic and sentiment classification for the news posts on Facebook, various pre-processing steps have been applied on all three datasets. The pre-processing activities for this study include several steps as following:

i. Remove manually the URLs that existed within the context of news posts.

ii. Tokenization that split the post's text into a sequence of tokens/words, where each token has represented along with its occurrence number in all documents.

iii. Transformation case which changed the case of all words in the posts into the lower case letters.

iv. Stop words removal that got rid of all insignificant common English words from the datasets by applying English stop words filter operator.

v. Stemming to return the word to its original root by removing the suffix or prefix parts. This process is executed using the Porter stemmer for English words. However, this stemmer algorithm sometimes produces not proper words [155] such as (police) which appeared as (polic) and etc. That's happened due to the conditions/action rules which the Porter algorithm consists of as well as it is necessary to note that any stemming algorithm cannot achieve perfection. The stemming algorithm purpose is not just to bring the word into its paradigm form, but to bring variant forms of a word together. This done by applying several conditions that fall into three classes: conditions on the stem, conditions on the suffix, and conditions on the rules [156].

In addition, the rules are distributed into steps, where the rules within one step are tested in sequence, and only one rule from a step can be implemented. Subsequently, the longest possible suffix is always removed because of the ordering of the rules within a step [156]. In any case, these few happening errors would not affect the classification accuracy so much, but it is matter of suggesting for future work that an additional rule could be merged in the stemmer algorithm to improve its performance [155].

vi. Generate different types of N-gram features such as (unigram, bigram, and trigram), in a document using generate N-Gram technique. N-gram is defined as a series of consecutive tokens of length n. Once the data has pre-processed, it is become ready for the next step which is a feature extraction phase.

3.4 Feature Extraction Phase

Feature extraction phase is an essential step that could lead to the determination of robust temporal features from a large number of news posts on Facebook. These features are essential to be recognized as it will contribute in identification of the correct topic and sentiment categories for the news post. Hence, this research has applied different feature extraction techniques such as TF-IDF, TF, BTO, Chi-Squared, and IG, more details about these techniques can be found in Chapter 2, Section 2.6.4. In addition, each feature extraction techinque has implemented on three n-grams (unigram, bigram, and trigram). These specific n-grams are selected due to their best performance in [130]. Furthermore, this study has extracted the robust temporal features for both topic/sentiment categories at the highest month's accuracy achieved for each news channel.

3.5 Classification Phase

After features extraction process is completed the classification phase has begun. The classification task for this study has involved both topic and sentiment classification of news posts into one of their appropriate categories. For topic classification, the news posts have been categorized into either conflict, natural disaster, disease, terrorism, or airplane crash. While for sentiment classification the posts have been classified into one of the eight categories such as sad, angry, happy, amused, inspired, annoyed, afraid, and don't know. Both classification processes have been done using the most famous machine learning classifier namely Support Vector Machine (SVM), which has been used by many previous studies and proved to be a better ML classifier compared to the others classifiers [8,15–19], for more details about this technique can refer to Chapter 2, Section 2.6.5.1. Furthermore, this study has used rapidminer's default settings for the SVM classifier.

3.6 Evaluation Phase

In order to evaluate the performance of both topic and sentiment classification for news posts, this study has used the identical validation technique for the SVM classifier, which is a 5-fold cross validation, whereby each dataset is randomly divided into 5 subsets (each subset has an equal number of examples). Then, five iterations will take place and each iteration involves a training model and a testing model. In addition, the evaluation for topic categorization has been done on original, resampled, unseen news posts. Unseen news posts have been selected randomly from the same dataset of the news channel. About 200 news posts have been selected randomly from each news channel at any time of the year 2014 in order to determine whether the classifier would be able to classify the new upcoming news

posts into their correct topic and sentiment categories and consequently achieve the third objective of this study. On the other hand, the evaluation for sentiment classification has done directly on resampled datasets as well as on unseen news posts.

Moreover, topic and sentiment classification have been evaluated by using various metrics, namely Accuracy (A), Weighted Mean Recall (WMR), Weighted Mean Precision (WMP), and Root Mean Squared Error (RMSE). Additionally, for further evaluation of the best feature extraction technique on unseen datasets for both topic categorization and sentiment classification addition metric has been used besides the previous ones called confidence score. Detailed information about these metrics can be found in Chapter 2, Section 2.6.6.

3.7 Data Resampling Phase

Usually the problem of imbalanced data is observed within text categorization and classification due to the overwhelmed of majority classes over minority classes. In other words, some classes may have a higher number of samples than other classes. This has led to a poor performance for the classifier on the minority class. Thus, this study has combined the optimal feature selection technique from the previous phase with a data over resampling technology called Bootstrapping OverSampling (BOS). BOS was implemented by creating a number of replicates of the minority class equal to the number of samples of the majority class. This has been done through the use of the RapidMiner bootstrapping operator. This operator generates a bootstrapped sample from the original dataset. Bootstrapped sampling utilizes sampling with replacement, hence the resampled dataset may not have all unique examples.

Furthermore, bootstrapping resampling algorithm works by separating each dataset into two sub-sets. One set for the majority class and the other for the minority class. Consequently, the minority class data are fed into the bootstrapping algorithm that leads to a multiple randomly resampled subsets which have the same volume as the volume of the original minority subset. Then, the resampled subset are combined with the majority class to build the resampled training data which is ready for training procedures. In addition, using this method was proved to be effective by many studies [31,124]. More detailed about oversampling bootstrapping technique can be found in Chapter 2, Section 2.7.1.

3.8 Graph Representation

Different types of graphs were generated to present the output results of this research which are presented in Chapter 4 and Chapter 5. The main three graphs for each news channel are topic, sentiment, and event associated with its emotions. Topic graph has illustrated the number of posts in each category (terrorism, airplane crash, natural disasters, diseases, and conflict). Topic graph for each news is presented in Chapter 4, Section 4.2. The sentiment graph on the other hand, has presented the number of posts in each category (sad, angry, happy, amused, inspired, annoyed, afraid, don't know). Meanwhile, Event graph has shown an event happened in the year 2014 and its associated sentiments. Both sentiment and event graphs for each news channel are presented in Chapter 5, Section 5.2. Other graphs also have been drawn in order to compare the classification performance based on accuracy metric through applying various FE techniques on the three n-grams terms. as can be seen in Chapter 4, Section 4.3 and Chapter 5, Section 5.3.

3.9 Chapter Summary

This chapter introduces the methodology of this study by which the problem mentioned in Chapter 1 can be solved and the objectives can be achieved. Then, a detail explanation about each phase included within the methodology is presented. Starting by the phases of data collection and labelling, pre-processing, feature extraction, classification, data resampling, evaluation, and ending by graph representation phase. Finally, a summary of the chapter is presented.

CHAPTER FOUR

RESULTS & DISCUSSION OF TOPIC CATEGORIZATION

4.1 Introduction

In this chapter, results and discussion for topic categorization of news posts on Facebook were represented for three news channel datasets namely Aljazeera, BBC and Al-Arabiya. Various feature extraction techniques have been used along with different types of features such as unigram, bigram, and trigram. The key idea of this research is to determine which feature extraction technique and type of n-gram feature that would lead to the robustness of the topic classification model. Furthermore, in order to enhance the topic classification performance as well as overcome the problem of imbalanced data, the realized best feature extraction technique has combined with an oversampling technology called Randomly OverSampling Bootstrapping technique. Moreover, well-known machine learning classifier was used in this research for the classification process named SVM.

4.2 Descriptive Analysis of Three News Channels

In this section, a descriptive analysis for all three news channels are presented as follows:

4.2.1 Al-Arabiya News Channel

According to Figure 4.1 the categories in this dataset are not well distributed throughout the year. There is a significant difference in term of number of posts between categories, whereby the number of news posts for conflict and terrorism categories are much higher compared to other categories (airplane crash, disease, and natural disaster) with the rate of 64%, 27%, respectively as it can be seen from the

Table 4.1. Furthermore, the total number of news posts for this channel is (3828) post and the highest number of posts is found in July where the largest number of posts belong to category airplane crash by (35) post. In contrary, other categories like disease and natural disaster contain only one post.

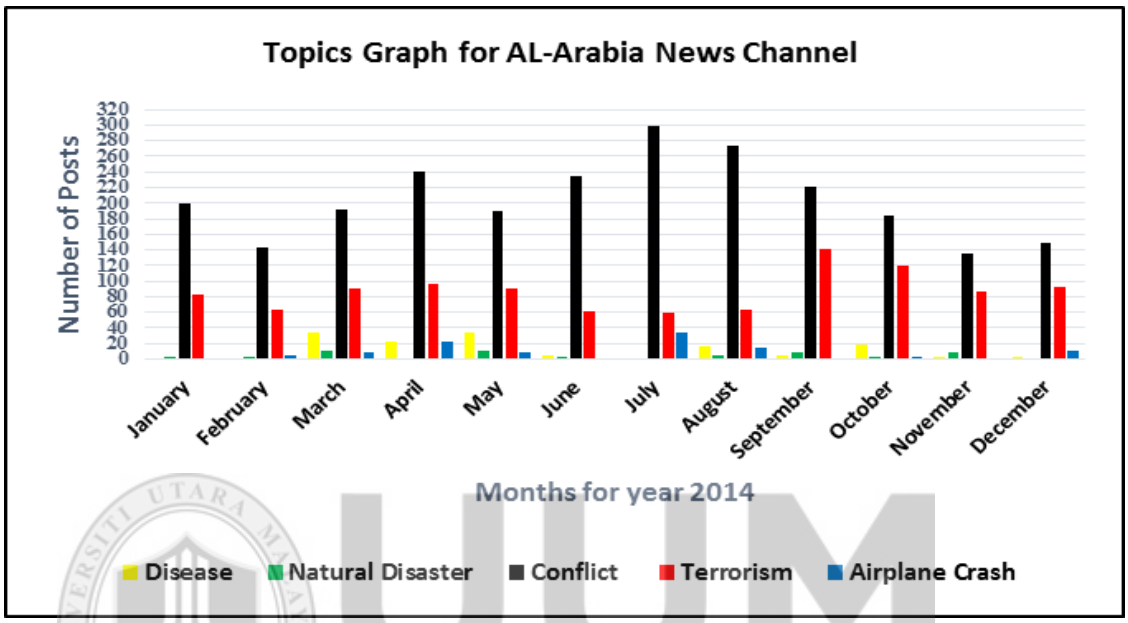

Figure 4.1.
Topic Graph for Al-Arabiya News Channel

Table 4.1
Number of Posts Per Topic Category for AL-ARABIYA News Channel

Month	Disease	Natural Disaster	Conflict	Terrorism	Airplane Crash	Total Per Month
January	1	4	199	82	2	288
February	2	4	143	63	5	217
March	34	11	191	90	8	334
April	23	2	241	96	23	385
May	34	11	190	90	8	333
June	5	4	234	62	2	307
July	**1**	**1**	**299**	**59**	**35**	**395**
August	16	5	273	63	15	372
September	5	8	221	142	2	378
October	18	3	184	120	3	328
November	4	8	135	86	0	233
December	4	2	149	92	11	258
Sum	**147**	**63**	**2459**	**1045**	**114**	**3828**
average	**4%**	**2%**	**64%**	**27%**	**3%**	

Table 4.1 shows no big difference in the number of news posts that have published in each month for the news channel and it is ranging between 200 to 400 posts per month. Furthermore, regarding to Figure 4.1 an increase is noticed in the number of posts for a disease category in the months (March, April, and May). This probably happened because the spread of the corona virus epidemic in the Middle East at that time. Similarly, same category has increased again in the period between August and October this perhaps due to the emergence of Ebola virus in West Africa. On the other hand, for the airplane crash an increasing number of posts is found in April due to the disappearance of the aircraft Malaysia MH370. Identical aviation accidents such as Malaysia MH17, Air Asia Q28501 and Algerian aircraft's crash have led to rise in number of posts once again in the three months named July, August, and December.

4.2.2 Al-Jazeera News Channel

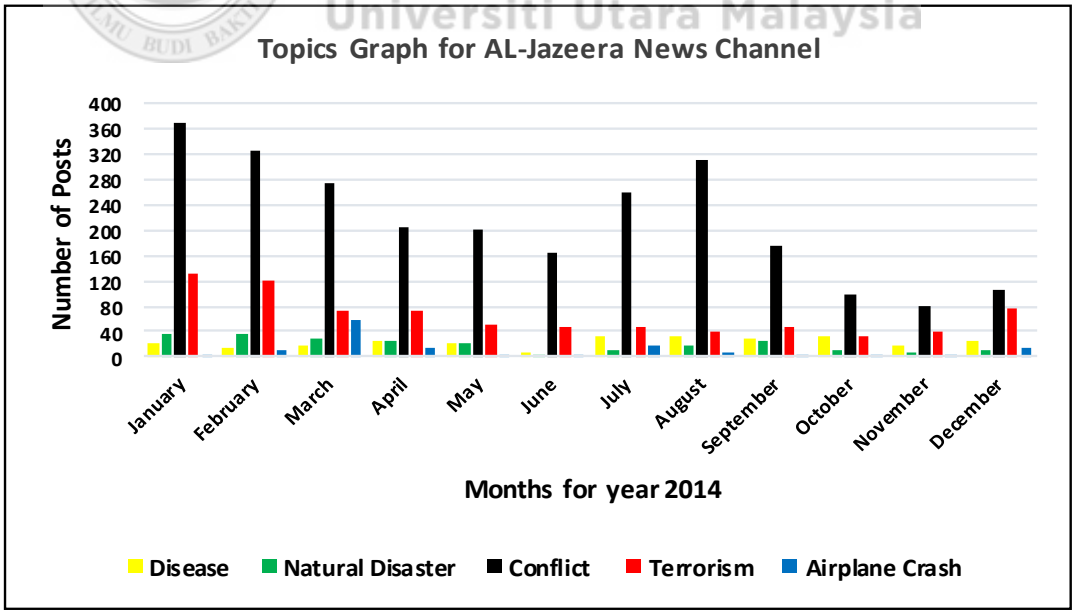

Figure 4.2.
Topic Graph for Al-Jazeera News Channel

Figure 4.2 indicates that conflict and terrorism are two events that mostly reported by this news channel same as Al-Arabiya channel, whereby the proportion of conflict represents a rate of 64%, similar to that of al-Arabiya channel. On the other hand, terrorism category has exceeded Al-Arabiya channel by 7%, followed by the other three categories (disease, natural disaster, and airplane crash) by a ratio of 7%, 6%, and 3%, respectively. In addition, it is also recognized that the highest number of posts has recorded for January and February which has reached 563 and 504 post, respectively. This is probably because these two months contain the largest number of posts exists in the categories such as natural disaster, conflict, and terrorism. In contrast, the lowest number of posts is identified in the month of November. Additionally, for airplane crash category a significant increase in the number of news posts has noticed in March due to the disappearance of the aircraft of Malaysia MH370.

Table 4.2.
Number of Posts Per Topic Category for AL-JAZEERA News Channel

Month	Disease	Natural Disaster	Conflict	Terrorism	Airplane Crash	Total Per Month
January	**21**	**37**	**371**	**131**	**3**	**563**
February	13	35	327	120	9	504
March	17	30	277	74	60	458
April	26	24	204	74	14	342
May	21	22	198	51	3	295
June	7	3	165	49	3	227
July	32	9	259	46	18	364
August	34	19	310	39	5	407
September	28	24	173	47	3	275
October	31	12	97	33	1	174
November	16	5	82	40	1	144
December	24	10	106	77	13	230
Sum	**270**	**230**	**2569**	**781**	**133**	**3983**
Average	**7%**	**6%**	**64%**	**20%**	**3%**	

Based on Table 4.2 there is no a significant difference in the distribution of the number of news posts for diseases and natural disasters classes during the year except for the month of June which has the minimum number of posts for the both categories. In addition, compared to al-Arabiya channel, Al-Jazeera channel has a higher number of posts for the natural disaster category for all months. This is probably because Al Jazeera is interested in covering natural disasters in various parts of the Earth more than Al Arabiya channel. Moreover, an increased number of news posts about the disease has been recorded for the period between July and December due to the occurrence of Ebola virus in West Africa's countries. While for the period between January and May the risen happened due to the emergence of corona virus in Middle East especially Saudi Arabia. Moreover, Aljazeera news channel has the same percentage as it is for al-Arabiya about the number of posts for the two categories conflict and airplane crash at 64% and 3%, correspondingly.

4.2.3 BBC News Channel

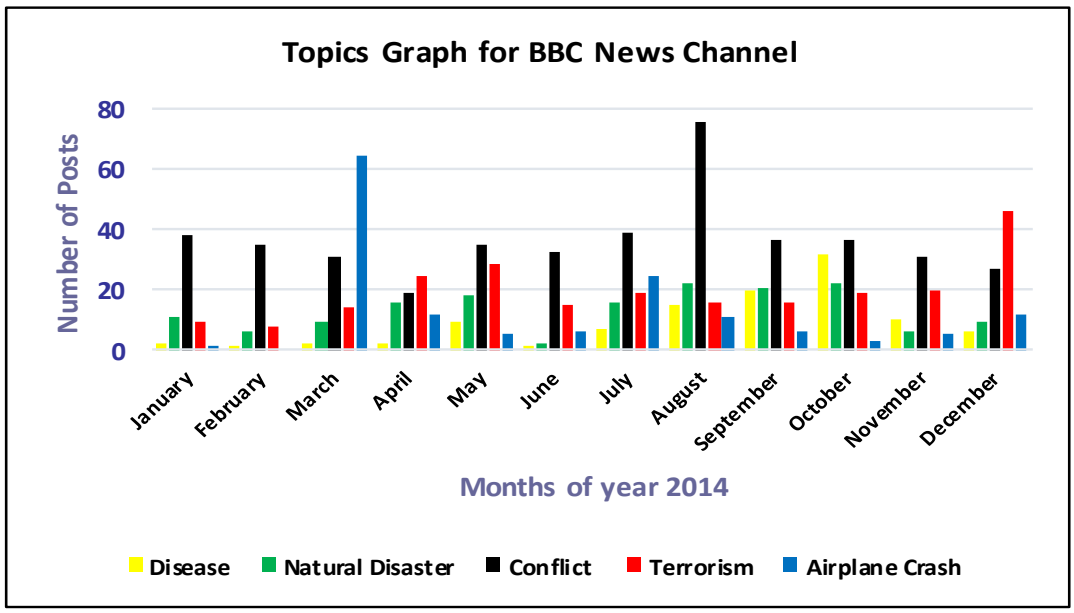

Figure 4.3.
Topic Graph for BBC News Channel

Figure 4.3 and with contradicting to previous two Figures for al-Arabiya and Aljazeera, it is noticed that there is no significant difference in term of number of posts between categories throughout the year except for the noticeable increase in conflict and airplane crash categories for the months March and August by 66 and 76 posts, respectively. In addition, BBC dataset is considered as the smallest dataset with only 1090 posts compared to the other datasets (Aljazeera, and l-Arabiya) which contain almost equal number of posts as it can be seen from Table 4.1 to Table 4.3. Furthermore, conflict and terrorism categories represent the highest percentages among other classes by 40% and 21%, respectively, followed by categories like airplane crash and natural disasters with the rate that is almost equal. Meanwhile, disease's class is considered to be the lowest category by only 10%.

Table 4.3.
Number of Posts Per Topic Category for BBC News Channel

Month	Disease	Natural Disaster	Conflict	Terrorism	Airplane Crash	Total Per Month
January	2	11	38	9	1	61
February	1	6	34	8	0	49
March	**2**	**9**	**31**	**12**	**66**	**120**
April	1	17	19	26	13	76
May	9	18	35	29	5	96
June	2	2	33	15	6	58
July	7	16	39	19	25	106
August	16	22	78	16	10	142
September	20	21	37	16	6	100
October	32	22	34	19	3	110
November	10	6	31	20	5	72
December	6	9	27	46	12	100
Sum	**108**	**159**	**436**	**235**	**152**	**1090**
Average	**10%**	**15%**	**40%**	**21%**	**14%**	

According to Table 4.3 the number of news posts published monthly, is close in number and ranging from 50 to 140 posts. Furthermore, similar to the other two news channels, categories like airplane crash and disease have a significant increase

in the number of news posts for some months due to the same events which were mentioned earlier in the Section 4.2.1.

4.3 Comparative Study of Feature Extraction Techniques (FET) for Topic Categorization

In order to achieve the first sub-objective of this study, which is to analyse FE techniques that could lead to the determination of temporal robust features for topic categorization from a large number of news posts. Various FE techniques and different n-gram features have been applied on the three news channel datasets namely, Al-Arabiya, Al-Jazeera, and BBC. The output results for each news channel is presented and discussed in the following sections.

4.3.1 Analyse Of (FET) on Three News Channels

4.3.1.1 Analyse of FET on Al-Arabiya News Channel

Based on the data of the Table A1 & Table A2 in Appendix A, the best accuracy has obtained in August at a rate of 86.83% for both IG and Chi-square which have used unigram as a feature set. However, this accuracy has a bit high RMSE of 0.604, while both WMR and WMP are considered to be quite good at 71% and 87%, respectively.

Furthermore, IG and Chi have identical accuracies for almost all months, which are represented as the highest accuracies achieved ranging between (69% - 86%) except for one month November in which TF has attained the top accuracy. Meanwhile, the other FE weighting techniques like TF-IDF, TF, and BTO have almost similar accuracies for the three n-gram types for all months, but the values are

considered low. Besides that, it is observed that their values for WMR and WMP are very low may reach 19% and 11%, respectively. Moreover, the lowest accuracy of 56% has accomplished for the month October using BTO as FE technique and a trigram as n-gram model.

In addition, it is remarked that the best accuracy for each month have achieved using unigram model as a feature set except for some months such as July and September in which bigram has accomplished higher accuracy while for December higher accuracy has achieved for both bigram and trigram features. According to the same tables, practically all five FE techniques have a high RMSE rate ranging from 0.5 to 0.8 except for July in which lowest value has recorded for Chi and IG using unigram feature set.

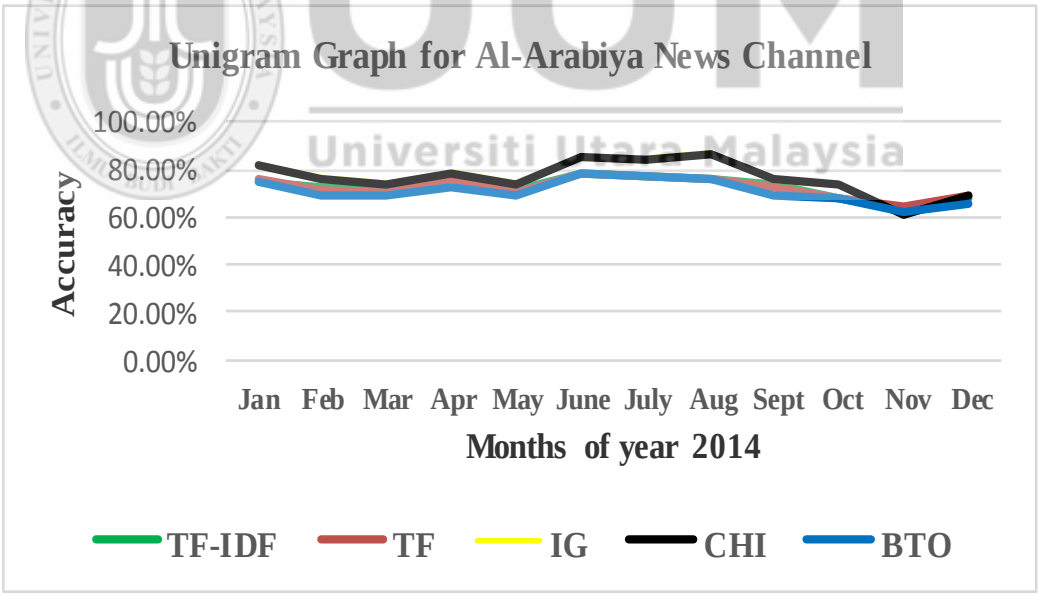

Figure 4.4.
Unigram Graph for Al-Arabiya News Channel

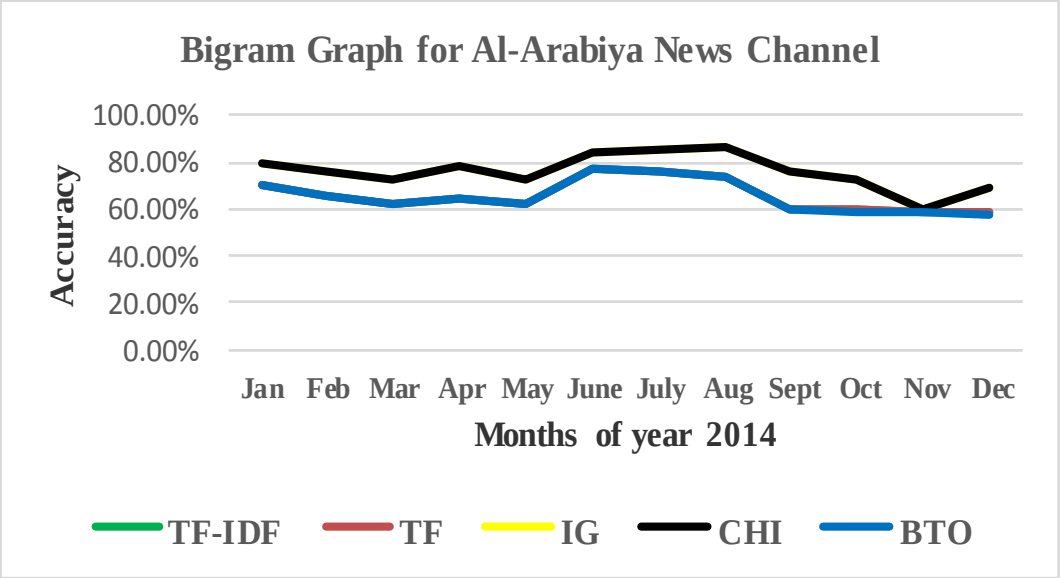

Figure 4.5.
Bigram Graph for Al-Arabiya News Channel

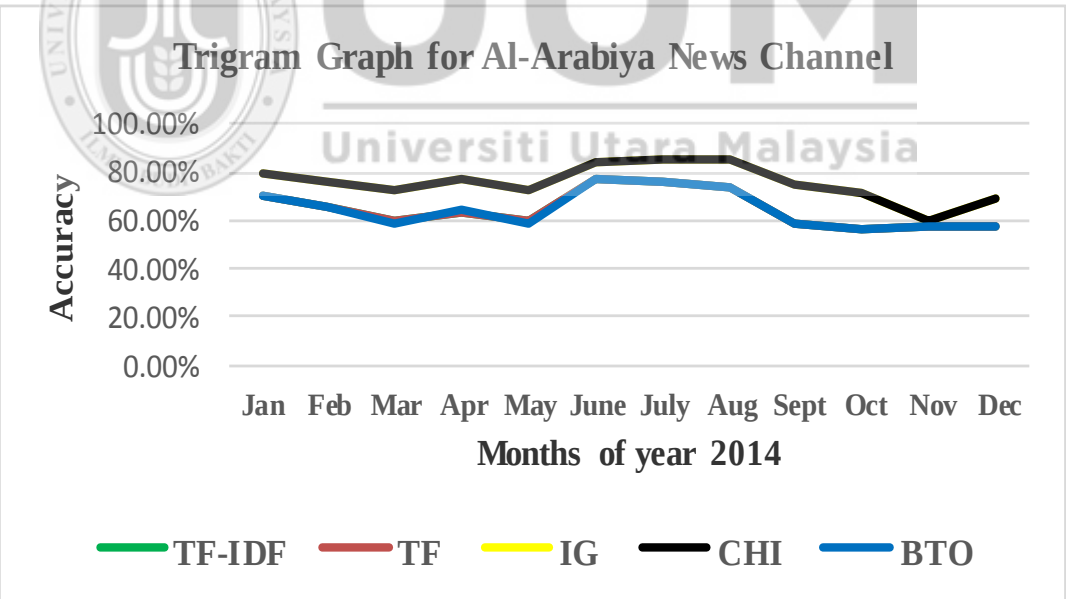

Figure 4.6.
Trigram Graph for Al-Arabiya News Channel

According to Figure 4.4, Figure 4.5, and Figure 4.6, it can be seen clearly that the accuracy's rates of Chi and IG are identical for three n-gram features. The same goes for the other three techniques (TF-IDF, TF, and BTO). However, a slight

increase was distinguished for TF-IDF's accuracy compared to other two techniques for some months, see Figure 4.4. In addition, Chi and IG have higher accuracies all over the year compared to the other techniques for the three n-gram types except for November month in which the accuracy has a significant decline and becomes almost equal with other techniques' accuracy as shown in Figure 4.5 and Figure 4.6. While for the unigram features it becomes lower compared to others accuracy, as can be seen from Figure 4.4. Moreover, categorization accuracy has reached its highest point for all three n-ngrams in August for Chi and IG techniques while for the other techniques it is obtained in June.

4.3.1.2 Analyse of FET on Al-Jazeera News Channel

Based on Table A3 & Table A4 in Appendix A, the highest accuracy has obtained in August same month as in al-Arabiya but with a higher rate by 1% at 87.96 %. This accuracy has acquired using IG and Chi-square techniques which have used unigram as a feature set. However, this accuracy has a little higher score for RMSE at 0.593, while both WMR and WMP values are quite good at 62% and 84%, respectively.

Furthermore, IG and Chi have achieved the highest accuracies ranging from 65% to 87% for almost all months except in October, whereby Chi has surpassed IG and accomplished a higher accuracy of 80% using unigram as feature type. On the other hand, FE techniques such as TF-IDF, TF, and BTO have almost comparable accuracies for three n-gram types for all months, but the values are considered to be low same as it was for al-Arabiya news channel. Additionally, their values for WMR and WMP are also very low compared to the values obtained by IG and Chi where the values have reached 20% and 11%, respectively.

Moreover, the lowest accuracy of 51% has attained using BTO as FE technique and a trigram as n-gram type for month December. In addition, it is noticed that the highest accuracy for almost all months have accomplished using unigram as a feature set except for some months such as February and June in which bigram has achieved higher accuracy. On the other hand, for months such as May, September, and November a higher accuracy has attained using trigram features. Furthermore, according to the same tables, almost all FE techniques have a high score of RMSE ranging from 0.5 to 0.8.

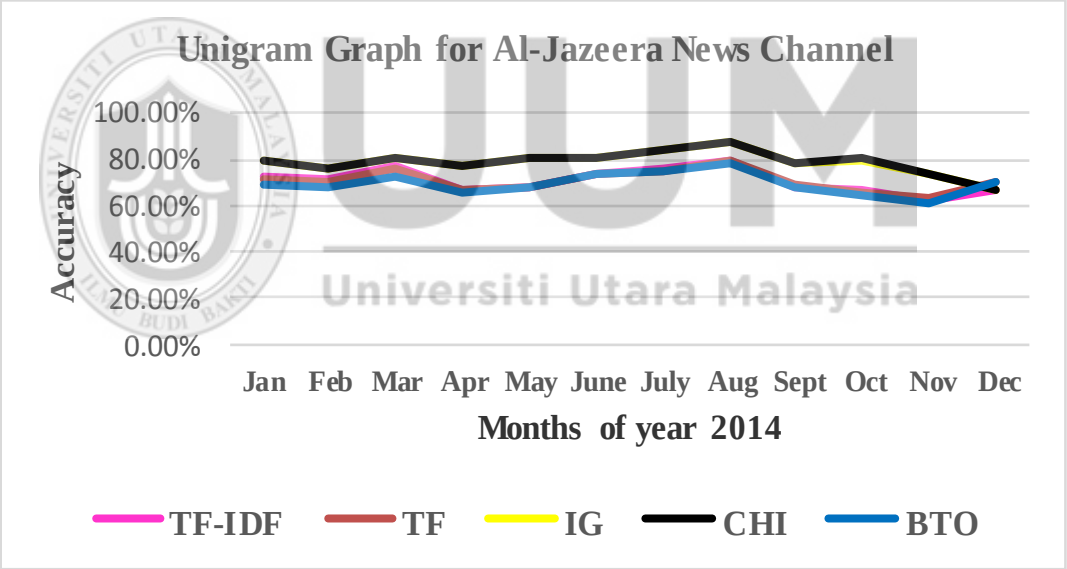

Figure 4.7.
Unigram Graph for Al-Jazeera News

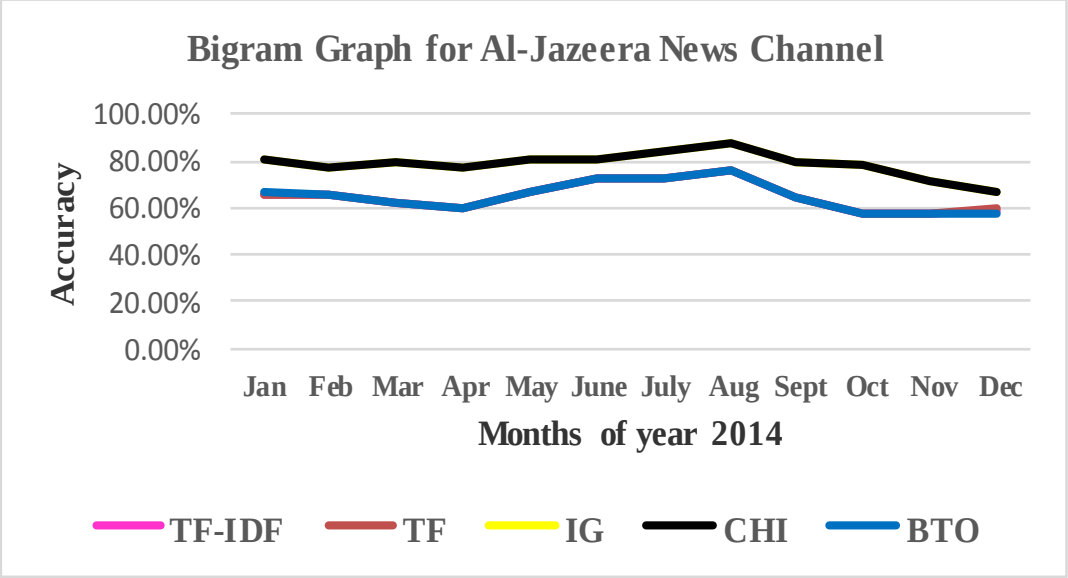

Figure 4.8.
Bigram Graph for Al-Jazeera News Channel

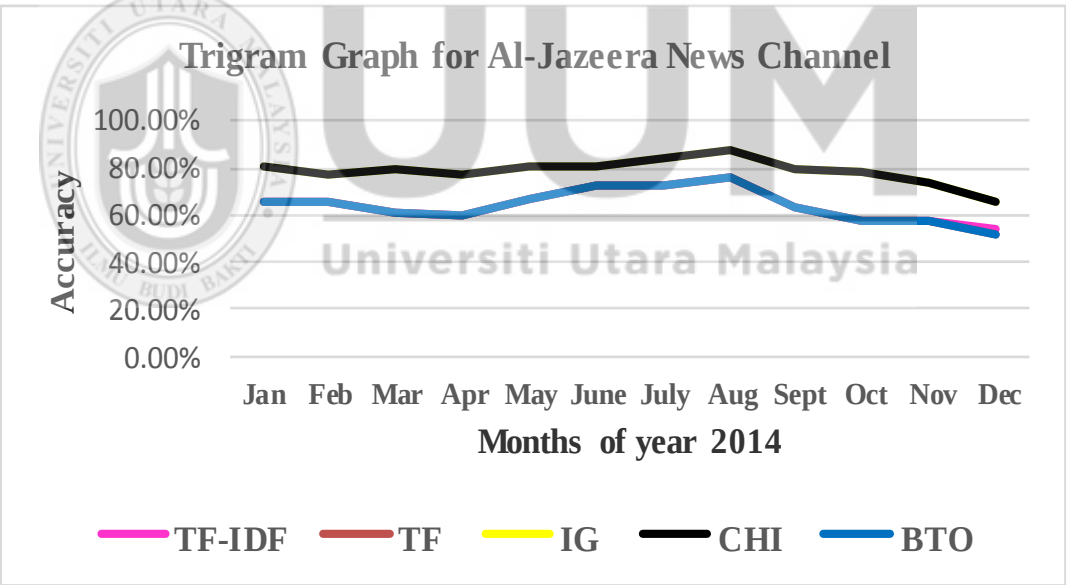

Figure 4.9.
Trigram Graph for Al-Jazeera News

Figure 4.7, Figure 4.8, and Figure 4.9 show that the accuracy's value for both Chi and IG are identical for all three n-gram features same as in al-Arabiya news channel. The same goes for the other three techniques (TF-IDF, TF, and BTO). However, as shown in Figure 4.7 for unigram type a minor increase is distinguished

for TF-IDF performance compared to other two techniques at the beginning of the year, while a slight decrease is noticed at the end. In addition, Chi and IG have higher accuracies all over the year compared to the other techniques for the three n-gram features. Furthermore, it can be seen from the same figures that a gradually increase in accuracy for all FE techniques has noted from the beginning of the year until it has arrived at the highest point in August and then begin to decline steadily until it reaches the lowest point in the month of November.

4.3.1.3 Analyse of FET on BBC News Channel

Table A5 in Appendix A indicate that the best accuracy has obtained in April at a rate of 88.89 %, which is higher than the best accuracies achieved for the previous two channels (Al-Arabiya and Al-Jazeera). This accuracy has acquired through using Chi-square technique and unigram as a feature set. However, this accuracy has a little high score of RMSE at 0.593, while both WMR and WMP values are quite good at 87% and 93%, respectively.

According to Table A5 & Table A6 in Appendix A, IG and Chi have same accuracies for almost all months, which are measured as the highest accuracies that have achieved ranging from 60% to 87%, excluding April month in which Chi has surpassed the IG's accuracy by 4%. However, for some months such as September, October, December, FE techniques like TF and BTO have obtained higher accuracies utilizing unigram as a feature features as shown in Table A2.

On the other hand, FE techniques such as TF-IDF, TF, and BTO have almost similar accuracies for all n-gram types for all months, but the values are lower compared to IG and Chi. Additionally, their values for WMR and WMP are correspondingly very

low same as for Aljazeera news channel where the values in some cases have reached 11% and 20%, respectively, Moreover, the lowest accuracy 46% had achieved in December using either TF-IDF or TF as FE techniques and trigram as n-gram features.

In addition, it is observed that the highest accuracy for each month has acquired using unigram as a feature type except for two months such as February and August, where bigram has obtained the highest accuracy at 79.17% and 74.78%, respectively. In contrast the lowest accuracy has obtained in May at ratio of 73.96% using trigram features. Furthermore, almost all five FE techniques have a high RMSE rate ranging from 0.5 to 0.8, similar to previous two news channel.

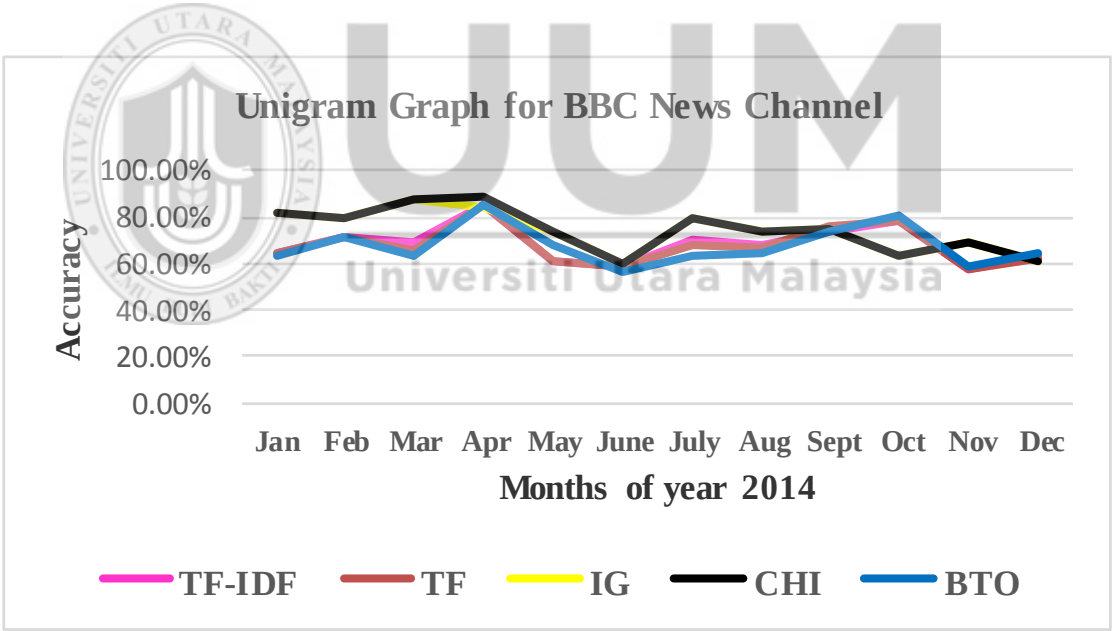

Figure 4.10.
Unigram Graph for BBC News Channel

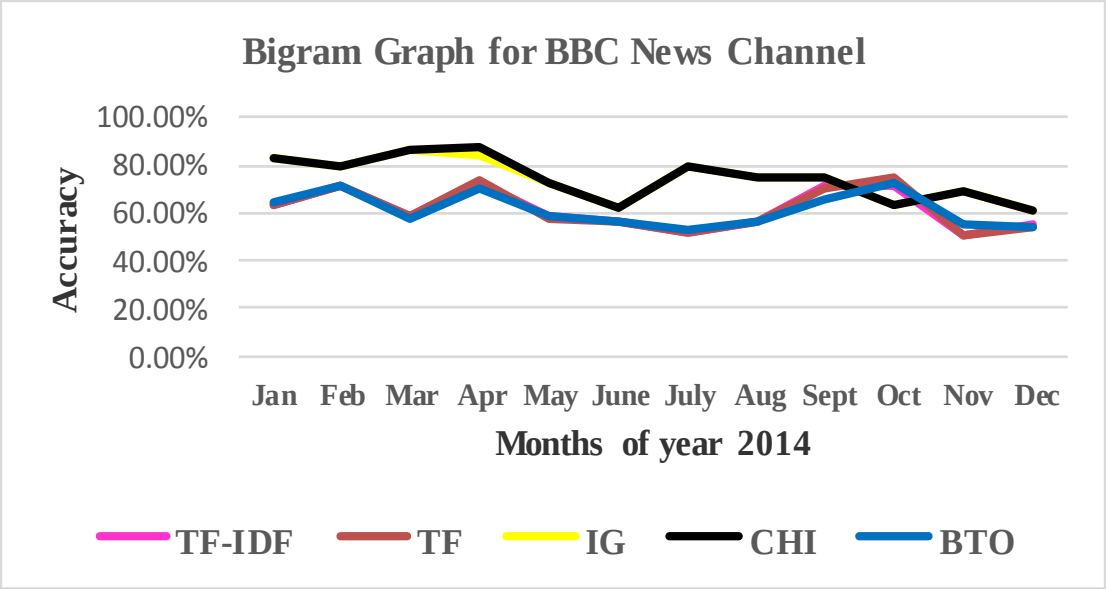

Figure 4.11.
Bigram Graph for BBC News Channel

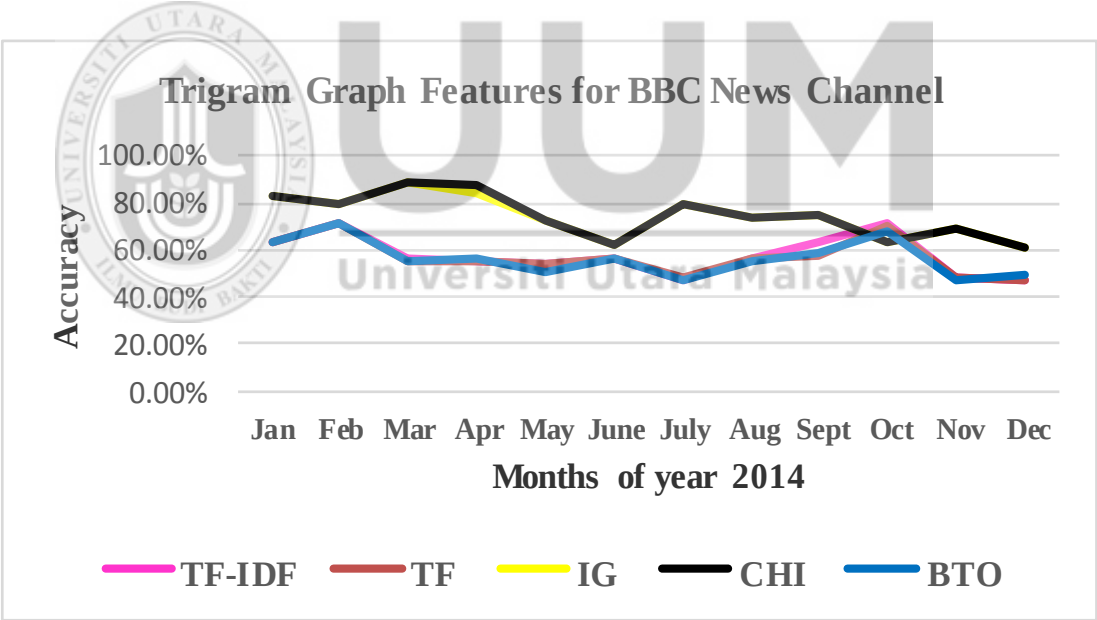

Figure 4.12.
Trigram Graph for BBC News Channel

According to Figure 4.10, Figure 4.11, and Figure 4.12 the accuracy for both Chi and IG are same for the three n-gram features, but a slight decrease in accuracy noticed for IG technique for all three n-gram types in April, oppositely Chi has achieved the highest accuracy for the same month. Also, a sharp decline in

performance has observed for the two techniques until reaching the lowest point in June. On the other hand, a slight difference in accuracy performance for the techniques such as TF-IDF, TF, and BTO has been noticed for both unigram and bigram features as can be seen from Figure 4.10 and Figure 4.11. Additionally, for all three n-gram features the accuracy of the former techniques has exceeded the accuracy of Chi and IG for the first time in month of October as shown from Figure 4.10 to Figure 4.12.

4.3.1.4 Overall Optimum Accuracy for three News Channels

To sum up, it has been concluded that for all three news channels, the highest accuracies have achieved using Chi- square or IG as feature extraction techniques with accuracy rates of 86% and 87% for al-Arabiya, Aljazeera respectively, while using only Chi-square for BBC. The optimal accuracy has been achieved for BBC dataset as there is no significant difference in the distribution of news posts over all categories. In addition, It is interesting to note that unigram have proved to be the most effective n-gram type for SVM classifier for obtaining a higher accuracy compared to other feature types as observed from Table 4.4 to Table 4.6. However, these high accuracies have pretty high values for RMSE nearly reach 0.6. In contrast, WMR and WMP have relatively good rates ranging from 62% to 93%, see Table A2 to Table A5 in Appendix A.

Based on Tables 4.4 till 4.6, both Chi and IG have identical accuracies for all months of the three news channels. Similarly, TF-IDF, TF, and BTO have almost same accuracies also, but their accuracies often lower compared to Chi and IG except for some months such as September, December for BBC news channel and November for Al-Arabiya channel where TF and BTO have surpassed Chi and IG

with a higher accuracy as it is shown from Table 4.4 and Table 4.6. Although TF-IDF has been reported in several studies as a prominent technique, it has not obtained highest accuracy in this experiment. This perhaps because the original TF-IDF is not suitable for extracting the features for single news domain [41]. These findings are aligned with other researchers and thus, a modified TF-IDF proposed by other studies have been applied.

Table 4.4
Optimum Accuracy for Al-Arabiya News Channel

Months	FE	N-Gram	Accuracy
JAN	**IG,CHI**	1	82.23%
FEB	**IG,CHI**	1	76.03%
MAR	**IG,CHI**	1	74.17%
APRIL	**IG,CHI**	1	78.57%
MAY	**IG,CHI**	1	74.17%
JUNE	**IG,CHI**	1	85.24%
JULY	**IG,CHI**	2	85.75%
AUG	**IG,CHI**	1	86.83%
SEPT	**IG,CHI**	2	76.46%
OCT	**IG,CHI**	1	73.48%
NOV	**TF**	1	64.24%
DEC	**IG,CHI**	1	69.26%

Table 4.5
Optimum Accuracy for Al-Jazeera News Channel

Months	FE	N-Gram	Accuracy
JAN	**IG,CHI**	2	80.08%
FEB	**IG,CHI**	2	77.18%
MAR	**IG,CHI**	1	80.99%
APRIL	**IG,CHI**	1	77.19%
MAY	**IG,CHI**	3	81.02%
JUNE	**IG,CHI**	2	80.63%
JULY	**IG,CHI**	1	84.62%
AUG	**IG,CHI**	1	87.96%
SEPT	**IG,CHI**	2	79.27%
OCT	**CHI**	1	80.34%
NOV	**IG,CHI**	3	73.95%
DEC	**IG,CHI**	1	67.77%

Table 4.6
Optimum Accuracy for BBC News Channel

Months	FE	N-Gram	Accuracy
JAN	**IG,CHI**	3	83.33%
FEB	**IG,CHI**	2	79.19%
MAR	**IG,CHI**	3	88.33%
APRIL	**CHI**	1	88.89%
MAY	**IG,CHI**	1	73.96%
JUNE	**IG,CHI**	2,3	62.11%
JULY	**IG,CHI**	1	79.05%
AUG	**IG,CHI**	2	74.78%
SEPT	**TF**	1	76.00%
OCT	**BTO**	1	80.96%
NOV	**IG,CHI**	1,2,3	69.44%
DEC	**BTO**	1	65.00%

Moreover, in some months such as April for BBC channel and October for Aljazeera channel, Chi-square has achieved the best accuracy compared to all other FE techniques as shown from Table 4.5 and Table 4.6. Furthermore, Table 4.7 shows the highest accuracy obtained for each news channel, whereby Chi and IG combined with unigram features have achieved the highest accuracy for two news datasets namely, al-Arabiya and Aljazeera but Chi outperforms IG for BBC dataset and achieved the highest accuracy at 88.89% compared to other two news channels datasets. Hence, this study has recognized Chi-square as the best FE technique which could lead to achieve a better performance for topic categorization.

Table 4.7
Highest Topic classification Accuracy achieved for Three News Channels On Original Datasets

News Channel Name	Type of Dataset	Feature Extraction Technique	N-gram	Accuracy
Al-Arabiya	Original	CHI,IG	1	86.83%
Al-Jazeera	Original	CHI,IG	1	87.96%
BBC	Original	CHI	1	88.89%

4.3.2 Analyse Of N-Gram Features Based On Chi-Square Technique for Resampled Datasets

The problem of imbalanced data usually occurs within text categorization due to the overwhelmed of majority classes over others and consequently lead to a poor performance for the classifier on the minority classes. Similarly, this study has encountered this problem while doing topic categorization process where some categories have a higher number of posts, than others, such as conflict and terrorism category as it can be seen from the Figure 4.1, Figure 4.2, and Figure 4.3 in Section 4.2. As a result, this study has combined the best feature selection technique Chi-square (which has realized from the previous experiment) with a data resampling technology called Random OverSampling (ROS), in order to overcome the problem of imbalanced data and improve the accuracy performance of SVM classifier for the topic categorization. The results of resampling has been introduced and discussed as below.

Table 4.7, Table 4.8, and Table 4.9 present the optimum accuracy obtained by giving the original and resampling datasets for each news channel (al-Arabiya, Aljazeera, and BBC) when Chi-square technique is employed over the three n-gram types (unigram, bigram, and trigram). In addition, it is clearly to see that after resampling the most effective n-gram features to achieve a higher accuracy is unigram feature for all datasets.

Table 4.8
Optimum Accuracy for Al-Arabiya News Channel Before & After Oversampling Using Chi-square

MONTHS	N-GRAM	BEFORE	N-GRAM	AFTER
January	1	82.23%	1	86.07%
February	1	76.03%	1	88.46%
March	1	74.17%	1	78.68%
April	1	78.57%	1	79.10%
May	1	74.17%	1	78.68%
June	1	85.24%	1	88.85%
July	2	85.75%	1	92.89%
August	1	86.83%	1	88.44%
September	2	76.46%	1	80.16%
October	1	73.48%	1	78.05%
November	1	61.21%	1	78.88%
December	2,3	69.26%	1	77.42%

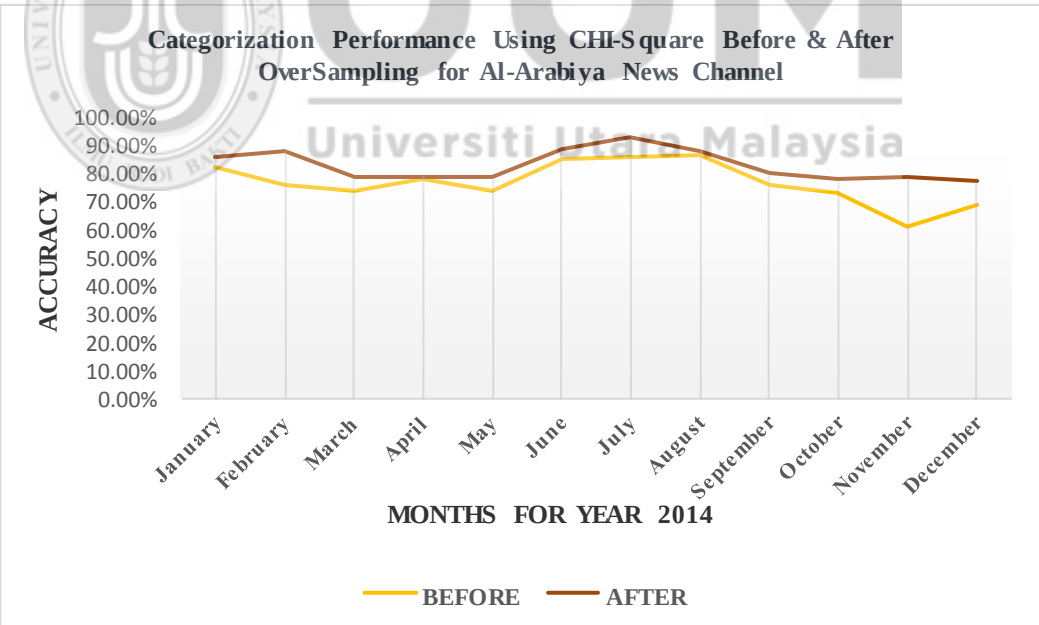

Figure 4.13.
Categorization Performance Using Chi-Square Before & After Oversampling for Al-Arabia News Channel

Table 4.9
Optimum Accuracy for Al-Jazeera News Channel Before & After Oversampling Using Chi-square

MONTHS	N-GRAM	BEFORE	N-GRAM	AFTER
January	2	80.08%	1	83.63%
February	2	77.18%	1	81.94%
March	1	80.99%	1	87.34%
April	1	77.19%	1	83.33%
May	3	81.02%	1	91.19%
June	2	80.63%	1	85.01%
July	1	84.62%	1	90.10%
August	1	87.96%	1	91.92%
September	2	79.27%	1	82.17%
October	1	80.34%	1	82.63%
November	3	73.95%	1	83.14%
December	1	66.77%	1	73.79%

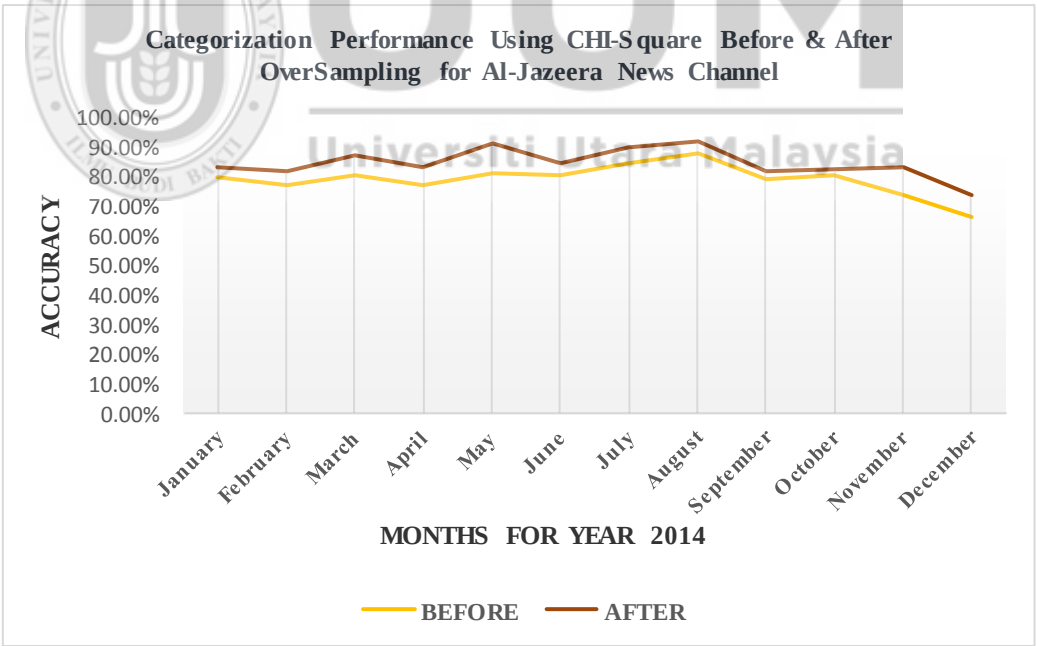

Figure 4.14.
Categorization Performance Using Chi-Square Before & After Oversampling for Al-Jazeera News Channel

Table 4.10
Optimum Accuracy for BBC News Channel Before & After Oversampling Using Chi-square

MONTHS	N-GRAM	BEFORE	N-GRAM	AFTER
January	3	83.33%	1	85.00%
February	2	79.17%	1	89.00%
March	3	88.33%	1	93.37%
April	1	88.89%	1	89.07%
May	1	73.96%	1	81.25%
June	2,3	62.11%	1	77.46%
July	1	79.05%	1	80.95%
August	2	74.78%	1	78.42%
September	1	74.96%	1	89.99%
October	1,2	63.59%	1	78.18%
November	1,2,3	69.44%	1	73.71%
December	1	60.99%	1	80.01%

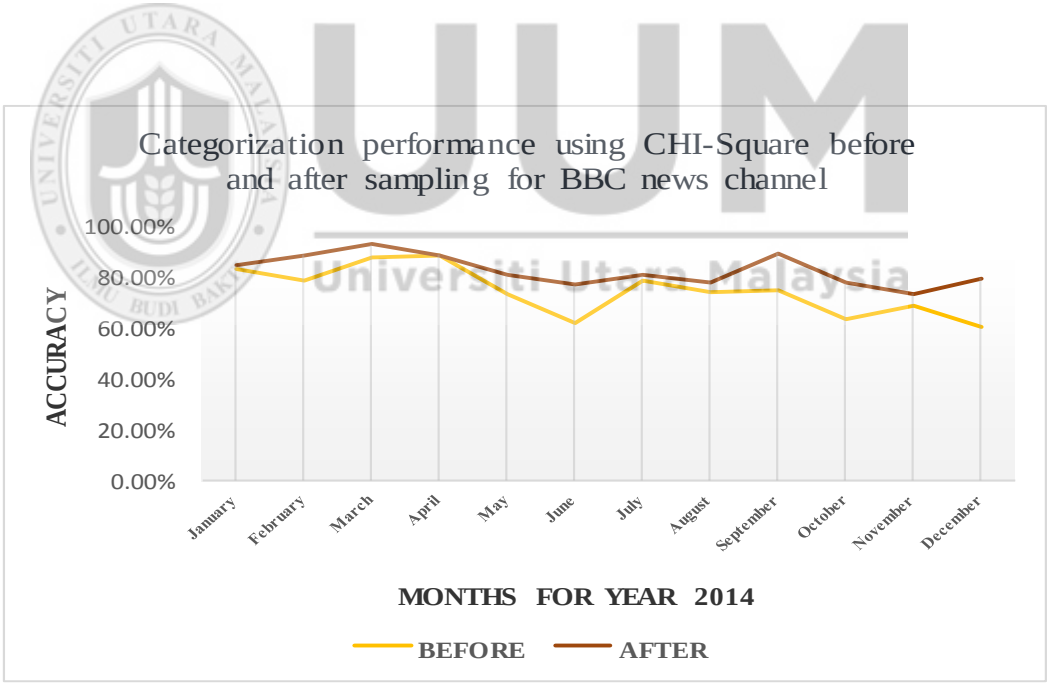

Figure 4.15.
Categorization Performance Using Chi-Square Before & After Oversampling for BBC News Channel

On the other hand, Figure 4.13, Figure 4.14, and Figure 15 plot all values of accuracy from these tables to provide a general view of the results. It is clear to see that the performance of classifier SVM has improved given resampling dataset for all

months for each news channel's dataset considered in this experiment, practically with respect to accuracy values. The comparison results confirm that performing resampling technology using oversampling technique (bootstrapping) can overcome the problem of imbalanced data and enhance minority class categorization as shown from Table B1, Table B2, and Table B3 in Appendix B.

Furthermore, according to Figure 4.13, Figure 4.14, and Figure 4.15 it can be seen that there is an enhancement in overall performance for each month of the year, but a significant improvement is noted for specific months for each dataset such as (November) for Al-Arabia, (June, September, October, December) for BBC while other months' accuracy has improved just slightly. This is probably because these months have unwell distribution of news posts over the five topic categories compared to the other months as can be seen from Table 4.1, and Table 4.3 in Section 4.2.1. Hence the idea of using oversampling technology comes to light as its impact can be more heavily on these kind of months as its primary advantage is to raise the performance of minority classes.

Additionally, and according to Table 4.11 the highest optimal accuracy for Al-Jazeera, Al-Arabiya, and BBC news channels have been achieved at 91.92%, 92.89%, 93.37% based on resampling datasets respectively. These accuracies have been increased by 4%, 6%, and 5% correspondingly, over that obtained for original datasets. Moreover, it is confirmed again that the combination of Chi-square and unigram features has led to achieve the highest optimum accuracies for all three news channels. Thus, this study has identified Chi-square technique with unigram features as best feature extraction technique that could lead to the determination of temporal

robust topic features from large number of news posts on Facebook, and hence achieve the first sub-objective of this study for topic categorization.

Table 4.11
Highest Topic classification Accuracy achieved for Three News Channels On Resampled Datasets Using Chi-square + Unigram

News Channel Name	Type of Dataset	Feature Extraction Technique	N-gram	Month	Accuracy
Al-Arabiya	Resampled	CHI	1	July	92.89%
Al-Jazeera	Resampled	CHI	1	August	91.92%
BBC	Resampled	CHI	1	March	93.37%

4.3.3 Determine Of Twenty Robust Features for Topic Categorization

Based on the highest optimal accuracies shown in Table 4.11, the top twenty robust features for topic categorization have been extracted from the three news channel datasets. The top twenty features for Al-Arabiya, Aljazeera and BBC are unigram features which have been extracted, ranked, selected using Chi-square technique as can be seen from Tables 4.12 till 4.14.

Based on Table 4.12, Table 4.13, and Table 4.14, it can be seen that some features are appearing in all three datasets such as kidnap, but having a diverse score based on the dataset they belong to. On the other hand, other features could be found in one dataset, but not in other like Ebola, virus, earthquake which can be found just in both Al-Jazeera and BBC but not in Al-Arabiya. This is perhaps because the month from which these features are extracted contains very few posts belong to disaster and disease categories as it is shown from Table 4.1 in Section 2.1. Hence, the top extracted features for each news channel depend on the events emerged in the relevant month. Furthermore, the highlighted values in the tables indicate the

maximum weight for these features since any feature can belong to more than one category. Thus, the highest weight could determine to which category the feature is fit. For example, in Table 4.14 feature (kill) has a predominant weight in terrorism (0.23), while (plane) has a predominant weight in airplane crash (0.54) compared to other categories.

Table 4.12
Top (20) Robust Temporal Features for Topic Categorization (Al-Arabiya News Channel)

Word	Conflict	Terrorism	Airplane Crash	Diseases	Natural Disaster
kidnap	0.00	0.00	0.48	0.00	0.00
lost	0.00	0.00	0.39	0.00	0.00
malaysia	0.00	0.00	0.37	0.00	0.00
plane	0.00	0.00	0.36	0.00	0.00
abduct	0.33	0.00	0.35	0.00	0.00
explose	0.00	0.32	0.00	0.00	0.00
aircraft	0.27	0.29	0.00	0.00	0.00
passenger	0.00	0.00	0.36	0.00	0.00
airline	0.27	0.00	0.22	0.00	0.00
blast	0.22	0.27	0.19	0.00	0.00
boko	0.21	0.27	0.00	0.00	0.00
carrier	0.13	0.23	0.27	0.00	0.00
victim	0.00	0.00	0.26	0.00	0.00
held	0.00	0.25	0.00	0.00	0.00
jet	0.00	0.25	0.00	0.00	0.00
below	0.22	0.00	0.24	0.00	0.00
build	0.21	0.24	0.23	0.00	0.00
car	0.16	0.24	0.00	0.00	0.00
wreckag	0.00	0.00	0.24	0.00	0.00
captive	0.19	0.00	0.00	0.00	0.00

Table 4.13

Top (20) Robust Temporal Features for Topic Categorization (Al-Jazeera News Channel)

Word	Conflict	Terrorism	Airplane Crash	Disease	Natural Disaster
Plane	0.39	0.00	0.54	0.00	0.00
Curfew	0.26	0.00	0.00	0.44	0.00
Infect	0.00	0.00	0.00	0.39	0.00
Rise	0.21	0.00	0.00	0.37	0.00
freed	0.00	0.37	0.00	0.00	0.22
earthquake	0.00	0.00	0.00	0.00	0.37
flood	0.23	0.00	0.00	0.00	0.35
crash	0.00	0.00	0.35	0.00	0.00
declare	0.00	0.00	0.00	0.35	0.26
death	0.31	0.00	0.00	0.34	0.25
kidnap	0.18	0.34	0.00	0.00	0.00
passenger	0.00	0.00	0.34	0.00	0.00
ebola	0.00	0.00	0.00	0.34	0.00
virus	0.00	0.00	0.00	0.34	0.00
dead	0.26	0.00	0.00	0.33	0.24
strong	0.00	0.29	0.00	0.00	0.33
malaysia	0.18	0.00	0.32	0.00	0.00
injur	0.312	0.00	0.00	0.00	0.31
landslide	0.00	0.00	0.00	0.00	0.31
vacate	0.00	0.00	0.00	0.00	0.31

Table 4.14

Top (20) Robust Temporal Features for Topic Categorization (BBC News Channel)

Word	Conflict	Terrorism	Airplane Crash	Disease	Natural Disaster
virus	0.00	0.00	0.00	0.37	0.00
attack	0.13	0.34	0.00	0.00	0.00
ebola	0.00	0.00	0.00	0.34	0.00
crisi	0.32	0.00	0.00	0.00	0.00
debri	0.00	0.00	0.29	0.00	0.00
dry	0.00	0.00	0.00	0.00	0.27
die	0.00	0.22	0.00	0.00	0.26
hit	0.00	0.26	0.00	0.00	0.23
kidnap	0.00	0.26	0.00	0.00	0.00
mudslide	0.00	0.00	0.00	0.00	0.25
flight	0.00	0.00	0.23	0.00	0.00
kill	0.08	0.23	0.00	0.11	0.00
protest	0.23	0.00	0.00	0.00	0.00
rebel	0.22	0.23	0.00	0.00	0.00
accident	0.00	0.00	0.22	0.00	0.00
earthquak	0.00	0.00	0.00	0.00	0.22
miss	0.00	0.00	0.22	0.00	0.00
activist	0.19	0.00	0.00	0.00	0.00
search	0.00	0.00	0.19	0.00	0.04
victim	0.00	0.00	0.00	0.16	0.19
suspect	0.00	0.10	0.18	0.16	0.00

4.4 News Post Classification on Topic Categorization Using SVM

In order to achieve the second sub-objective of this study for topic categorization which is to classify the news post into its corresponding topic categories for the social events, famous machine learning technique called SVM has been applied to classify the news post to its appropriate topic category. Tables 4.15 till 4.17 show the output results of performing the topic classification process for the three news channels on resampled datasets using Chi-square as feature extraction technique and unigram as n-gram features and which subsequently leads to achieve the highest topic classification accuracy.

In addition, it can be seen from the same Tables that most news posts been correctly predicted to their categories and this consequently led to obtain high proportion of precision for all five categories where the precision values were ranging between 90 and 100%. Similarly, the recall rates are also quite good ranging between 60 and 100%. In addition, it is noticed that only three categories (conflict, terrorism, airplane crash) were presented in the classification process for Al-Arabiya dataset that is because for that month there was no any news post released for the other two categories (disease, natural disaster). On the other hand, all five topic categories have presented for Al-Jazeera and BBC datasets.

Moreover, Furthermore, it is concluded that the obtained classification accuracies are in order of 93.37%, 92.89%, 91.92%, for BBC, Al-Arabiya, Al-Jazeera news channels. The highest accuracy belong to BBC dataset as there is no significant difference in the distribution of news posts over the five categories involved in the process. Hence, the value of accuracy depend basically on the number on the distribution of news posts over these categories.

Table 4.15

News Posts Classified On Topic Categorization Using SVM (Al-Arabiya News Channel)

ACCURACY: 92.89%

	True Conflict	True Terrorism	True Airplane Crash	Class Precision
Pred. Conflict	302	19	9	91.52%
Pred. Terrorism	0	45	0	100.00%
Pred. Airplane Crash	0	0	19	100.00%
Class Recall	100.00%	70.31%	67.86%	

Table 4.16

News Posts Classified On Topic Categorization Using SVM (Al-Jazeera News Channel)

ACCURACY: 91.92%

	True Conflict	True Terrorism	True Natural Disaster	True Disease	True Airplane Crash	Class Precision
Pred. Conflict	308	16	5	10	1	**90.59%**
Pred. Terrorism	0	27	0	0	1	**96.43%**
Pred. Natural Disaster	0	0	13	0	0	**100.00%**
Pred. Disease	0	0	0	20	0	**100.00%**
Pred. Airplane Crash	0	0	0	0	8	**100.00%**
Class Recall	**100.00%**	**62.79%**	**72.22%**	**66.67%**	**80.00%**	

Table 4.17
News Posts Classified On Topic Categorization Using SVM (BBC News Channel)

ACCURACY: 93.37%

	True Conflict	True Airplane Crash	True Disease	True Terrorism	True Natural Disaster	Class Precision
Pred. Conflict	17	0	0	0	0	**100.00%**
Pred. Airplane Crash	3	81	0	3	2	**91.01%**
Pred. Disease	0	0	2	0	0	**100.00%**
Pred. Terrorism	0	0	0	8	0	**100.00%**
Pred. Natural Disaster	0	0	0	0	5	**100.00%**
Class Recall	**85.00%**	**100.00%**	**100.00%**	**72.73%**	**71.43%**	

4.5 Evaluation of Chi+Unigram on Topic Categorization Using Random Selected Dataset

In order to achieve the third sub-objective of this study for topic categorization which is to evaluate the performance of the best feature extraction technique that has been realized (Chi-square+Unigram) on randomly selected news posts. The evaluation process has been done using the same validation technique that used for original datasets and resampled datasets. The validation technique is an automatic 5-fold cross validation for the SVM classifier. Each evaluating process for each news channel has been done using various performance metrics such as Accuracy (A), Weighted Mean Recall (WMR), Weighted Mean Precision (WMP), Root Mean Squared Error (RMSE), and confidence score.

About two hundreds (200) news posts released at any time in the year (2014) are selected randomly from each news channel's dataset namely, al-Arabiya, Aljazeera, BBC in order to build three individual testing datasets. Then, the

optimization topic categorization models would run on these unseen dataset. The optimization model is which has been realized for each news channel in months such as July, August, March for al-Arabiya, Aljazeera, and BBC, respectively and has achieved an optimal performance accuracy for topic categorization on resampled datasets. Each model uses Chi-square as FE technique combine with unigram features and utilizes SVM as a machine learning classifier.

Table 4.18
Evaluation Metrics of Topic Categorization For Unseen Data

Channel	Accuracy	WMR	WMP	RMSE
Al-Arabiya	**70.00%**	27.73%	46.54%	**0.644**
Al-Jazeera	67.00%	34.87%	64.58%	0.662
BBC	46.00%	**42.66%**	**74.20%**	0.728

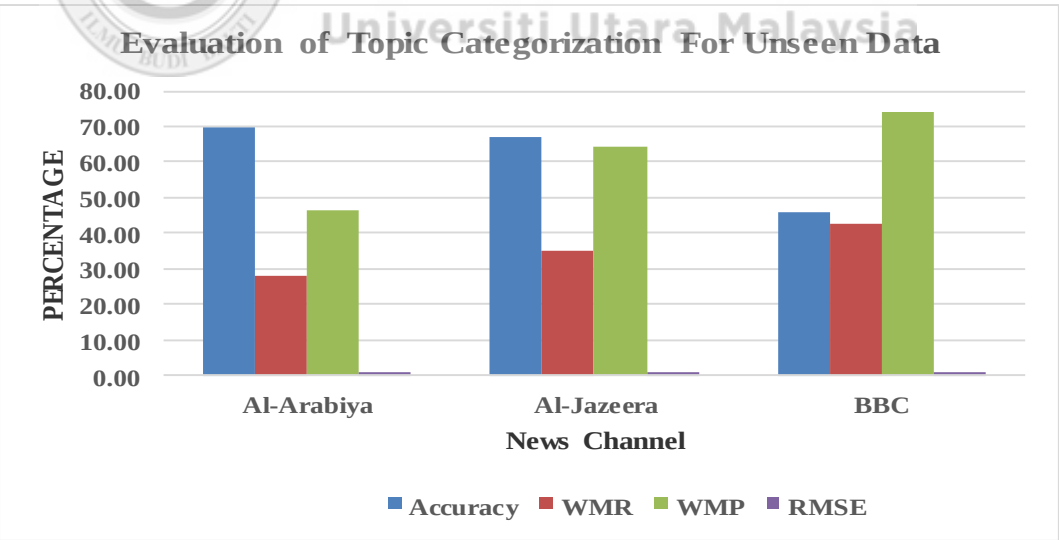

Figure 4.16.
Randomly selected Data Evaluation for Three News Channels

Table 4.18 shows the results of topic categorization for unseen data for each news channel based on four evaluation metrics (accuracy, WMR, WMP, RMSE). The highlighted values indicate the maximum value for each matrices. Figure 4.16 presents a comparison between all these metrics, where it can be seen that the highest accuracy belong to al-Arabiya news channel at 70%, though it has the worst values for other metrics such as WMR, WMP which is in contrary, quite good for BBC news channel but has the lowest accuracy rate and high score of RMSE. On the other hand, the metrics' rate for Aljazeera is reasonable especially for the accuracy and WMP.

Furthermore, it is observed that these accuracies are quite lower compared to that achieved by the cross validation on either original or resampled datasets. This probably due to a change of topics over time, whereby the training dataset of each news channel may not contain features that can contribute in detecting the events emerged in the randomly selected news posts. Thus, could not classify the news posts to their topic category correctly and consequently decrease the confidence scores as shown in Table 4.19.

Table 4.19
Confidence Score of Each Category Prediction for Each News Channel

Channel	Conflict	Terrorism	Airplane Crash	Disease	Natural Disaster
Al-Arabiya	0.419	0.149	0.151	0.138	0.143
Al-Jazeera	0.414	0.152	0.138	0.154	0.142
BBC	0.194	0.206	0.285	0.158	0.157

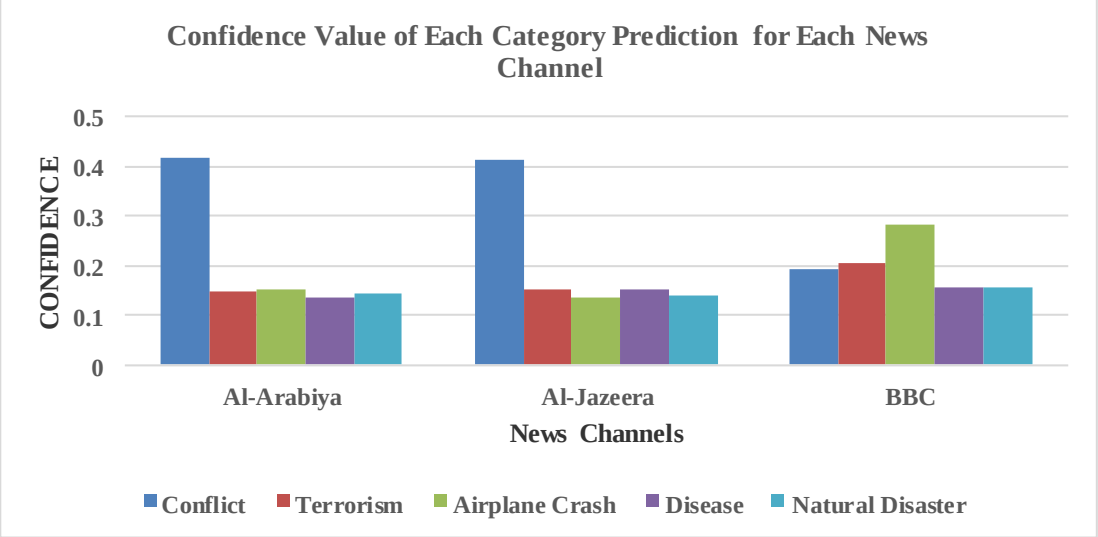

Figure 4.17.
Confidence Value of Each Category Prediction for Each News Channel

Table 4.19 presents the confidences score obtained for the prediction of each topic category for each news channel. Based on Figure 4.17, it's clear that almost equally high confidence scores of conflict class have attained for both al-Arabiya and Aljazeera news channels at 0.419 and 0.414, respectively. In addition, there is no big difference between confidence scores for the other four categories of all three news channels except for airplane crash, where BBC has the highest score at 0.285. This happened probably because of the large number of posts exist in its training dataset of month March, which contains about 66 posts for airplane crash. Hence, the training dataset has more features of that category which has contributed in predicting the unseen posts that belong to the same category and subsequently raise the confidence score.

4.6 Chapter Summary

In this chapter, topic categorization experimental results and discussion were presented. Started with a descriptive analysis for each news channel's dataset

followed by introducing and discussing the analysing of feature extraction techniques for topic categorization. Then, analysis of n-gram features based on Chi-square technique for resampled datasets and the determination of robust temporal features for topic categorization. Finally, show the results of evaluating the best realized feature extraction technique of topic categorization using randomly selected dataset from each news channel. In next chapter, the results of sentiment classification experiments and discussion will be presented.

CHAPTER FIVE

RESULTS & DISCUSSION OF SENTIMENT CLASSIFICATION

5.1 Introduction

In this chapter, results and discussion for sentiment classification of news posts on Facebook are presented for each news channel datasets namely Aljazeera, BBC and Al-Arabiya. For this part the best feature extraction technique which has been recognized from the previous chapter named Chi-square has been used as FE technique and applying with various n-gram such as unigram, bigram, and trigram in order to determine which combination of Chi-square and n-gram features that could lead to the robustness of the sentiment classification. In addition, this experiment has performed on already resampled datasets for all three news channels using the oversampling technique Bootstrapping since the problem of imbalanced data which has encountered within topic categorization has emerged here also for sentiment classification as can be seen in the following sections. In addition, the same machine learning classifier as for topic categorization has been used for sentiment classification named SVM.

5.2 Descriptive Analysis of Three News Channels

In this section, a descriptive analysis for each news channel is introduced as follows:

5.2.1 Al-Arabiya News Channel

Figure 5.1 shows the distribution of the sentiment categories in this dataset throughout the year. There is a significant difference in term of number of posts between categories, whereby the number of news posts belonging to afraid, sad and

angry categories are much higher with the rate of 49%, 28%, and 13%, compared to other categories such as inspired, amused, annoyed, happy, and don't care as it can be seen from the Table 5.1.

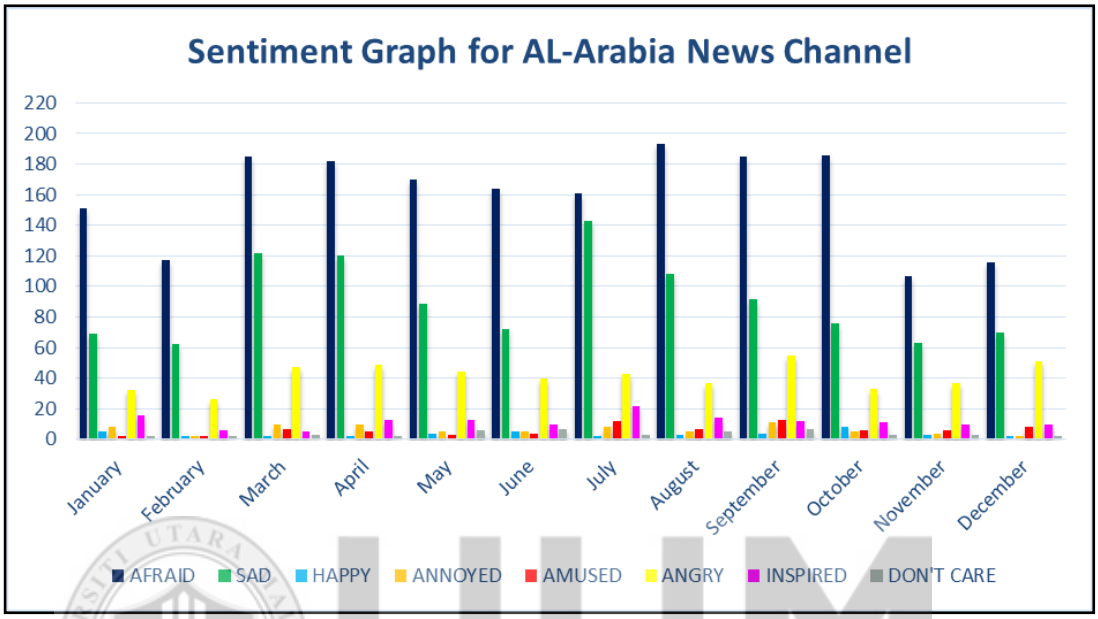

Figure 5.1.
Sentiment Graph for Al-Arabiya News Channel

As shown from Table 5.1 it can be noticed that the largest number of posts for each month belong to the feelings of fear, sadness, and then anger, respectively. That is because conflict and terrorism have dominated the events in each month as it can be seen from Figure 4.1 in Chapter 4, Section 4.2.1. Additionally, the highlighted cells in the table represent the highest number of posts exist in each sentiment class. According to Figure 5.1 the posts by the category sad represent almost half of class afraid for all months except for July in which it is noted that the categories are almost have the same number of posts that's probably due to the high number of posts of airplane crash that released in that month and consequently led to increase in sad news. Similarly, class angry is also about half the class of sad. Sentiment inspired represents just 4% of total posts while sentiments like annoyed and amused have the

same percentage at 2%, followed by happy and don't care which is representing the least ratio of 1% as shown in Table 5.1.

Table 5.1
Number of Posts Per Sentiment Category for AL-ARABIYA News Channel

Month	AFRAID	SAD	HAPPY	ANNOYED	AMUSED	ANGRY	INSPIRED	DON'T CARE
January	151	69	5	8	2	32	16	2
February	117	62	2	2	2	26	6	2
March	185	122	2	10	7	47	5	3
April	182	120	2	10	5	49	13	2
May	170	89	4	5	3	44	13	6
June	164	72	5	5	4	40	10	7
July	161	143	2	8	12	43	22	3
August	193	108	3	5	7	37	14	5
September	185	92	4	11	13	55	12	7
October	186	76	8	5	6	33	11	3
November	107	63	3	4	6	37	10	3
December	116	70	2	2	8	51	10	2
Sum	**1917**	**1086**	**42**	**75**	**75**	**494**	**142**	**45**
Average	**49%**	**28%**	**1%**	**2%**	**2%**	**13%**	**4%**	**1%**

5.2.2 Al-Jazeera News Channel

Unlike al-Arabiya news channel, the number of posts in the category sad is almost equal to the number of posts in category fear in every month and sometimes has exceeded it for certain months like July and December as shown from Figure 5.2. This happened, maybe because news is overwhelmed by conflicts and terrorism events in every month as well as each month contains more posts for the events (diseases and natural disasters) compared with al-Arabiya as shown from Tables (5.1, 5.2), whereby these events often contain sad news. Similarly, a noticeable increase for sentiment angry is observed for some months, especially in (January and

February) that is may be due to the large number of posts belong to the events like conflict, terrorism which often associated with angry emotion.

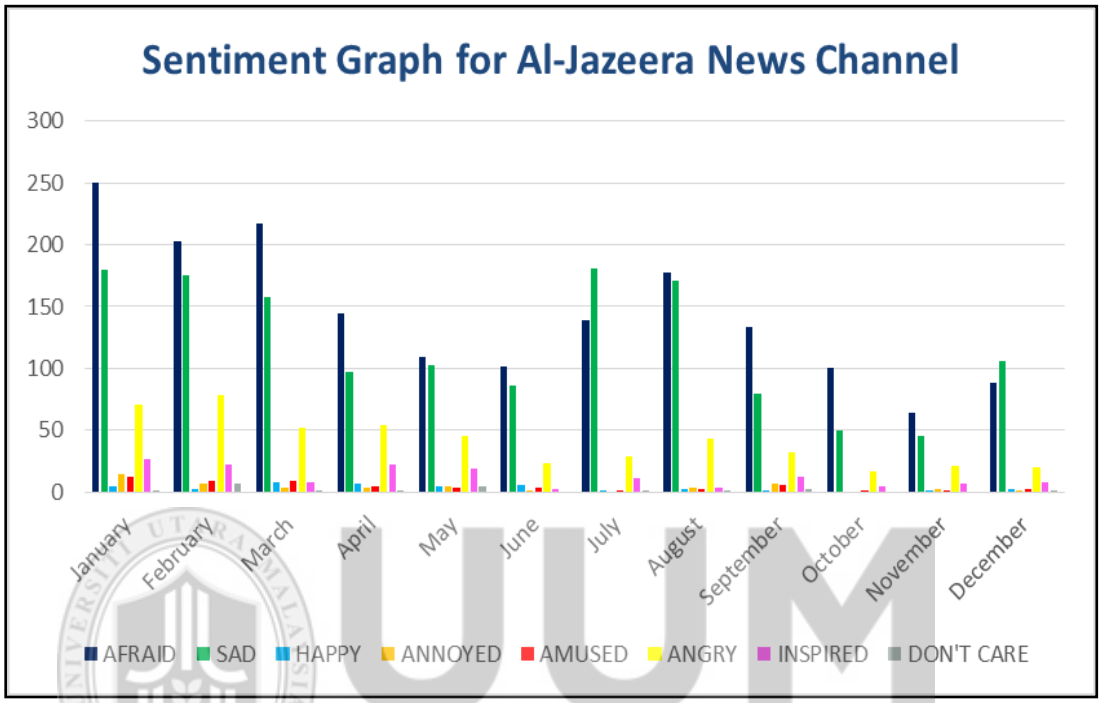

Figure 5.2.
Sentiment Graph for Al-Jazeera News Channel

Based on the data of Table 5.2, the distribution of the sentiments during the year is similar to al-Arabiya news channel. The most emotions used are afraid, sad, and angry at ratios of 43%, 36%, and 12% respectively. Followed by inspiration sentiment by 4% same as for previous channel. The least number of posts in this channel belongs to (happy, annoyed, don't care) sentiments by only 1%.

Table 5.2
Number of Posts Per Sentiment Category for Al-Jazeera News Channel

Month	AFRAID	SAD	HAPPY	ANNOYED	AMUSED	ANGRY	INSPIRED	DON'T CARE
January	250	180	5	15	13	71	27	2
February	203	175	3	7	9	78	22	7
March	217	158	8	4	9	52	8	2
April	145	97	7	4	5	54	22	2
May	109	103	5	5	4	45	19	5
June	102	86	6	2	4	24	3	0
July	139	181	2	0	2	29	11	2
August	178	171	3	4	3	43	4	2
September	134	80	2	7	6	32	12	3
October	100	50	0	0	2	17	5	0
November	64	45	2	3	2	21	7	0
December	88	106	3	2	3	20	8	2
Sum	**1729**	**1432**	**46**	**53**	**62**	**486**	**148**	**27**
Average	**43%**	**36%**	**1%**	**1%**	**2%**	**12%**	**4%**	**1%**

5.2.3 BBC News Channel

According to Figure 5.3, like the previous two channels, two categories of emotions such as afraid and sad dominate most emotions during the year for this channel also. However, with a little bit different in that the rates for both emotions are almost convergent by 47%, 40%, respectively as can be seen from Table 5.3. The large number of news posts for six months belong to the category of afraid, likewise the same number of months for class sad. While at third place sentiment angry comes with a rate of 8%, followed by the least categories such as happy, annoyed, and amused by 1%. While don't care class contains a negligible number of posts, only

four posts during the whole year. In addition, the highlighted cells in the same table indicate to the large number of news posts in each sentiment category.

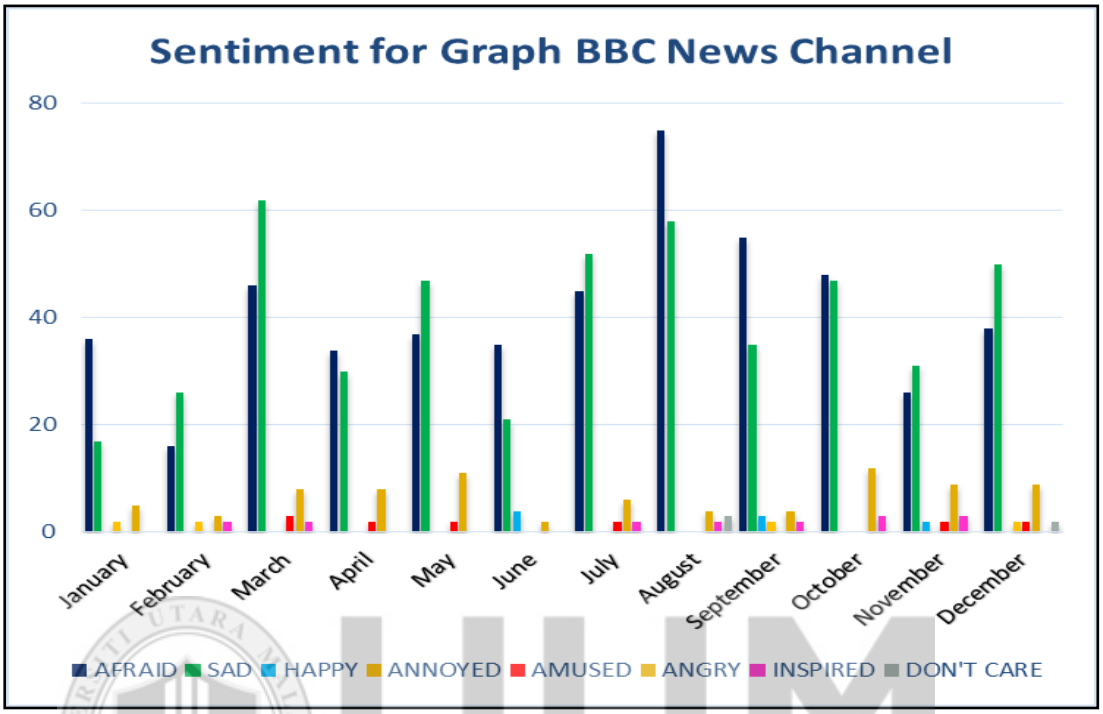

Figure 5.3.
Sentiment Graph for BBC News Channel

Moreover, according to Figure 5.3 a significant increase in afraid category has noticed in August by (75) posts. That is probably because this month contains a large number of posts that belong to natural disaster and conflict class as shown from Table 4.3 in Chapter 4, Section 4.2.2, whereby these events often associated with sentiments of fear. Similarly, a considerable rise in sad class has noticed for month March, perhaps due to a higher number of posts belong to airplane crash category which usually hold sad emotion.

Table 5.3
Number of Posts Per Sentiment Category for BBC *News Channel*

Month	AFRAID	SAD	HAPPY	ANNOYED	AMUSED	ANGRY	INSPIRED	DON'T CARE
January	36	17	0	2	0	5	0	0
February	16	26	0	2	0	3	2	0
March	46	62	0	0	3	8	2	0
April	34	30	0	0	2	8	0	0
May	37	47	0	0	2	11	0	0
June	35	21	4	0	0	2	0	0
July	45	52	0	0	2	6	2	0
August	75	58	0	0	0	4	2	2
September	55	35	3	2	0	4	2	0
October	48	47	0	0	0	12	3	0
November	26	31	2	0	2	9	3	0
December	38	50	0	2	2	9	0	2
Sum	**491**	**418**	**9**	**8**	**13**	**81**	**16**	**5**
Average	**47%**	**40%**	**1%**	**1%**	**1%**	**8%**	**2%**	**0%**

Figure 5.4, Figure 5.5, and Figure 5.6 plot the values from the Table 5.4, Table 5.5, and Table 5.6, respectively, to show the events happened in the year 2014 and its associated sentiments for each news channel. As can be seen from the three figures that for the two events such as conflict and terrorism the most dominated sentiments are afraid, sad, and angry. While sadness is overwhelming the airplane crash event followed by fear, especially for the BBC news channel where there is a significant increase in saddens emotions for this event. On the other hand, the events like disease and natural disaster for all news channels have a higher proportion of sadness and fear sentiments compared with other emotions that are barely visible.

Table 5.4

Number of Posts Per Sentiment Category for Each Topic (Al-Arabiya News Channel)

Event	Afraid	Sad	Happy	Annoyed	Angry	Inspired	Amused	Don't Care
Conflict	1257	677	35	67	308	109	46	33
Terrorism	507	249	6	7	183	27	22	11
Airplane Crash	42	118	6	1	3	1	6	1
Disease	81	15	0	1	0	6	5	1
Natural Disaster	30	27	0	0	0	0	0	0
Total	**1917**	**1086**	**47**	**76**	**494**	**143**	**79**	**46**

Table 5.5

Number of Posts Per Sentiment Category for Each Topic (Al-Jazeera News Channel)

Event	Afraid	Sad	Happy	Annoyed	Angry	Inspired	Amused	Don't Care
Conflict	1056	926	41	58	347	91	39	20
Terrorism	357	265	1	2	122	19	11	7
Airplane Crash	28	88	0	2	8	3	3	0
Disease	168	56	5	0	8	28	10	1
Natural Disaster	120	97	1	1	3	7	2	0
Total	1729	1432	48	63	488	148	65	28

Table 5.6

Number of Posts Per Sentiment Category for Each Topic (BBC News Channel)

Event	Afraid	Sad	Happy	Annoyed	Angry	Inspired	Amused	Don't Care
Conflict	213	165	5	6	40	0	9	6
Terrorism	85	111	0	4	32	5	2	0
Airplane Crash	26	110	2	0	7	3	4	0
Disease	82	18	0	0	2	8	2	0
Natural Disaster	85	72	2	0	0	2	0	0
Total	491	476	9	10	81	18	17	6

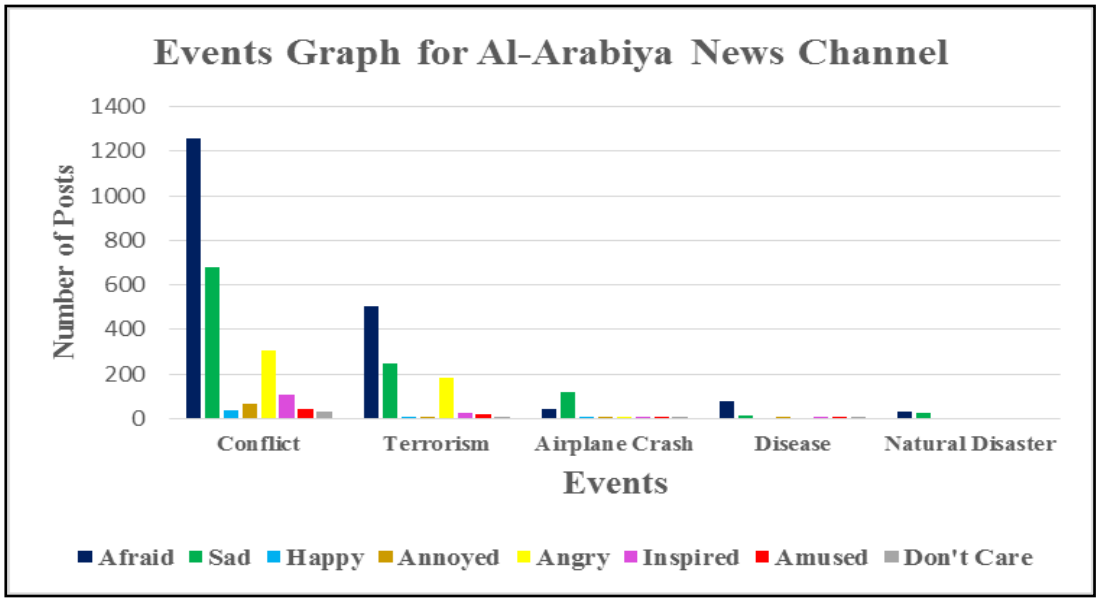

Figure 5.4.
Event Graph for Al-Arabiya News Channel

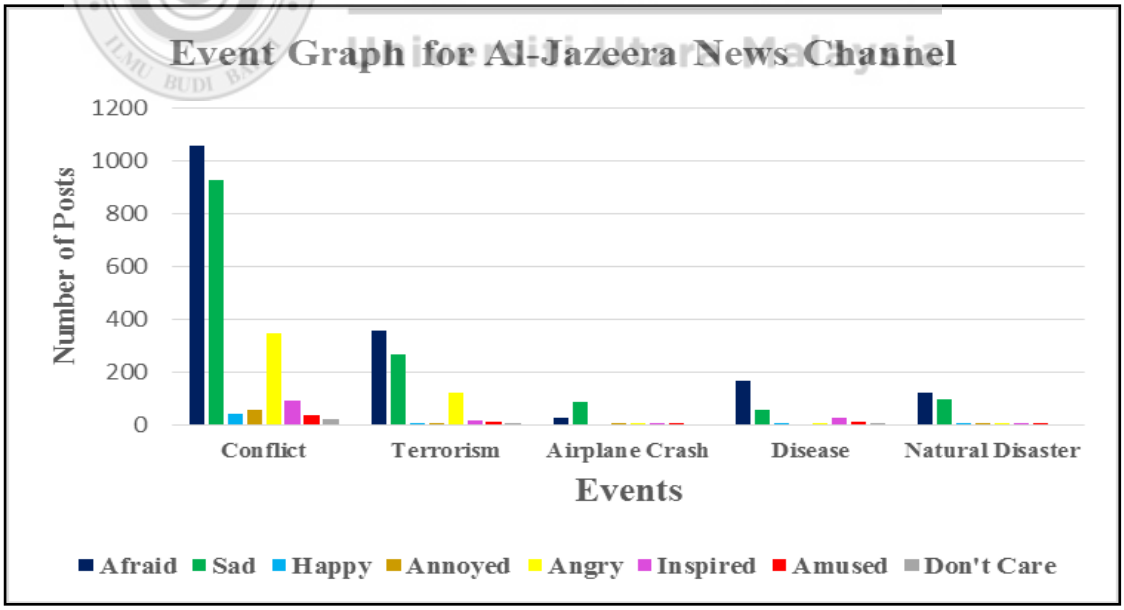

Figure 5.5.
Event Graph for Al-Jazeera News Channel

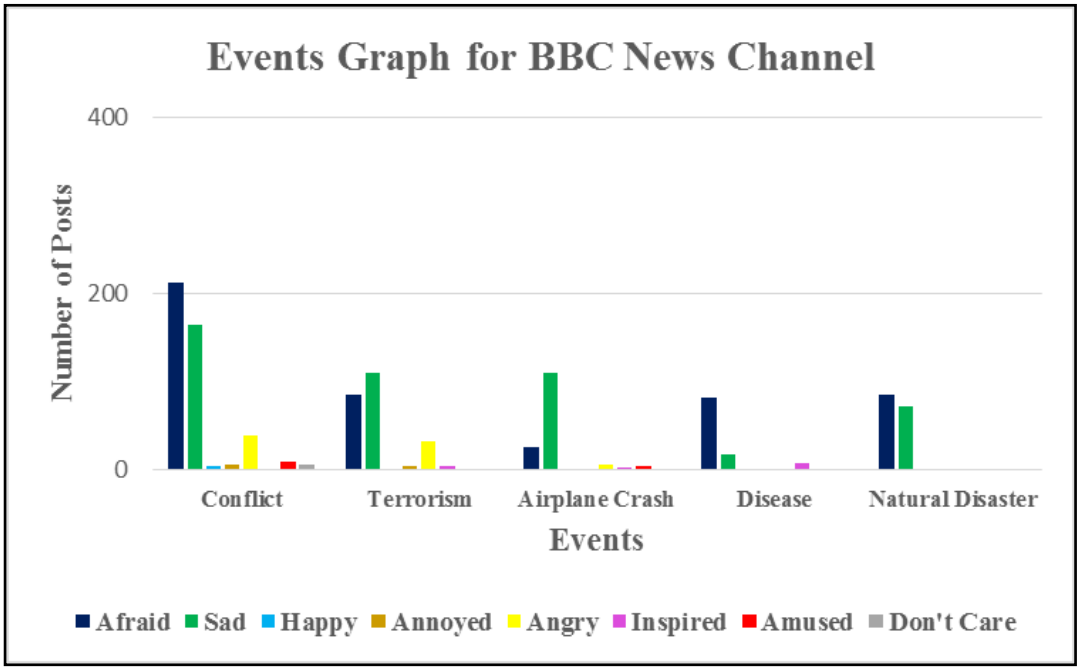

Figure 5.6.
Event Graph for BBC News Channel

In addition, almost all events in all news channels have very low number of posts for the sentiments such as happy, annoyed, inspired, amused, and don't care. This may be happened because usually the emerging of events such as conflict, terrorism, airplane crash, disease, and natural disaster which used for this study lead to a rise in feelings of sadness, fear, and anger in people more than the others emotions. However, except two events conflict and terrorism from al-Arabiya and Aljazeera news channels which contain quite number of posts for these feelings as shown in Figure 5.4 and Figure 5.5.

5.3 Analyse Of N-Gram Features Based On Chi-Square Technique for Resampled Datasets

In order to achieve the first sub-objective of this study for sentiment classification, which is to analyse FE techniques that could lead to the determination of temporal robust features from a large number of news posts. This study has

combined Chi-square (which has realized from topic categorization experiments as best feature extraction technique) with a data resampling technology called Random OverSampling (ROS), in order to overcome the problem of imbalanced data and improve the accuracy performance of SVM classifier for the sentiment classification. In addition, Chi-square technique has applied on the three n-gram features (unigram, bigram, and trigram) on already resampled datasets of the three news channel namely, Al-Arabiya, Al-Jazeera, and BBC. The outcome results for each news channel is presented and discussed as follows:

5.3.1 Al-Arabiya News Channel

Based on the data of the Table C1 in Appendix C, the best accuracy has achieved in January using unigram as a feature set at rate of 70.36%. However, this accuracy has a bit high score of RMSE at 0.756, while both WMR and WMP are considered to be low at 45.93% and 52.97%, respectively.

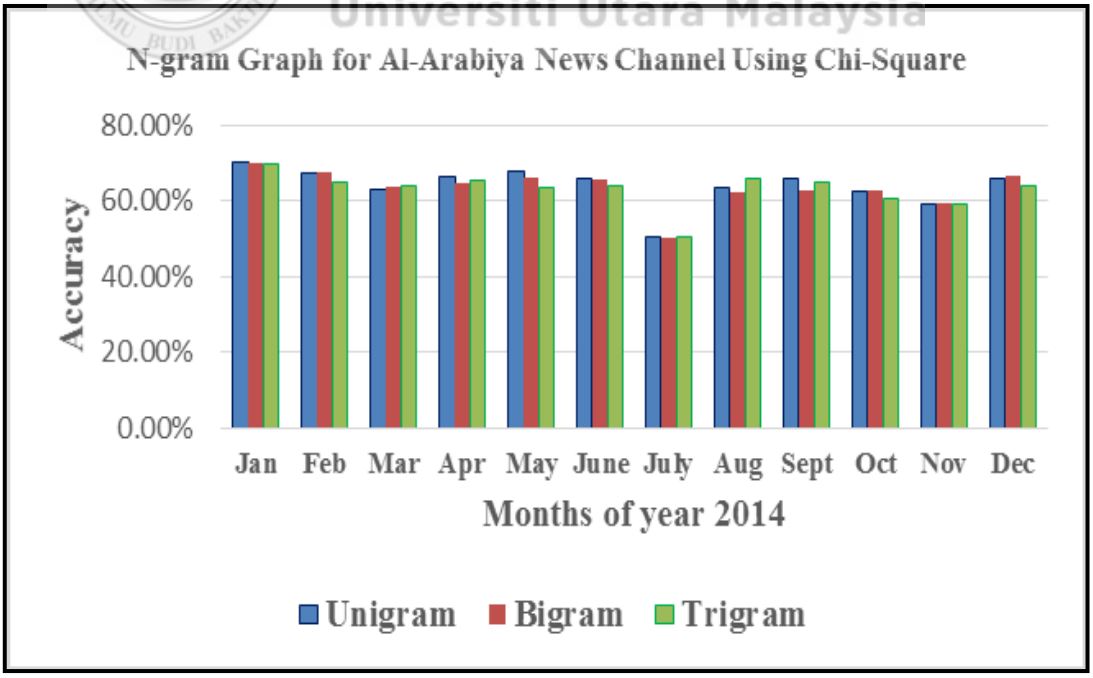

Figure 5.7.
N-gram Graph for Al-Arabiya News Channel

According to Figure 5.7, the accuracy achieved by the three n-gram features for all months, almost similar. Chi has attained the lowest accuracy in July though its rates for WMR and WMP were higher compared to what has been achieved for optimum accuracy in month January as it is observed from Table C1. In general, the rates for WMR and WMP for almost all months are very low that may reach 24% and 36%, respectively. In addition, the optimum accuracy for six months over the year have achieved using the unigram as a feature set, while for the other six months bigram has obtained the optimum accuracy for four months while trigram for just two months. Thus, unigram has recognized as the effective n-gram features that led to achieve the highest accuracies with Chi-square technique for this news channel dataset.

5.3.2 Al-Jazeera News Channel

As shown from Table C2 in Appendix C, the highest accuracy has acquired in October with a rate of 77.01% that is higher than al-Arabiya channel by 7%. This accuracy has achieved using unigram as a feature set, and has a reasonable value of RMSE at 0.593, while both WMR and WMP have quite good values 61.55% and 72.41%, respectively. On the other hand, according to Figure 5.8 the lowest rates for accuracies has achieved in two months; February using bigram and May by applying unigram at ratio of 36% and 20%, respectively.

In addition, except the two months February and May the other months have a high accuracy rate ranging between (60% and 80%) for the three types of n-gram features. It can be seen from Table C2 that unigram feature has the supreme accuracy for five months while both the bigram and trigram features came in second place

with an equal number of months, three months for each. Hence, it is concluded that unigram consider to be the most effective n-gram feature for this dataset as well.

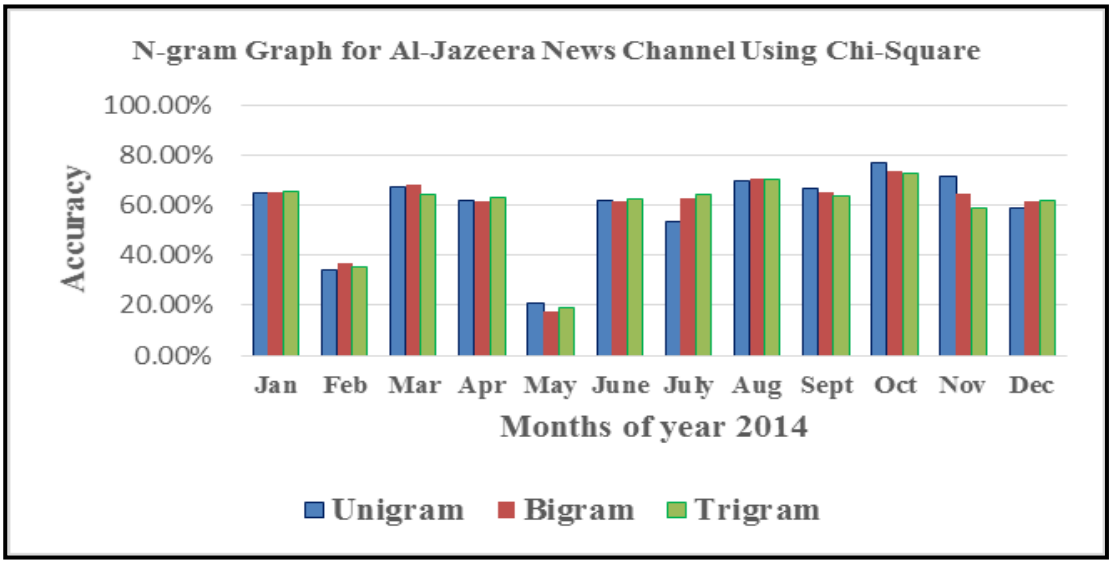

Figure 5.8
N-gram Graph for Al-Jazeera News Channel

5.3.3 BBC News Channel

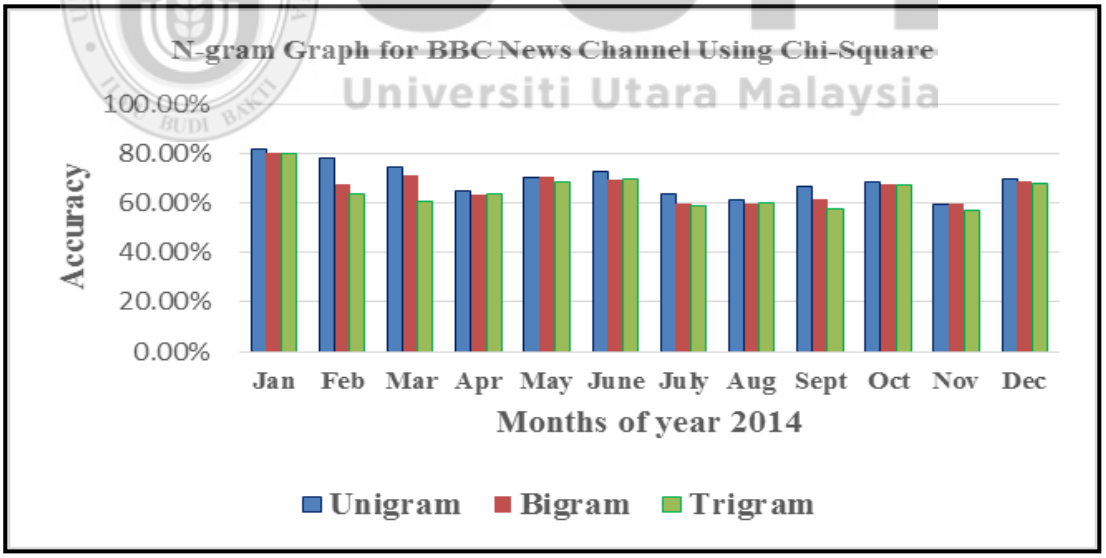

Figure 5.9.
N-gram Graph for BBC News Channel

Form Table C3 in Appendix C, it can be noticed that the best accuracy has achieved in January by 81.67%, which is higher than the best accuracies for the previous two news channels. This accuracy has used unigram also as a feature set.

However, this accuracy has a little high RMSE of 0.569, but the rates for both WMR and WMP are considered to be low at 61% and 64%, respectively. In contrast, the lowest value of accuracy has attained in November by 59% with low rates also for both WMR and WMP at 50% and 45%, individually.

According to Figure 5.9, Chi has almost similar accuracies for all three n-gram features over the year, except for two months named February and March, whereby there is a quite difference between the accuracies but with best accuracy belong to unigram features. Furthermore, the highest accuracies that have been achieved through the year are ranging from 60% to 80%. In addition, it is noticed that the optimum accuracy for each month has been achieved using unigram as a feature set. Thus, unigram represents the best n-gram features for this channel's dataset as it was for the other previous news channels' datasets.

5.3.4 Overall Optimum Accuracy for Three News Channels

To sum up, it has been concluded that almost all optimum accuracies for all months for the three news channels have been achieved using Chi- square with unigram as n-gram features. The optimum accuracies for al-Arabiya, Aljazeera, and BBC are 70%, 77% and 81%, respectively, as shown from Tables 5.7 to Table 5.9. These best accuracies have been achieved since there is no significant difference in the distribution of news posts over all categories for the related months as well as the number of sentiment categories involved in the classification process is small.

For example, for Aljazeera news channel the optimal accuracy has been obtained in October, whereby in this month there is just five categories involved in the classification process. Furthermore, there is no major difference in the term of

number of news posts in each category. Oppositely, for same news channel the lowest accuracy has been achieved in May as there is a significant variation in news posts distribution over all eight classes which involved in the classification task and consequently has led to a very low accuracy.

Table 5.7
Optimum Accuracy for Al-Arabiya News Channel Using Chi-square

MONTH	N-GRAM	ACCURACY
JAN	1	70.36%
FEB	1	67.28%
MARCH	2	63.73%
APRIL	1	66.31%
MAY	1	67.94%
JUNE	1	65.80%
JULY	2	50.27%
AUG	3	65.86%
SPET	1	65.96%
OCT	2	62.50%
NOV	3	59.23%
DEC	2	66.28%

Table 5.8
Optimum Accuracy for Al-Jazeera News Channel Using Chi-square

MONTH	N-GRAM	ACCURACY
JAN	3	70.36%
FEB	2	67.28%
MARCH	2	63.73%
APRIL	3	66.31%
MAY	1	67.94%
JUNE	1	65.80%
JULY	3	50.27%
AUG	2	65.86%
SPET	1	65.96%
OCT	1	77.01%
NOV	1	71.29%
DEC	2	61.56%

Table 5.9
Optimum Accuracy for BBC News Channel Using Chi-square

MONTH	N-GRAM	ACCURACY
JAN	1	81.67%
FEB	1	77.78%
MARCH	1	74.33%
APRIL	1	64.76%
MAY	1	70.47%
JUNE	1	72.69%
JULY	1	63.55%
AUG	1	61.21%
SPET	1	66.38%
OCT	1	68.38%
NOV	1	59.52%
DEC	1	69.83%

In addition, according to same tables, unigram features have proved to be the most effective n-gram feature set for SVM classifier as it has led to the highest accuracy for the most months for all three news channels. However, these optimal accuracies have pretty high values for RMSE ranging from 0.6 to 0.8. In contrast, WMR and WMP have relatively better rates fluctuating from 62% to 72% for both Aljazeera and BBC news channels compared to al-Arabiya, which has a lower ratio of 46% and 53% for WMR and WMP, respectively refer to Tables C1 to C3 in Appendix C. Practically all three types of n-gram features have almost similar accuracies for all months of the three news channels as shown from Figure 5.7 to Figure 5.9.

Furthermore, according to Table 5.10 the highest optimal accuracy for Al-Jazeera, Al-Arabiya, and BBC news channels have been achieved at 77.01%, 70.36%, 81.87% based on resampling datasets respectively. Moreover, it is confirmed that the combination of Chi-square and unigram features has led to achieve the highest optimum accuracies for all months of the three news channels.

Thus, this study has identified Chi-square technique with unigram features as best feature extraction technique that could lead to the determination of temporal robust sentiment features from large number of news posts on Facebook, and hence achieve the first sub-objective of this study for sentiment classification.

Table 5.10
Highest Sentiment classification Accuracy achieved for Three News Channels On Resampled Datasets Using Chi-square + Unigram

News Channel Name	Type of Dataset	Feature Extraction Technique	N-gram	Month	Accuracy
Al-Arabiya	Resampled	CHI	1	January	70.36%
Al-Jazeera	Resampled	CHI	1	October	77.01%
BBC	Resampled	CHI	1	January	81.67%

5.3.5 Determine of Fifteen Robust Features for Sentiment Classification

Based on the highest optimal accuracies shown in Table 5.10, the top fifteen robust features for sentiment classification have been extracted from the three news channel datasets. The top fifteen features for Al-Arabiya, Aljazeera and BBC are unigram features which have been extracted, ranked, selected using Chi-square technique as can be seen from Tables 5.11 till 5.13. According to Table 5.11, Table 5.12, and Table 5.13, it can be seen that some features are appearing in more than one news channel dataset such as violence which appears in both BBC and Aljazeera datasets, but having a different score values based on the dataset they belong to. Furthermore, typically each database for each channel contains different features from other channels based on the sentiments that emerged in the relevant month from which these features have been extracted.

Table 5.11
Top (15) Robust Temporal Features for sentiment classification (Al-Arabiya News Channel)

Feature	AFRAID	SAD	HAPPY	ANNOYED	AMUSED	ANGRY	INSPIRED	DON'T CARE
Army	0.21	0.17	0.14	0.00	0.26	0.00	0.00	0.00
Base	0.23	0.00	0.25	0.00	0.00	0.00	0.00	0.00
Border	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00
Checkpoint	0.23	0.00	0.22	0.24	0.00	0.00	0.00	0.00
Clash	0.23	0.18	0.19	0.00	0.00	0.21	0.00	0.00
Decide	0.22	0.23	0.26	0.00	0.00	0.00	0.00	0.00
End	0.18	0.17	0.00	0.17	0.00	0.00	0.00	0.00
Leave	0.32	0.19	0.00	0.22	0.00	0.00	0.00	0.00
militari	0.19	0.17	0.21	0.00	0.00	0.00	0.00	0.00
Oust	0.29	0.24	0.00	0.00	0.00	0.33	0.00	0.00
Ride	0.26	0.00	0.00	0.26	0.00	0.00	0.00	0.00
support	0.25	0.19	0.00	0.00	0.00	0.28	0.00	0.00
Troop	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00
Death	0.00	0.22	0.18	0.00	0.00	0.00	0.00	0.00
Hit	0.27	0.00	0.00	0.00	0.31	0.00	0.00	0.00

Table 5.12
Top (15) Robust Temporal Features for sentiment classification (Al-Jazeera News Channel)

Feature	AFRAID	SAD	HAPPY	ANNOYED	AMUSED	ANGRY	INSPIRED	DON'T CARE
Achieve	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32
Credit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32
Crowd	0.15	0.00	0.26	0.00	0.00	0.00	0.00	0.00
Declar	0.28	0.00	0.00	0.28	0.00	0.00	0.00	0.00
Ebola	0.24	0.16	0.07	0.17	0.00	0.00	0.00	0.15
Fighter	0.18	0.14	0.00	0.00	0.00	0.00	0.00	0.00
Free	0.00	0.00	0.00	0.37	0.00	0.00	0.00	0.00
Grave	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00
impress	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32
infect	0.25	0.00	0.00	0.31	0.00	0.00	0.00	0.00
Miss	0.00	0.45	0.00	0.00	0.00	0.00	0.00	0.00
release	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00
student	0.09	0.39	0.20	0.00	0.00	0.00	0.00	0.21
Troop	0.12	0.23	0.00	0.00	0.00	0.00	0.00	0.00
violence	0.16	0.29	0.22	0.00	0.00	0.00	0.00	0.00

Table 5.13
Top (15) Robust Temporal Features for sentiment classification (BBC News Channel)

Feature	AFRAID	SAD	HAPPY	ANNOYED	AMUSED	ANGRY	INSPIRED	DON'T CARE
accused	0.00	0.15	0.00	0.00	0.00	0.00	0.30	0.00
Block	0.00	0.09	0.16	0.00	0.00	0.00	0.00	0.00
Car	0.13	0.09	0.00	0.00	0.00	0.00	0.00	0.00
End	0.11	0.14	0.00	0.00	0.00	0.00	0.00	0.00
Force	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00
Freez	0.25	0.00	0.12	0.00	0.00	0.00	0.00	0.00
Group	0.17	0.00	0.33	0.00	0.00	0.00	0.00	0.00
Held	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00
impose	0.33	0.00	0.44	0.00	0.00	0.00	0.00	0.00
resign	0.00	0.33	0.12	0.00	0.00	0.00	0.00	0.00
resolution	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00
sanction	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00
surround	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00
violence	0.11	0.14	0.00	0.00	0.00	0.00	0.00	0.00
Warn	0.00	0.10	0.12	0.00	0.00	0.00	0.00	0.00

In addition, the highlighted values in the tables indicate the highest weight for these features since any feature can belong to more than one sentiment. Thus, the optimal weight would determine which category the feature belongs. For example, in Table 5.11 feature (clash) has a predominant weight in afraid (0.23), while (death) has a predominant weight in sad (0.54) compared to other sentiments. As can be seen from the same tables, that the sentiment features are indeed informative words rather than sentiment words. This perhaps leads to poor performance of the SVM classifier in learning the features of each sentiment category and subsequently caused in very low accuracy for the classification process.

5.4 News Post Classification on Sentiment Classification Using SVM

In order to achieve the second sub-objective of this study for sentiment classification which is to classify the news post into its corresponding sentiment categories for the social events, famous machine learning technique called SVM has been applied to classify the news post to its appropriate sentiment category. Tables 5.14 till 5.16 show the output results of performing the sentiment classification process for the three news channels on resampled datasets using Chi-square as feature extraction technique and unigram as n-gram features and which subsequently leads to achieve the highest sentiment classification accuracy.

Table 5.14 *News Posts Classified on Sentiment Classification using SVM (Al-Arabiya News Channel)*

ACCURACY: 70.36%

	True Afraid	**True Sad**	**True Angry**	**True Inspired**	**True Happy**	**True Don't Care**	**True Amused**	**True Annoyed**	**Class Precision**
Pred. Afraid	159	28	27	12	3	0	1	5	**67.66%**
Pred. Sad	2	26	1	0	0	0	0	0	**89.66%**
Pred. Angry	3	1	4	0	0	0	0	1	**44.44%**
Pred. Inspired	0	0	0	3	0	0	0	0	**100.00%**
True Happy	0	0	0	0	3	0	0	0	**100.00%**
True Don't Care	0	0	0	0	0	3	0	0	**100.00%**
True Amused	0	0	0	0	0	0	4	0	**100.00%**
True Annoyed	0	0	0	0	0	1	0	0	**0.00%**
Class Recall	**96,95%**	**47.27%**	**12.50%**	**20.00%**	**50.00%**	**75.00%**	**80.00%**	**0.00%**	

Table 5.15 *News Posts Classified on Sentiment Classification using SVM (Al-Jazeera News Channel)*

ACCURACY : 77.01%

	True Sad	True Afraid	True Angry	True Inspired	True Amused	Class Precision
Pred. Sad	27	1	0	0	0	**96.43%**
Pred. Afraid	27	95	9	2	0	**71,43%**
Pred. Angry	0	1	6	0	0	**85,71%**
Pred. Inspired	0	0	0	4	0	**100.00%**
True Amused	0	0	0	0	2	**100.00%**
Class Recall	**50.00%**	**97.94%**	**40.00%**	**66.67%**	**100.00%**	

Table 5.16 *News Posts Classified on Sentiment Classification using SVM (BBC News Channel)*

ACCURACY : 81.67%

	True Angry	True Afraid	True Sad	True Annoyed	Class Precision
Pred. Angry	5	0	0	0	**100.00%**
Pred. Afraid	0	39	10	1	**78.00%**
Pred. Sad	0	0	4	0	**100.00%**
Pred. Annoyed	0	0	0	1	**100.00%**
Class Recall	**100.00%**	**100.00%**	**28.57%**	**50.00%**	

In addition, it can be concluded from the same Tables that almost all categories have high precision rates for the three news channels' datasets except for angry and annoyed categories in Al-Arabiya dataset which have very low proportion as shown in Table 5.14. Oppositely, most categories have low recall rates especially some categories such as sad in BBC dataset as well as angry, inspired, and annoyed classes in Al-Arabiya dataset as seen in Tables 5.14 and Figure 5.16. Additionally,

afraid category has quite good rates for both precision and recall metrics for all three news channels' datasets.

Furthermore, it is concluded that the obtained classification accuracies are in order of 81.87%, 77.01%, 70.36%, for BBC, Al-Jazeera, Al-Arabiya news channels. The highest accuracy belong to BBC dataset as there is few number of classes included in the classification process as well as no significant difference in the distribution of news posts over the categories involved in the process. Hence, the value of accuracy depend basically on the number of classes involved in the classification process and on the distribution of news posts over these categories.

5.5 Evaluation of Chi+Unigram on Sentiment Classification Using Random Selected Dataset

In order to achieve the third sub-objective of this study for sentiment classification which is to evaluate the performance of the best feature extraction technique that has been realized (Chi-square+Unigram) on randomly selected news posts. The validation technique is an automatic 5-fold cross validation for the SVM classifier. Each evaluating process for each news channel has been done using various performance metrics such as Accuracy (A), Weighted Mean Recall (WMR), Weighted Mean Precision (WMP), Root Mean Squared Error (RMSE), and confidence score.

About (200) news posts released at any time in the year (2014) are selected randomly from each news channel's dataset (al-Arabiya, Aljazeera, BBC) in order to build three individual testing datasets. Then, the optimization sentiment classification models would run on these unseen news posts. The optimization model is which has

been realized for each news channel and has achieved an optimal performance accuracy in previous sections in months such as January for both al-Arabiya and BBC, while November for Aljazeera. Each model uses Chi-square as FE technique combine with unigram model and utilizes SVM as a machine learning classifier.

Table 5.17
Evaluation Metrics Sentiment Classification For Randomly selected Data

Channel	Accuracy	WMR	WMP	RMSE
Al-Arabiya	51.00%	13.16%	12.55%	0.800
Al-Jazeera	43.50%	14.44%	17.31%	0.733
BBC	41.00%	16.65%	15.15%	0.713

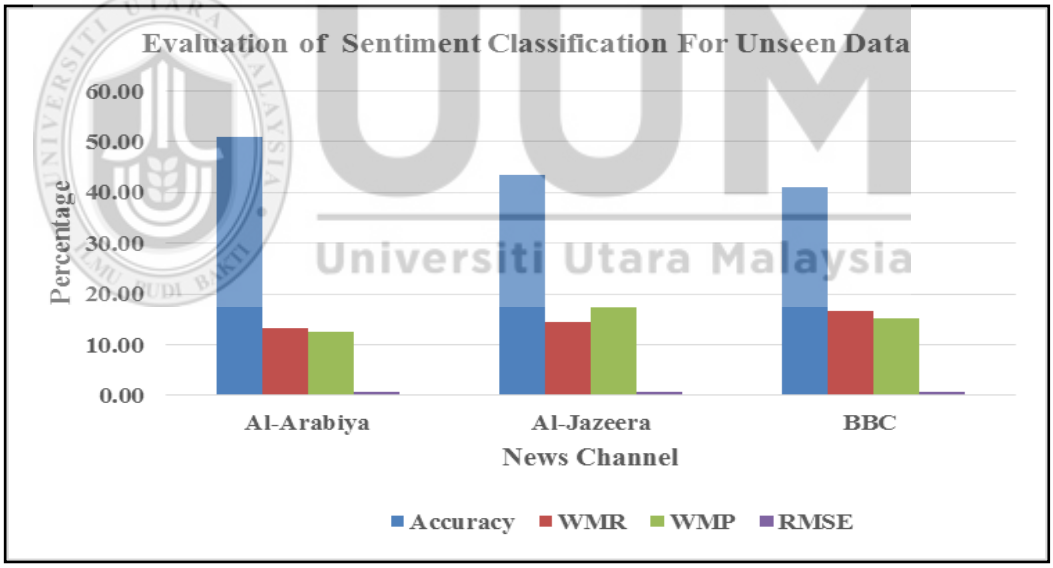

Figure 5.10.
Randomly Selected Data Evaluation for all News Channels

Table 5.14 shows the results for sentiment classification on unseen data for each news channel based on four evaluation metrics (Accuracy, WMR, WMP, and RMSE). The highlighted values in the table represent the highest value for each metric. The accuracies obtained for each news channel, namely Al-Arabiya, Al-

Jazeera, BBC are 51%, 43%, 41%, respectively. These accuracies are considered to be very low compared with what have been accomplished using cross-validation evaluation on resampled datasets as can be seen from Table 5.10 in Section 5.3.4. This is probably happened because the changes of sentiment over time. In other words, the training dataset of each news channel may not contain features that can contribute in detecting the sentiments emerged in the randomly selected news posts. Thus, could not classify the news posts to their sentiment category correctly and consequently decrease the confidence scores as shown in Table 5.15.

Figure 5.10 plots all values in Table 5.14 in order to provide a comparison between all these metrics. Consequently, it can be seen that the highest accuracy belongs to al-Arabiya channel, although it has the poorest values for other metrics (WMR, WMP) and the highest ratio for RMSE at 0.8. In contrary, BBC news channel has the lowest accuracy rate and by 10% difference from al-Arabiya.

Table 5.18
Confidence Value of Each Sentiment Category Prediction for Each News Channel

Channel	Afraid	Sad	Happy	Annoyed	Amused	Angry	Inspired	Don't Care
Al-Arabiya	0.287	0.108	0.105	0.103	0.104	0.103	0.104	0.087
Al-Jazeera	0.366	0.152	0.162	0	0.151	0.168	0	0
BBC	0.503	0.165	0.156	0	0	0	0.176	0

Table 5.15 presents the confidences score obtained for the prediction of each sentiment category for each news channel. According to Figure 5.14, it's clear that a higher confidence score of afraid class has attained for BBC news channels in 0.503. Additionally, BBC dataset has confidence scores for all eight sentiment classes while the other channels datasets named Aljazeera, Al-Arabiya contain zero confidence values for some categories such as annoyed, amused, angry, and don't care. This is

because these categories have zero number of news posts belong to them for the relevant month of the news channel dataset.

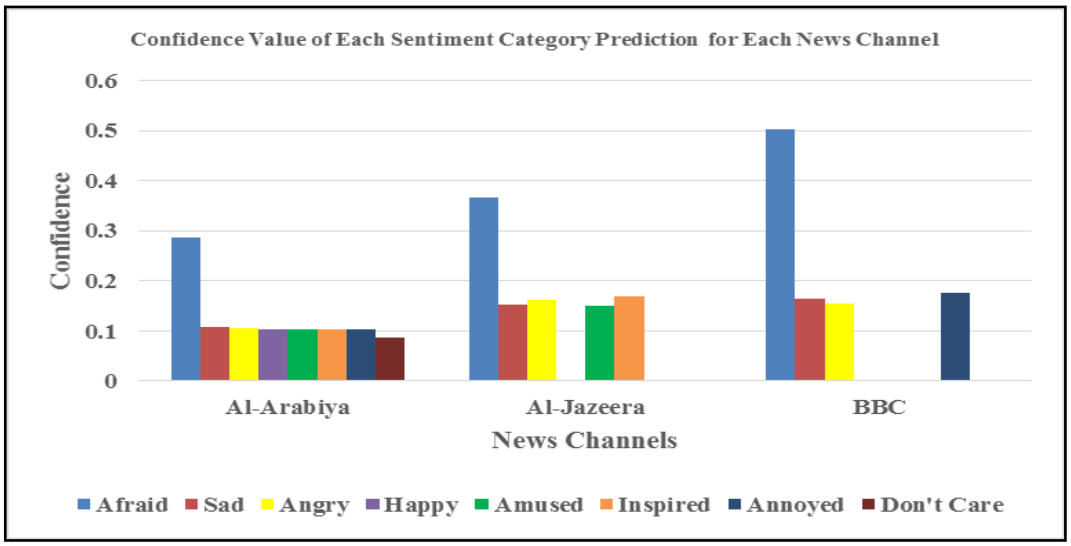

Figure 5.11.
Confidence Value of Each Sentiment Category Prediction for Each News Channel

5.6 Chapter Summary

In this chapter, sentiment classification experimental results and discussion were presented. Started with a descriptive analysis of each news channel's dataset, followed by introducing and discussing the results of applying various n-gram features with Chi-square feature extraction technique for sentiment classification on resampled datasets. Then, followed by the results of evaluating the best feature extraction technique (Chi+Unigram) on sentiment classification using randomly selected news posts from each news channel. In the next chapter, the conclusion, contribution, limitation, and future recommendation of this study has presented.

CHAPTER SIX

CONCLUSION & RECOMMENDATIONS

6.1 Conclusion

Based on the main objective of this study, which is to develop a comparative analysis on feature extraction techniques in order to recognize topics and sentiment classes for social event detection. This study has performed the following experiments for both topic and sentiment classification.

On one hand, for topic categorization, this study has performed the following three experiments. In first experiment, this study has examined different feature extraction techniques on a large amount of news posts from three Facebook news channels, namely Al-Jazeera, BBC and Al-Arabiya. Five feature extraction techniques which include TF-IDF, TF, BTO, IG, and Chi-square have been used for extracting different features types to determine optimal accuracy for topic classification. Additionally, these techniques have applied on diverse n-gram types such as unigram, bigram, and trigram. The results of this experiment have shown that Chi-square has proved to be a better feature extraction technique compared to other techniques as it leads to the highest classification accuracies on the original datasets at 88.89%, 87.96%, and 86.83% for BBC, Aljazeera, and Al-Arabiya dataset, respectively. Furthermore, unigram features have proved to be the most effective feature type which has assisted in obtaining the highest classification accuracy for topic classification.

The second experiment has studied the effectiveness of using a resampling technology named OverSampling technique Bootstrapping provided by RapidMiner application. In other words, the aim of this experiment is to enhance the accuracy performance of topic classification model by resampling the datasets as well as to do an analysing of n-gram features based on Chi-square technique for resampled datasets. Thus, this experiment has used Chi-square technique (which has been realized as best feature extraction technique from the first experiment), and has applied it on the three n-grams (unigram, bigram, and trigram). The evaluation results of this experiment have proved that applying resampling technique, namely, OverSampling has increased the classification performance accuracy by 5%, 6%, 4% for BBC, Al-Arabiya, and Al-Jazeera, respectively, and subsequently achieved higher accuracy of 93.37%, 92.89%, 91.92, compared to what have been obtained for the original datasets.

In the third experiment, this study has evaluated the performance of the best feature extraction technique that has been realized from the previous two experiments, where Chi-square+Unigram has applied on randomly selected news posts. The evaluation results for this experiment has shown a relatively very low accuracies at rates of 46%, 70%, 67% for BBC, Al-Arabiya, and Al-Jazeera, respectively. This low accuracies have obtained may be due to the changes in topics over time. In other words, the training dataset of each news channel may not contain features that can contribute in detecting the topics emerged in the randomly selected news posts. Thus, could not classify the news posts to their topic category correctly and consequently decrease the confidence scores.

On the other hand, for sentiment classification propose this study has performed two experiments. In first experiment, the best realized feature extraction technique Chi-square has been applied with different features types (unigram, bigram, and trigram) on already resampled datasets of news posts from three news channels on Facebook, namely Al-Jazeera, BBC and Al-Arabiya. The aim of this experiment is, to determine which combination of Chi-square and n-gram features could lead to achieve an optimal accuracy for sentiment classification. The experimental results have shown that Chi-square with unigram dataset has proved to be a better extracting technique compared to other combinations as it leads to the highest classification accuracies for the resampled datasets at rates of 81.67%, 77.01%, and 70.36% for BBC, Aljazeera, and Al-Arabiya dataset, respectively.

In the second experiment, the performance of the best feature extraction technique (Chi-square+Unigram) that has been realized from the previous experiment has been evaluated on randomly selected news posts. The evaluation results for this experiment has shown a very low accuracies at rates of 41%, 51%, 43.50% for BBC, Al-Arabiya, and Al-Jazeera, respectively. This low accuracies have obtained may be due to the changes in sentiments associated with topics over time. In other words, the training dataset of each news channel may not contain features that can contribute in detecting the sentiments emerged in the randomly selected news posts. Thus, could not classify the news posts to their sentiment category correctly and consequently decrease the confidence scores.

To sum up the results of all experiments, it is been concluded that Chi-square proved to be the best feature extraction technique as it contributes in achieving the highest classification accuracies for both topic categorization and sentiment

classification. Additionally, unigram features proved to be the most effective n-gram dataset as it assists in acquiring the optimal classification accuracies for both topic and sentiment classification. Furthermore, resampling the original datasets has enhanced the classification performance for minority classes of topics and sentiments. However, evaluating the optimal feature extraction technique (Chi-square+Unigram) on randomly selected news posts has obtained very low accuracies for topic and sentiment classification because of changes in topics and sentiments overtime.

6.2 Contribution of Study

Although, various feature extraction techniques have been implemented and examined with different n-grams models, but there is still needs to discover which combination of feature extraction technique and n-gram that would give better performance results. Thus, this study has developed a comparative analysis on feature extraction Techniques which has examined various feature extraction techniques (TF-IDF, TF, BTO, IG, Chi-square) with three different n-gram features (Unigram, Bigram, Trigram), and using SVM as a classifier. The aim of this study is to discover the optimal Feature Extraction Technique (FET) that could achieve optimum accuracy results for both topic and sentiment by determining the temporal robust features from large number of news posts on Facebook from the three news channels namely (Al-Jazeera, Al-Arabiya, and BBC). The experimental results has been concluded that the combination of Chi-square+Unigram is the best feature extraction technique as it contributes in achieving the highest classification accuracies for both topic categorization and sentiment classification.

As a result these temporal features would contribute in classifying the news posts on Facebook into their appropriate topics such as conflict, terrorism, airplane crash, disease, natural disaster as well as classifying news posts into their suitable sentiment categories like afraid, sad, happy, amused, inspired, don't care, angry, and annoyed. Subsequently, this study provides three different types of graphs (topic, sentiment, and event) which could be beneficial for anyone interested in a particular category/topic of news channels posts as well as investigating the changes in the opinions associated with these topics. These graphs have presented in Descriptive Analysis in Chapter 4 and Chapter 5 in Sections 4.2, 5.2, respectively. Those beneficiaries could be news channels owner, journalists, government, stakeholder or analyst, and for research area. More detailed about the benefits of this study for each type of the beneficiaries can be found in Chapter 1, Section 1.6.

Furthermore, based on our literature review there is no evidence shows that any research has done a comparative analysis study of feature extraction techniques for news posts on Facebook. Whereas most of previous studies have been conducted on news articles that usually collected randomly from various news resources as well as most of studies on news have performed only topic categorization or sentiment classification but not both of them. Thus, based on our literature review, this study is consider to be a stepping stone for other studies to perform further investigation on this field.

6.3 Limitations of Study

Although this study has reached its objectives, there is still some unavoidable limitations. First, due to time limit, this study has implemented only on three news channels as well as has involved just five categories of news for topic categorization.

Therefore, to generalize the results for larger categories, the research should have included more news channels and more news categories. Second, the feature extraction techniques which are involved in the study are just five. Third, this research used manual labelling to assign the suitable topic category to a news post which consider to be a time consuming method and not sufficiently reliable because of the influence of the human's education and cultural background. Fourth, this study uses a free online lexicon named DepecheMood for labelling the news posts to their appropriate sentiments. Additionally, the sentiments restricted to eight sentiments categories provided by the free online site. Finally, only one resampling technology called OverSampling Bootstrapping technique has been used in order to overcome the problem of imbalanced datasets and consequently improve the classification performance. Thus, other resampling techniques can be used to get a more general comparison for the improvement of classification performance.

6.4 Future Recommendations

Future works could be focused on extending the scope of the study so it can include more news channels as well as more categories of news such as entertainment, political, crimes, sports and so on. In addition, the combinations of FE techniques and n-gram models could be extended to include other FE techniques such as Pointwise Mutual Information.

Furthermore, implementing these combinations on another dataset such as news channels posts on twitter in order to validate the results and findings of this study. Moreover, other labelling techniques can be used such as for topic categorization like clustering instead of using the manual labelling. Similarly, other methods could be used for labelling the news posts to their appropriate sentiments

such as based on the writer's perspective through using the free online emotion lexicons on the web, or from a reader perspective by building our own voting system interface so it will be labelled by the readers to gain deeper understanding of their perspective instead of using free online labelling systems. Additionally, reduce the number of sentiment categories as well as apply other sampling techniques like undersampling or a combination of oversampling and undersampling in order to overcome the problem of imbalanced data and enhance the classification performance.

REFERENCES

[1] J. Akaichi, Z. Dhouioui, and M. J. Lopez-Huertas Perez, "Text mining facebook status updates for sentiment classification," in *System Theory, Control and Computing (ICSTCC), 2013 17th International Conference*, 2013, pp. 640–645.

[2] E. N. Neumann, *The spiral of silence Public opinion–our social skin*. Chicago: University of Chicago Press, 1993.

[3] E. Bjørkelund and T. Burnett, "Temporal Opinion Mining," no. June, p. 122, 2012.

[4] T. Fukuhara, H. Nakagawa, and T. Nishida, "Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events.," in *ICWSM*, 2007.

[5] R. Chakraborty, "D OMAIN K EYWORD E XTRACTION T ECHNIQUE : A N EW W EIGHTING M ETHOD," pp. 109–118, 2013.

[6] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, "Sentiment Analysis on Social Media," *2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min.*, pp. 919–926, 2012.

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

[8] J. Zhang, Y. Kawai, S. Nakajima, Y. Matsumoto, and K. Tanaka, "Sentiment bias detection in support of news credibility judgment," in *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, 2011, pp. 1–10.

[9] J. Allan, "Introduction to topic detection and tracking," in *Topic detection and tracking*, Springer, 2002, pp. 1–16.

[10] D. Clarke, P. Lane, and P. Hender, "Developing robust models for favourability analysis," in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 2011, pp. 44–52.

[11] P. D. Turney, "Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews," no. July, pp. 417–424, 2002.

[12] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Syst.*, vol. 24, no. 7, pp. 1024–1032, Oct. 2011.

[13] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 251–258, 2011.

[14] Z. Fu, X. Sun, J. Shu, and L. Zhou, “Plain Text Zero Knowledge Watermarking Detection Based on Asymmetric Encryption,” vol. 48, no. Cia, pp. 126–134, 2014.

[15] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, “Opinion Extraction, Summarization and Tracking in News and Blog Corpora.,” in *AAAI spring symposium: Computational approaches to analyzing weblogs*, 2006, vol. 100107.

[16] I. P. Cvijikj and F. Michahelles, “Monitoring trends on Facebook,” *Proc. - IEEE 9th Int. Conf. Dependable, Auton. Secur. Comput. DASC 2011*, pp. 895–902, 2011.

[17] M. Cataldi, U. Torino, L. Di Caro, U. Torino, C. Schifanella, and U. Torino, “a4-Cataldi,” 2010.

[18] J. Weng, Y. Yao, E. Leonardi, F. Lee, and B. Lee, “Event Detection in Twitter Event Detection in Twitter ∗ ,” 2011.

[19] G. Burnside, D. Milioris, and P. Jacquet, “One Day in Twitter: Topic Detection Via Joint Complexity,” *Www*, 2014.

[20] S. Greener and A. Rospigliosi, *ePub - European Conference on Social Media: ECSM*, vol. 7. Academic Conferences Limited, 2014.

[21] D. Richter, P. D. D. K. Riemer, and J. vom Brocke, “Internet social networking,” *Wirtschaftsinformatik*, vol. 53, no. 2, pp. 89–103, 2011.

[22] S. Setty, R. Jadi, S. Shaikh, C. Mattikalli, and U. Mudenagudi, “Classification of facebook news feeds and sentiment analysis,” in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, 2014, pp. 18–23.

[23] J. K. Ahkter and S. Soria, “Sentiment analysis: Facebook status messages,” *Unpubl. master’s thesis, Stanford, CA*, 2010.

[24] A. E.-D. A. Hamouda and F. E. El-taher, “Sentiment Analyzer for Arabic Comments System,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 3, 2013.

[25] N. Bansal and N. Koudas, “BlogScope: a system for online analysis of high volume text streams,” *Proc. 33rd Int. Conf. Very large data bases*, pp. 1410–1413, 2007.

[26] H. Choi and H. Varian, “Predicting the present with google trends,” *Econ. Rec.*, vol. 88, no. s1, pp. 2–9, 2012.

[27] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, “Event extraction using behaviors of sentiment signals and burst structure in social media,” *Knowl. Inf. Syst.*, vol. 37, no. February, pp. 279–304, 2013.

[28] S. Bai, Y. Ning, S. Yuan, and T. Zhu, “Predicting Reader ’ s Emotion,” pp. 16–27, 2013.

[29] B. Thomas, “Exploration of Robust Features for Multiclass Emotion Classification,” pp. 1704–1709, 2014.

[30] J. Zhang, Y. Kawai, and T. Kumamoto, “Extracting Similar and Opposite News Websites Based on Sentiment Analysis,” in *Proc. of International Conference on Industrial and Intelligent Information (ICIII 2012)*, 2012, pp. 24–29.

[31] L. U. Ye and R. Xu, “E Motion Prediction of News Articles From Reader ’ S Perspective Based on Multi-Label Classi Fication,” pp. 15–17, 2012.

[32] C. G. Patil, “Use of Porter Stemming Algorithm and SVM for Emotion Extraction from News Headlines,” vol. 2, no. 7, pp. 9–13.

[33] G. Li and F. Liu, “A clustering-based approach on sentiment analysis,” *Proc. 2010 IEEE Int. Conf. Intell. Syst. Knowl. Eng. ISKE 2010*, pp. 331–337, 2010.

[34] J. Kamps, M. J. Marx, R. J. Mokken, and M. De Rijke, “Using wordnet to measure semantic orientations of adjectives,” 2004.

[35] J. Staiano and M. Guerini, “DepecheMood: a Lexicon for emotion analysis from crowd-annotated news,” *arXiv Prepr. arXiv1405.1605*, 2014.

[36] R. Akbani, S. Kwek, and N. Japkowicz, “Applying support vector machines to imbalanced datasets,” in *Machine Learning: ECML 2004*, Springer, 2004, pp. 39–50.

[37] F. Chang, J. Guo, W. Xu, and K. Yao, “A Feature Selection Method to Handle Imbalanced Data in Text Classification.,” *J. Digit. Inf. Manag.*, vol. 13, no. 3, p. 169, 2015.

[38] A. Zughrat, M. Mahfouf, Y. Y. Yang, and S. Thornton, “Support Vector Machines for Class Imbalance Rail Data Classification with Bootstrapping-based Over-Sampling and Under-Sampling,” in *19th World Congress of the International Federation of Automatic Control. Cape Town*, 2014.

[39] P. G. Preethi and V. Uma, “Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction,” *Procedia Comput. Sci.*, vol. 48, pp. 84–89, 2015.

[40] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, “Sentiment analysis in the news,” *Proc. Seventh Int. Conf. Lang. Resour. Eval.*, pp. 2216–2220, 2010.

[41] S.-M. Kim and E. Hovy, “Extracting opinions, opinion holders, and topics expressed in online news media text,” in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, 2006, pp. 1–8.

[42] R. H. W. Pinheiro, G. D. C. Cavalcanti, R. F. Correa, and T. I. Ren, "A global-ranking local feature selection method for text categorization," *Expert Syst. Appl.*, vol. 39, no. 17, pp. 12851–12857, 2012.

[43] "Top News on Facebook | Fan Page List." [Online]. Available: http://fanpagelist.com/category/news/view/list/sort/fans/page1. [Accessed: 06-Dec-2015].

[44] "BBC World News achieves major distribution milestone, reaching more than 330m households worldwide," 2012.

[45] "About BBC World News TV," 2011.

[46] "Media Use in the Middle East 2013 | Northwestern University in Qatar." [Online]. Available: http://menamediasurvey.northwestern.edu/. [Accessed: 09-Dec-2015].

[47] T. Johnson and S. Fahmy, "Who is winning the hearts and minds of the Arab public?," *Int. Commun. Res. J.*, vol. 45, no. 1–2, pp. 24–48, 2010.

[48] "Major Events in 2014, What Happened in 2014." [Online]. Available: http://www.mapsofworld.com/events/year-2014/. [Accessed: 09-Dec-2015].

[49] "The Biggest News Stories of 2014 - ABC News." [Online]. Available: http://abcnews.go.com/International/biggest-news-stories-2014/story?id=27466867. [Accessed: 09-Dec-2015].

[50] "The 10 Biggest International Stories of 2014 - The Atlantic." [Online]. Available: http://www.theatlantic.com/international/archive/2014/12/the-10-biggest-international-stories-of-2014/383935/. [Accessed: 09-Dec-2015].

[51] "2014 Year in Review | Facebook Newsroom." [Online]. Available: http://newsroom.fb.com/news/2014/12/2014-year-in-review/. [Accessed: 09-Dec-2015].

[52] "Facebook's most talked-about topics of 2014 - CBS News." [Online]. Available: http://www.cbsnews.com/news/facebooks-most-talked-about-topics-of-2014/. [Accessed: 09-Dec-2015].

[53] "Twitter's top tweets and retweets of 2014: Ellen, World Cup score big - TODAY.com." [Online]. Available: http://www.today.com/money/twitters-top-tweets-retweets-2014-ellen-world-cup-score-big-1D80349123. [Accessed: 09-Dec-2015].

[54] "Twitter And Facebook Launch Their 2014 'Year In Review' With Top Content, Trends & More." [Online]. Available: http://marketingland.com/twitter-facebook-launch-2014-year-review-top-content-trends-110643. [Accessed: 09-Dec-2015].

[55] "Facebook," 2011. [Online]. Available: http://www.facebook.com/.

[56] "Facebook," 2011. [Online]. Available: http://www.facebook.com/press/info.php?statistics.

[57] "Harpsocial," p. http://www.harpsocial.com/2011/04/social–medias–sh, 2011.

[58] "Twitter," p. http://ww.twitter.com/, 2011.

[59] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," *VAST 10 - IEEE Conf. Vis. Anal. Sci. Technol. 2010, Proc.*, pp. 115–122, 2010.

[60] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, 2009.

[61] A. Shrivastava and B. Pant, "Opinion extraction and classification of real time Facebook Status," *Glob. J. Comput. Sci. Technol.*, vol. 12, no. 8, 2012.

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

[63] M. R. Morris, J. Teevan, and K. Panovich, "What Do People Ask Their Social Networks, and Why?," *Chi*, vol. 69, p. 1739, 2010.

[64] A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei, "Detecting Comments on News Articles in Microblogs."

[65] C. Lin, Y. He, and R. Everson, "Sentence subjectivity detection with weakly-supervised learning," pp. 1153–1161, 2011.

[66] S. Amer-Yahia, S. Anjum, A. Ghenai, A. Siddique, S. Abbar, S. Madden, A. Marcus, and M. El-Haddad, "Maqsa: a system for social analytics on news," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 653–656.

[67] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.

[68] B. Liu, "Sentiment analysis and subjectivity," *Handb. Nat. Lang. Process.*, vol. 2, pp. 627–666, 2010.

[69] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012.

[70] A. Montoyo, P. MartíNez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments," *Decis. Support Syst.*, vol. 53, no. 4, pp. 675–679, 2012.

[71] M. Sadegh, R. Ibrahim, and Z. A. Othman, "Opinion mining and sentiment analysis: A survey," *Int. J. Comput. Technol.*, vol. 2, no. 3, pp. 171–178, 2012.

[72] P. Case and G. D. V, "Opinion Mining and Classification of User Reviews in Social Media," vol. 7782, pp. 37–41, 2014.

[73] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Min. Knowl. Discov.*, vol. 24, no. 3, pp. 478–514, 2012.

[74] L.-W. Ku, L.-Y. Lee, T.-H. Wu, and H.-H. Chen, "Major topic detection and its application to opinion summarization," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 627–628.

[75] S.-M. Kim and E. Hovy, "Automatic detection of opinion bearing words and sentences," in *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2005, pp. 61–66.

[76] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," *arXiv Prepr. arXiv1309.6202*, 2013.

[77] G. Vinodhini and R. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 6, pp. 282–292, 2012.

[78] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[79] N. Isa, M. Puteh, R. Mohamad, and H. Raja, "Sentiment Classification of Malay Newspaper Using Immune Network (SCIN)," vol. III, 2013.

[80] G. Jaganadh, "Opinion mining and Sentiment analysis CSI communication," 2012.

[81] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An On-line Lexical Database *," *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, Jan. 1990.

[82] S. Mohammad, C. Dunne, and B. Dorr, "Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 2009, pp. 599–608.

[83] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *Proc. Conf. Empir. Methods Nat. Lang. Process. July 6-7, 2002, Philadephia, Pennsylvania, USA*, pp. 79–86, 2002.

[84] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10760–10773, 2009.

[85] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.," in *LREC*, 2010, vol. 10, pp. 2200–2204.

[86] D. Das, A. K. Kolya, A. Ekbal, and S. Bandyopadhyay, "Temporal analysis of sentiment events–a visual realization and tracking," in *Computational Linguistics and Intelligent Text Processing*, Springer, 2011, pp. 417–428.

[87] G. Mishne and M. De Rijke, "MoodViews: Tools for Blog Mood Analysis.," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 153–154.

[88] G. B. Tran, M. Alrifai, I. A. Intelligence, and N. Language, "Predicting Relevant News Events for Timeline Summaries," *Www*, pp. 91–92, 2013.

[89] D. Bhattacharya and S. Ram, "Sharing news articles using 140 characters: A diffusion analysis on twitter," *Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012*, pp. 966–971, 2012.

[90] "Top 10 Arabic YouTube Channels in the Middle East | IstiZada." [Online]. Available: http://istizada.com/blog/top-10-arabic-youtube-channels/. [Accessed: 09-Dec-2015].

[91] "Al_Jazeera_English," 2015. [Online]. Available: https://en.wikipedia.org/wiki/Al_Jazeera_English. [Accessed: 06-Dec-2015].

[92] "Al_Arabiya," 2015. [Online]. Available: https://en.wikipedia.org/wiki/Al_Arabiya#cite_note-cablegatesearch1-17. [Accessed: 06-Dec-2015].

[93] "Media Use in the Middle East:An Eight-Nation Survey - NU-Q." [Online]. Available: http://www.scribd.com/doc/137906439/Media-Use-in-the-Middle-East-An-Eight-Nation-Survey-NU-Q. [Accessed: 09-Dec-2015].

[94] J. Kleinnijenhuis, F. Schultz, D. Oegema, and W. van Atteveldt, "Financial news and market panics in the age of high-frequency sentiment trading algorithms," *Journalism*, p. 1464884912468375, 2013.

[95] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Syst.*, vol. 41, pp. 89–97, Mar. 2013.

[96] J. Teevan, D. Ramage, and M. R. Morris, "# TwitterSearch: a comparison of microblog search and web search," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 35–44.

[97] "Billion-Dollar Weather and Climate Disasters: Overview | National Climatic Data Center (NCDC)." [Online]. Available: http://www.ncdc.noaa.gov/billions/. [Accessed: 03-Apr-2015].

[98] Wikipedia, "Diseases and disorders," 2015. [Online]. Available: http://en.wikipedia.org/wiki/Disease. [Accessed: 17-Apr-2015].

[99] Wikipedia, "Terrorism," 2015. [Online]. Available: http://en.wikipedia.org/wiki/Terrorism. [Accessed: 19-Apr-2015].

[100] Ask.com, "what are the effects of terrorist attacks," 2015. [Online]. Available: http://www.ask.com/. [Accessed: 01-Jan-2015].

[101] Wikipedia, "Conflict_(process)," 2014. [Online]. Available: http://en.wikipedia.org/wiki/Conflict_(process). [Accessed: 22-Dec-2014].

[102] Collinsdictionary, "plane-crash," 2015. [Online]. Available: http://www.collinsdictionary.com/dictionary/english/plane-crash. [Accessed: 01-Jan-2015].

[103] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556–1560.

[104] R. Plutchik, "A general psychoevolutionary theory of emotion," *Theor. Emot.*, vol. 1, 1980.

[105] "List of Human Emotions - List of Human Emotions." [Online]. Available: http://www.listofhumanemotions.com/listofhumanemotions. [Accessed: 21-May-2015].

[106] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 37–45.

[107] A. Kontostathis, L. M. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps, "A survey of emerging trend detection in textual data mining," in *Survey of Text Mining*, Springer, 2004, pp. 185–224.

[108] "Poll results: Top languages for analytics/data mining programming." [Online]. Available: http://www.kdnuggets.com/2012/08/poll-analytics-data-mining-programming-languages.html. [Accessed: 23-Dec-2015].

[109] "RapidMiner at CeBIT 2010: the Enterprise Edition, Rapid-I and Cloud Mining - Data Mining - Blog.com." [Online]. Available: http://www.data-mining-blog.com/cloud-mining/rapidminer-cebit-2010/. [Accessed: 09-Dec-2015].

[110] "RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series): Markus Hofmann, Ralf Klinkenberg: 9781482205497: Amazon.com: Books." [Online]. Available: http://www.amazon.com/RapidMiner-Analytics-Applications-Knowledge-Discovery/dp/1482205491. [Accessed: 09-Dec-2015].

[111] "KDnuggets Annual Software Poll:RapidMiner and R vie for first place." [Online]. Available: http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html. [Accessed: 09-Dec-2015].

[112] "Rexer Analytics 5th Annual Data Miner Survey - 2011." [Online]. Available: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2011.html. [Accessed: 09-Dec-2015].

[113] "German Predictive Analytics Startup Rapid-I Rebrands As RapidMiner, Takes $5M From Open Ocean, Earlybird To Tackle The U.S. Market." [Online]. Available: http://techcrunch.com/2013/11/04/german-predictive-analytics-startup-rapid-i-rebrands-as-rapidminer-takes-5m-from-open-ocean-earlybird-to-tackle-the-u-s-market/. [Accessed: 09-Dec-2015].

[114] F. Ben Abdesslem, I. Parris, and T. Henderson, "Reliable online social network data collection," in *Computational Social Networks*, Springer, 2012, pp. 183–210.

[115] B. Rieder, "Studying Facebook via data extraction: the Netvizz application," in *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, pp. 346–355.

[116] C. Troussas, M. Virvou, K. Junshean Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," in *Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on*, 2013, pp. 1–6.

[117] A. Shrivatava and B. Pant, "Opinion Extraction and Classification of Real Time Facebook Status," vol. 12, no. 8, 2012.

[118] R. Rogers, "The end of the virtual," 2009.

[119] C. Cesarano, B. Dorr, A. Picariello, D. Reforgiato, A. Sagoff, and V. Subrahmanian, "Oasys: An opinion analysis system," in *AAAI spring symposium on computational approaches to analyzing weblogs*, 2004.

[120] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[121] "TheySay PreCeive API Demo." [Online]. Available: http://apidemo.theysay.io/. [Accessed: 10-Dec-2015].

[122] "Python NLTK Sentiment Analysis with Text Classification Demo." [Online]. Available: http://text-processing.com/demo/sentiment/. [Accessed: 10-Dec-2015].

[123] "DepecheMood - Try Our Online Demo!" [Online]. Available: http://www.depechemood.eu/DepecheMood.html. [Accessed: 10-Dec-2015].

[124] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007, pp. 70–74.

[125] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[126] D. Meyer, K. Hornik, and I. Feinerer, "Text mining infrastructure in R," *J. Stat. Softw.*, vol. 25, no. 5, pp. 1–54, 2008.

[127] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[128] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[129] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.

[130] Y. Mejova and P. Srinivasan, "Exploring Feature Definition and Selection for Sentiment Classifiers," pp. 546–549, 2011.

[131] S. Li and C. Zong, "A new approach to feature selection for text categorization," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, 2005, pp. 626–630.

[132] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 80–89, 2004.

[133] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguist.*, vol. 37, no. 2, pp. 267–307, 2011.

[134] J. S. Kessler and N. Nicolov, "Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations.," in *ICWSM*, 2009.

[135] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Min. Knowl. Discov.*, vol. 7.4, pp. 373–397.

[136] K. Balog, G. Mishne, and M. De Rijke, "Why are they excited?: identifying and explaining spikes in blog mood levels," in *Proceedings of the Eleventh*

Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations, 2006, pp. 207–210.

[137] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[138] D. Maynard and A. Funk, "Automatic detection of political opinions in Tweets," pp. 81–92.

[139] B. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," in *9th. IT & T Conference*, 2009, p. 13.

[140] P. Chaovalit and L. Zhou, "Movie review mining: A comparison between supervised and unsupervised classification approaches," in *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, 2005, p. 112c–112c.

[141] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[142] K. T. Durant and M. D. Smith, "Mining sentiment classification from political web logs," in *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA*, 2006.

[143] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Proj. Report, Stanford*, pp. 1–12, 2009.

[144] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing- style features and classification techniques," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, 2006.

[145] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6527–6535, 2009.

[146] H. Cui, V. Mittal, and M. Datar, "Comparative experiments on sentiment classification for online product reviews," in *AAAI*, 2006, vol. 6, pp. 1265–1270.

[147] E. Airoldi, X. Bai, and R. Padman, "Markov blankets and meta-heuristics search: Sentiment extraction from unstructured texts," in *Advances in Web Mining and Web Usage Analysis*, Springer, 2006, pp. 167–187.

[148] B. Xu, T.-J. Zhao, D.-Q. Zheng, and S.-Y. Wang, "Product features mining based on conditional random fields model," in *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, 2010, vol. 6, pp. 3353–3357.

[149] D. K. Kirange and R. R. Deshmukh, "Emotion Classification of News Headlines Using Svm," vol. 5, pp. 104–106, 2012.

[150] N. V Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[151] Y. Y. Yang, M. Mahfouf, G. Panoutsos, Q. Zhang, and S. Thornton, "Adaptive neural-fuzzy inference system for classification of rail quality data with bootstrapping-based over-sampling," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, 2011, pp. 2205–2212.

[152] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.

[153] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *ICML*, 1999, vol. 99, pp. 258–267.

[154] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004.

[155] M. Porter, "The Porter stemming algorithm, 2005," *See http//www. tartarus. org/~ martin/PorterStemmer*.

[156] W. B. Frakes, "Information Retrieval: CHAPTER 8: STEMMING ALGORITHMS." [Online]. Available: http://dns.uls.cl/~ej/daa_08/Algoritmos/books/book5/chap08.htm. [Accessed: 17-Dec-2015].