Mining and Analysing Social Network in the Oil Business: Twitter Sentiment Analysis and Prediction Approaches

Hanaa Ali Aldahawi

2015

Cardiff University

School of Computer Science and Informatics

A thesis submitted in partial fulfilment of the

requirement for the degree of Doctor of Philosophy

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

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

Signed ..... (candidate)

Date .....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed ..... (candidate)

Date .....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

To my parents

With profound gratitude to Mum - I hope I made you proud.
In loving memory of Dad; you would have been proud of me.

Abstract

Twitter is a rich source of data for opinion mining and sentiment analysis that companies can use to improve their strategy with the public and stakeholders. However, extracting and analysing information from unstructured text remains a hard task. The aim of this research is to investigate the use of Twitter by "controversial" companies and other users. In particular, it looks at the nature of positive and negative sentiment towards oil companies and shows how this relates to cultural effects and the network structure. This has required the evaluation of existing automated methods for sentiment analysis and the development of improved methods based on user classification. The research showed that tweets about oil companies were noisy enough to affect the accuracy. In this thesis, we analysed data collected from Twitter and investigated the variance that arises from using an automated sentiment analysis tool versus crowd sourced human classification. Our particular interest lay in understanding how users' motivation to post messages affected the accuracy of sentiment polarity. The dataset used Tweets originating from two of the world's leading oil companies, BP America and Saudi Aramco, and other users that follow and mention them, representing Western and Middle Eastern countries respectively. Our results show that the two methods yield significantly different positive, natural and negative classifications depending on culture and the relationship of the poster of the tweet to the two companies. This motivated the investigation of the relationship between sentiment and user groups extracted by applying machine learning classifiers. Finally, clustering based on similarities in the network structure was used to connect user groups, and a novel technique to improve the sentiment accuracy was proposed. The analytical technique used here provided structured and valuable information for oil companies and has applications to other controversial domains.

Acknowledgements

I would like to express my gratitude to my supervisors, Dr Stuart Allen and Professor Roger Whitaker for their support throughout this research. In particular, I would like to acknowledge my debt to Dr Allen, my main supervisor, and express to him my special appreciation and thanks. He has been a tremendous mentor for me, with his patience and knowledge. His guidance, advice and continual encouragement at all stages of my PhD helped me tackle the challenges of this research and shaped my ideas, keeping me on track and helping me become an independent researcher. I am much blessed at being under his supervision and I have learned a lot from him.

I would like to extend my thanks to my sponsor in Saudi Arabia, King Abdul-Aziz University (KAU), for the scholarship and the continuous support throughout the years of my study in the UK. I am also thankful to the head of the Information Science Department and all the staff for their support, encouragement and friendship. Special thanks and appreciation also go to the UK Saudi Arabian Cultural Bureau for their help and support.

I would like, too, to thank Dr Martin Chorley for his continuously helpful crowdsourcing expertise, which is an important part of this thesis. I am also grateful to Dr Matt Williams for his help in learning Python language in the early stages of this research. Thanks also to all members of the School of Computer Science and Informatics for their helpful discussions, comments, feedback, events and facilities. Special thanks to Dr Rob Davies, and Mrs Helen Williams for technical and administrative support.

I am deeply grateful to my family who always encourage and support me. To my mum, who provided endless support, encouragements and love, to make me who I am today. Words cannot express how grateful I am to her for all the sacrifices that she has made for my sake. Her prayer for me has been what has sustained me thus far. To my brother Mohammed who was a good companion during all my study period in the UK and was very supportive and caring. To my sisters Wafaa, Rajaa and Asmaa, who believed in me. To my niece Hala and my nephews Ziyad and Yossef for their love. I am also thankful to my bigger family, my aunts and uncles, particularly uncle Mohammed and aunt Saleha. They all deserve my utmost thanks.

My deep gratitude goes to my friends. To Liqaa Nwaf for her support, kindness and care during the difficult period of my PhD. I am very lucky to have found a big-hearted person like Liqaa to be my close friend. To Fatima Alrayes for her support and caring. To Shada Alsalamah and Haya Almagwashi for their support, feedback, and precious friendship. To all my colleagues who have been positive and supportive during the PhD journey.

Finally, I am extremely grateful to all those who made my study journey easier and finally successful.

Contents

List of Figures

List of Tables

List of Acronyms

AMT - Amazon Mechanical Turk

HIT - Human Intelligent Tasks

RT - Retweet

BBS - Bulletin Board Systems

URL - Uniform Resource Locator

ROC - Receiver Operating Characteristic

MLP - MultiLayer Perceptron

CCT - Correctly Classified Tweets

NLP - Natural Language Processing

NT - Number of Tweets

TF - Term Frequency

CT - Characterization Toolkit

SVM - Support Vector Machine

KNN - K-Nearest Neighbors

NB - Naïve Bayes

DT - Decision Tree

ROC - Receiver Operating Characteristic

List of Publications

The work in this thesis has contributed to the following refereed publications:

[1] H. Aldahawi and S. Allen, "An approach to tweets categorization by using machine learning classifiers in oil business," in *Computational Linguistics and Intelligent Text Processing*, pp. 535–546, Springer, 2015

[2] H. Aldahawi and S. M. Allen, "Twitter mining in the oil business: A sentiment analysis approach," in *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pp. 581–586, IEEE, 2013

This thesis also contributed to the following refereed paper whose workshop proceedings were not published:

[3] H. Aldahawi and S. M. Allen, "Analysing cultural effects of social network usage in business," in *8th UK Social Network Analysis Conference. UKSNA*, 2012

Chapter 1

Introduction

Social networking has dramatically changed our lives and the way that we interact with the world [4]. Recent research shows that millions of people are using social network services such as Facebook and Twitter for various purposes, such as finding and sharing information, making friends and entertaining themselves [5]. A social network is a web-based utility which allows people to register under an identity so as to make new friends, socialize with old friends, and post news, pictures and videos using a central location [6, 7]. As a rule, new users connect with their friends and other friends whom these friends recommend. However, as their experience grows, they make friends with quite unknown people who have similar interests. This can be observed in both Facebook and Twitter. In Twitter, new users follow popular users and their friends, before gaining enough confidence to choose others more carefully on the basis of similar interests. These social networks are managed by international companies who generate their revenues through paid advertisements [8]. A major advantage of social networks is that they can tell us other users' opinions and the enormous amount of data present on the social sites can be useful for purposes such as analysing the current suitation and predicting the future [9]. There are multiple ways through which a user can interact with others on social networks, such as chatting, messaging, audio, video, voice and file sharing. Today all the activities that are conducted on special online services are included in social network platform. These services provide software and tools that allow users to build a network of people [6]. Recently, research into social networks has been carried out on data collected from online platforms such as Facebook, Twitter, LinkedIn and Flickr in different sectors such as business, education, politics and medicine.The capacity to collect such data is very important in social network research and has led to the appearance of the new field of "computational social science" [10, 11].

The use of social networks in business is an area which considers the communication between humans in order to create long-term relationships between companies and their customers, build trust between these two and keep people informed when new events, products or services are offered in the market [12]. It uses two-way communications that target suitable listeners among the wider audience. Social networks in business can have great influence on the marketing and success of corresponding businesses [13]. The use of social networks in marketing and busi-

nesses will help consumers to get to know about any relevant information about products and services [14]. The media help network members to share their ideas and trusted information about different products or services. When businesses use social networks, whatever information presented is not necessarily interesting to everyone, but it is considered important that people should learn about all the products and services offered by these businesses [13, 14]. If one interested customer responds, then it may encourage more customers to respond similarly [14]. If nobody responds at first, it does not mean that the application is wrong; it could be a matter of time, until someone respond positively. Social networks today exhibit the principle of "word of mouth at the speed of light" [15], in which a bad experience can be propagated more quickly than a good experience. Social networks are, therefore, the effective media for enhancing business performance and any company that sells business-to-business services is strongly recommended to use social networks [16]. Twitter, introduced in October 2006, is one of the most popular social network sites. It is a micro-blogging site, which allows the users to post or receive instant messages up to 140 characters long. These messages can be viewed or composed on Twitter websites, smartphones or tablets, as well as in applications such as Facebook, TweetDeck or a web page aggregator [15]. An example of the positive effect of using social network channels such as Twitter, Facebook, YouTube and Flickr in business is studied in [17]. This study observed and analysed how an oil company, BP, used the above channels to implement a new strategy (following the theory of image restoration discourse) to restore their image in the wake of the 2010 Gulf Coast oil spill. Micro-blogging in particular helps business owners to share intensive information more quickly with the public about their products and services. Customer satisfaction can also be surveyed using social networks as it demonstrates a company's commitments to its consumers to improve the quality of service offered [16]. Conversely, social networks offer consumers an opportunity for both negative and positive feedback and comments, which lead companies to consider ways of dealing with their consumers [16].

1.1 Research Problem and Motivation

Mining information from text is a very important research area which aims to discover and extract knowledge from unstructured text. In this era of Web 2.0, text is the most common vehicle for the formal and informal exchange of information, but automating the extraction of meaningful information from the text is still difficult. The huge number of unstructured data generated by social networks are called big data. Gartner IT glossary defines big data as"high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [18]. Many researchers tend to focus on automated methods across large datasets and consequently

miss the subtleties that may nevertheless be significant. Other people have found problems with big data but in this work problems arise from little data. This research investigates the sentiment issues within individual tweets and tries to resolve them, finally moving from little data to big data by improving the sentiment in groups of tweets clustered together, but not individual tweets.

The aim of this research is to investigate the use of Twitter by 'controversial' companies and other users. In particular, it looks at the nature of positive and negative sentiment towards oil companies and asks how this relates to the cultural effects of the network structure. This has required an evaluation of the current automated tools for sentiment analysis, and the development of improved methods based on user classification. Machine learning algorithms are used, both supervised and unsupervised, for learning and prediction purposes. In this research text mining method was used for learning the diverse emotions/sentiments of different groups of users from their tweets/postings scraped from Twitter. One purpose was to predict these groups and investigate the role of groups in the nurturing of sentiment. Additionally, with the help of these intelligent learning algorithms it was a motivation to show the relationship between user groups by clustering them and evaluating the sentiments in different clusters. To improve the accuracy of sentiment prediction a novel technique was proposed.

1.2 Hypothesis and Research Questions

This research proposes the following main hypothesis to be examined:

Automated Twitter analysis tools can be a reliable and effective means of data interpretations for companies to make proactive or reactive decisions regarding communications with their stakeholders, provided they take account of their local cultural environment and personal motivation.

In order to verify this hypothesis, a number of important research questions were addressed.

1. Tweets Sentiment Analysis

   RQ1. Are automated sentiment analysis tools a reliable and effective means of data interpretation for companies?

   (a) Are automated sentiment analysis tools suitable for all companies' tweets and mentions?

   (b) Is there any need for manual sentiment analysis for companies' tweets and mentions?

(c) How can automated and manual tools be combined to get accurate results efficiently?

2. Twitter Users Classification (Supervised Learning Techniques)

   RQ2: Can the type of users/followers be used to improve automated sentiment analysis?

   (a) How accurately can the type of users (e.g. media, government or non-government organisations, environmentalist, politicians, business analysts, oil company employee and general public) be predicted by using machine learning classifiers with simple features?

3. Twitter Users Clustering (Unsupervised Learning Techniques)

   RQ3. How do the clusters in the network structure of friends and followers relate to negative and positive sentiment?

   (a) Is the accuracy of automated sentiment analysis related to the structure/ clustering of users?

   (b) Can a closely connected group of users and a contrasting group be easily observed within each cluster?

4. Cultural Differences

   RQ4. Is there any difference between oil companies in the Middle East and Western countries in the structure of users, sentiment accuracy and the quality of prediction?

1.3 Case Study

In order to achieve the research aim and address the research questions, two oil companies, British Petroleum (BP) [1] based in America (Western), and Saudi Aramco [2] based in Saudi Arabia (Middle East). Oil industry was a preferred case study because of the controversial nature of their business. They often face more backlash and negative public comments in the community in which they operate. Both companies have active Twitter accounts and the rationale for choosing these companies is based on factors such as: similar business size, industry, objectives and the fact that they are in culturally distinct countries. Twitter was a preferred choice of social media platform because these companies are more active on Twitter than others. Twitter data also makes it easier to accurately categorize the sources and targets of their activities on social

[1] `https://www.bp.com/`
[2] `http://www.saudiaramco.com/en/home.html`

media. All data used in this work were solely sourced from Twitter. It is also noteworthy to mention that these companies were personally contacted for information about their communication strategy via social media that could aid this research but they both decline to provide any useful information for security purposes. Saudi Aramco as a company have only one official Twitter account (@Saudi_Aramco) and BP has different Twitter accounts dedicated to each of the countries where it operates but (@BP_America) account being the most active of them all.

1.4 Research Contributions

The contribution of this research lies in building a predictive model and discovering relationships between tweets posted by people belonging to different categories; by extracting sentiment information and applying supervised and unsupervised machine learning techniques. In answering the research questions, the main contributions made in this research work are outlined below:

- Analysis of public sentiment towards oil companies on social media and investigation of the limitations of such analysis.
- Investigation of the role of groups based on the accuracy of the sentiment - multi-text categorization models are used for categorizing incoming tweets automatically into a number of pre-defined classes.
- Proposal of a novel technique to improve the accuracy of automated sentiment prediction using clustering, multi-class clustering models of tweets, used by applying hard clustering algorithm.

1.5 Thesis Structure

This thesis is organized as follows: Chapter 2 gives an overview of social networking in general in business and Twitter analysis in particular, reviewing the relevant literature across the range of topics addressed in the thesis.

Chapter 3 describes the collection datasets used in this work. It also provides primary analysis about Twitter and the features of its use by oil companies.

Chapter 4 presents an evaluation of automated sentiment analysis for oil companies on Twitter and discusses the initial results.

Chapter 5 proposes the use of machine learning algorithms to group the tweets according to different types of user.

Chapter 6 introduces a clustering algorithm that can be used to find the groups in the given tweets in an unsupervised way. The accuracy measures of the sentiment prediction within different clusters is also presented.

Chapter 7 the final contributory chapter, concludes the thesis by underlining the major contributions and suggesting some directions for future research.

Chapter 2

Background and Literature Review

Introduction

This chapter serves three main purposes. First, it provides a general overview of social networks and a background to the techniques used in social network analysis. Second, it investigates the use of Twitter analysis by businesses. Finally, various Twitter and related data analysis techniques are discussed. Social networks can be broadly divided into those that focus on the content (content-centric) and those that focus on user profiles (user-centric). Twitter, according to this means of categorization, falls into both categories. The present analysis of Twitter content by companies focuses on its benefits and the influence of cultural differences on such analysis. It has been found that most of the tweets posted by users are positive, negative or neutral with respect to situations, circumstances, products or brands. These tweets are usually mined and analysed to determine the orientation of the users, who may be customers, stakeholders or the general public. This process is called *sentiment analysis*. The importance and problems identified with this process in Twitter are examined in the present chapter. The use of supervised learning techniques (classification) and unsupervised learning ones (clustering) in Twitter are also examined. Several algorithms of each are discussed and their uses in business are highlighted.

Section 2.1 presents an overview of definitions and a brief history of social network. In Section 2.2 Twitter analysis in business is briefly discussed. The benefits of social networks in business and the impact of cultural differences on social networks are summarised in this section. This chapter concludes with the discussion of Twitter and data analysis techniques in Section 2.3. Sentiment analysis is important in data analysis; therefore, its importance and problems are discussed with respect to Twitter. Classifications in Twitter, a supervised learning technique, with respect to its applications in business, are outlined. Various machine learning classifiers are also briefly examined. Clustering in Twitter, an unsupervised learning technique, is discussed. A summary of various clustering algorithms is presented. This chapter concludes with a report on precision, recall and the F-measure as evaluation measures related to this project in section

2.4.

2.1 Social Network Background

The evolutionary process of "*social networks*" can be traced back to the late 1970s when Bulletin Board Systems (BBS) were in vogue. Popular platforms for BBS included vBulletin and PhP BB. These BBS allowed data and software to be uploaded and downloaded; reading news and bulletins and exchanging messages with other users through emails and public message boards were also possible. Forums were the direct descendants of BBS and they in turn played an important role in the evolution of social networks [19]. Forums, like public message boards, allow people to hold conversations on a topic, or *thread*. Unlike BBS, forums are widely available since the advent of inexpensive dial-up modems and modern mosaic browsers. Social networks are built on the foundations of BBS and forums, to which they add features such as multi-user chatting; selecting some of the users as friends or followings, as in Twitter; and creating or joining groups of interest, among others. **Six Degree**, which was launched in 1997, is generally accepted as the first form of modern social media. In 2002, **Friendster** was launched for the US public while **MySpace** and **LinkedIn** were launched in 2003. **Facebook** and **Twitter** became available to users in 2004 and 2006 respectively. A large amount of data can now be found on social media because of the increasing number of users. The available data on these social networks is of great importance when mined and used for such purposes as analysis and prediction.

2.1.1 Definition

Although social networks have become an integral part of our everyday lives, it is important to give a formal definition that that encapsulate them all. A social network can be defined as "*a group of internet-based applications that build on the ideological and technological foundations of web 2.0 and allow the creation and exchange of user-generated content*" [5]. They offer a web-based service that helps users to create their own profile within a bounded system [6]. Safko [15] views social networks as a set of tools and a technology that allows people to connect with others and build relationships more efficiently. In the view of Hogan [20], a social network can be defined through its network - a set of nodes linked with each other by relations. These nodes can be individuals, groups of individuals or web pages. Hogan [20] observes that these networks can be grouped into whole, personal and partial networks. A whole network is used to define the relationships between people within a limited population, where useful active data are collected in order to make a subjective assessment of individuals' views and to examine their

behaviours. A personal network, however, is used to compare the characteristics of people, such as the size, shape and quality of their personal networks; while a partial network is recognised as an efficient solution for data collection, as shown in Figure 2.1.

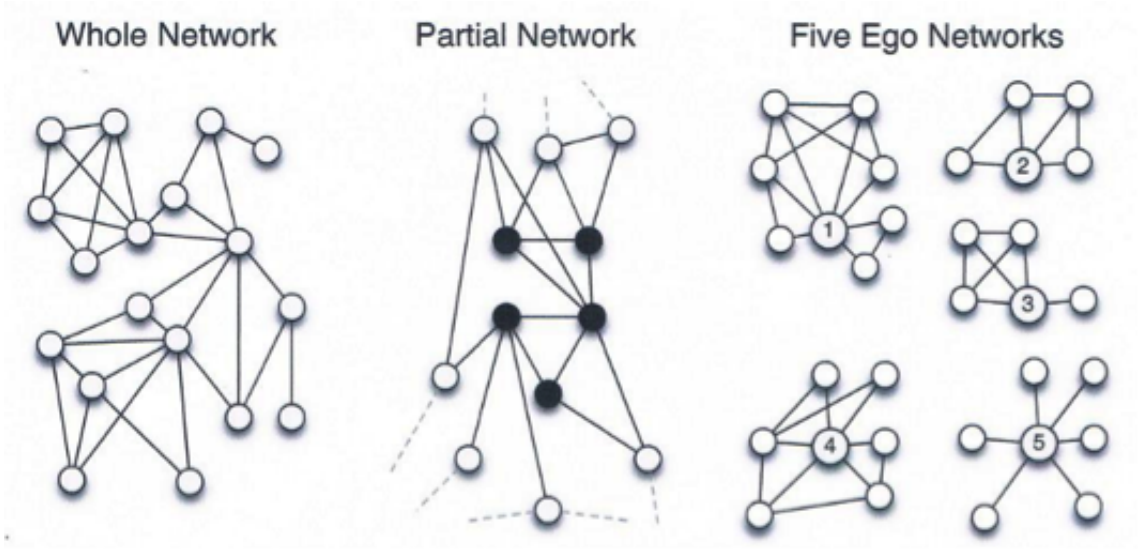


Figure 2.1: Whole, partial and personal network types [20]

From another perspective, social networks can be seen as large-group interventions. In the popular micro-blogging social network, Twitter, the features can be related closely to the characteristics of a large-group intervention [21] as shown in Table 2.1:

Table 2.1: Similarity between large-group interventions and corresponding Twitter features

Large-Group Interventions	**Corresponding Twitter Features**
Comprehending the requirement for change (such as business expansion, training needs etc.).	Making decisions based on the analysed opinionated tweets of customers.
Inspecting the existing facts and determining what should be changed (in the case of previous ineffective training sessions).	Predicting the consequences of business actions through the sentiment analysis of customers' tweets.
Effecting a change to current procedures (such as hiring a new trainer with the required expertise).	Changing or continuing business actions based on the desired outcomes predicted.
Applying and encouraging change and ensuring that it works (re-tweets, circulating tweets).	Carrying out the business actions and analysing the feedback generated.

Although large-group intervention exercises are growing, many analysts are of the opinion that the theory supporting them is not effectively articulated. Large-group intervention is "*a structured process for engaging large numbers of people to enhance the amount of relevant information brought to bear on a problem, to build commitment to problem definitions and solutions,*

to fuse planning and implementation, and to shorten the amount of time needed to conceive and execute major policies, programs, services or projects" [22].

Manning and Binzgar [23] report that most large-group intervention techniques involve a great number of individuals representing all organizational stakeholder groups that get together in a 2 or 3-day conference to evaluate data, make informed decisions and produce action plans. They also observe that the effort of large-interventions was often formally earmarked for small groups at senior levels in the organizational chain of command. Holman and Devane [21] highlighted six conditions that were they employed to determine large-group interventions. Each method of doing so had to:

i involve individuals in an important way;

ii find out and generate contributed assumptions;

iii come with fundamental exploration;

iv continue to be used for a minimum of 5 years;

v offer a methodological procedure for change; and

vi be competent at producing remarkable benefits from a modest level of resources.

2.1.2 Types of Social Network

Generally, social networks can be divided into content-centric and user-centric platforms.

1. **Content-centric**: These are informational platforms where users post on a variety of topics. These posts can be commented on or forwarded by other people. Usually, the comments add more information than the original posts held. Popular examples of a content-centric platform are blogs such as Twitter, Tumblr and WordPress. Twitter is a microblog that allows users to interact and communicate with others. Users are able to broadcast real-time messages called tweets of up to 140 characters in length. These tweets can take the form of texts, pictures or videos in which people express their opinions, inform others of breaking news, activities or events. These tweets can be used for a variety of research in different fields [24]. Other examples of a content-centric platform include YouTube and Flickr, where users share images, videos and discuss contents posted by other people. On Last.fm (a music website), users listen to audio songs and receive suggestions on related tracks, based on the choices made by previous users [25].

2. **User-centric**: These platforms emphasise the identity of the users, supplying profile and personal interests of each [12]. Users can update their status, share pictures, videos and links. A popular example of this type of social network is Facebook. It is estimated that 57% of online American adults use Facebook and 64% of these adults check their profile on a daily basis [26]. Similar platforms include MySpace and Bebo. Another method of representing users as they are is by depicting their appearance and personality as avatars in a virtual world [27]. An example of this is "second life". Users can interact through voice chats and messages. These user-centric platforms can be used to select candidates of interest, conduct interviews and recruit specific people by checking or sorting users' profiles. This method of recruitment is more cost-effective than the conventional recruitment process [28].

Twitter, both a user-centric and content-centric platform, has been chosen as the social medium to be analysed in the present work, since its data are publicly available and each tweet is small enough to crawl, store and process easily. People share news and opinions in real time on Twitter. Petrovic et al. [29] observe that Twitter communicates news faster than a conventional news system. They analysed tweets and news articles over a 77-day period and found that local and minor events were reported in more detail in Twitter than in other news sources. Although the increasingly large data due to its characteristics of speedily updating and disseminating information provides a more accurate view of users in sentiment analysis, the latter can be a difficult task. As David Ediger et al. [30] observe, analysis of Twitter data poses a challenge both from the hardware and software points of view, because of their abundance. These writers used the Graph Characterization Toolkit (CT) to analyse a cloud of conversations and ranked famous personalities such as actors and actresses from their analysis.

2.2 Twitter Analysis in Business

Organizations can engage in online dialogues to enhance their prosperity by using social network platforms such as Twitter. Lovejoy and Saxton [31] show how non-profit organizations are using micro-blogging applications to analyse their business activities; they classify these organizations into three categories: information, community and action. Waters and Jamal [32] identify four models of public relations by which non-profit organizations communicate through Twitter: public information, press agentry and both one-way and two-way communication. The study concludes that these non-profit organizations are more likely to use one-way communication than the more rewarding two-way communication. The creation and management of internal and external relational networks are critical factors in the success of innovative and small companies. Relational networks represent the aggregation of interactions through

membership in official organizations and relational encounters that business owners create and nurture with providers, distributors, consultants and customers, or some of the wide range of additional social contacts including friends, loved ones and acquaintances [33].

In a study conducted by Barnes and Andonian [34], it was observed that the top 200 of the Fortune 500 companies on the 2011 listing had an active Twitter account - being active was defined as having had an update in the past thirty days. An extended study was made by Rybalko & Seltzer [35] in order to investigate how the 500 richest companies benefited from building an online relationship with the stockholders when they started using popular social networks such as Twitter. A dialogue can be fostered between the company and the public through social networks, in order to present useful information about the organizations and to arrange dialogic loops and continuous online customers care. Wigley & Lewis [36] examine how the engagement between the companies and the stakeholders can be made through Twitter. It shows that positive Tweet were received from the highly engaged companies in particular when the company holds a practised dialogical communication. On the other hand, less engaged companies can receive more negative tweets because of lack of communication. Jansen et al. [37] examine two approaches related to micro-blogging: the first one is companies' use of micro-blogging as electronic 'word-of-mouth' in order to share consumers' opinions and concerns about products, services or brands. The second one examines the overall structure of a micro-blog posting, the types of expressions and the movement in positive and negative sentiment. It has been observed that companies can use micro-blogging in order to explore their branding strategy.

Web communication and social network activities and trends are closely related to the concept of co-creation as a marketing and business activity. Co-creation focuses on the generation and consistent realization of shared company-consumer value. Because co-creation is generally assumed to be a personal activity, little is known about the collaborative customer participation process and the implications for the customer and the brand of the combined effort of social networks and co-creation [38].

2.2.1 The Benefits of Social Networks to Business

Social networks offer many benefits to businesses which know how to mine and process vast amounts of available data on the social media. They thus obtain useful information which can help in decision-making and learning the required actions. A few of these benefits are discussed below.

2.2.1.1 Analytical resources

Companies use social networks to analyse the efficiency and effectiveness with which they meet their customers' needs [39].This in turn influences their overall strategy. The feedback of customers can be consulted to learn their expectations and to monitor whether these are being met. Companies can also create awareness of new products or services and monitor the acceptability rate of these innovations. This is particularly useful during the development phase [34], since it makes customers feel that they are part of the developments if their concerns and suggestions are appropriately addressed. This also increases the acceptability rate. For example, companies can monitor the buying patterns of customers or groups of customers to enable them to discover which products - their key products - occupy the largest percentage of their sales and the customers or group of customers - key customers - who contribute the largest percentage to their revenues. This enables the companies to target their marketing activities towards the key customers and focus on their key products. This task is mostly handled by social business analysts [34].

2.2.1.2 Crises management and reputation protection

Paul [40] suggests that the rapid adoption of new media has led to a shift in the practice of crisis management. According to Wright and Hinson [41] "*blogging and other aspects of social media have the potential to bring about dramatic changes to many aspects of public relations*". In a study carried out by Wigley and Lewis [36] a company in crisis can keep its reputation if it deeply engages its stakeholders and the public through two-way communication. When crises ensue, companies need to explain to the stakeholders and the general public what the root causes are and what current actions they themselves are taking to correct and prevent any repetition of the occurrence. This can be done through the conventional news media - print, radio and TV - and the social media. It has been found that social media 'travels' faster than the conventional news media and is more economical. During the 2010 oil spill in the Gulf of Mexico which affected the coasts of Louisiana, Mississippi, Alabama and Florida - triggered by an explosion on a platform operated by British Petroleum (BP) - BP did not have a Twitter presence. It thus could not meet the demand for an explanation through the conventional social media that it adopted. This led to the severe damage of BP's corporate image [17]. The lack of presence in the social media did not allow BP to respond to stakeholders and other people's concerns in real time. This had a negative effect on their crisis management. But a new study by Watson [42] compares the coverage of the BP oil spill by the Gulf Coast journalists and the Twitter users. The study found that the journalists' and Twitter users' attitudes to the crises was similar but their thematic frames differed. While the Twitter users used a thematic frame

and focused more on the government's role, the journalists were found less likely to use a thematic frame and focused on BP's role. This analysis reveals that Twitter represented a good alternative medium. American Airlines in 2008 were confronted with a crisis brought about by numerous delayed flights. In an attempt to manage the crisis, the airline issued a press release informing readers how to access the advisory section on its official website; sent updated emails to its customers and compensated stranded passengers. They claimed to have been monitoring people's responses on social media but they failed to engage people whose posts were either condemnations or commendations. The consequent of the cancellations was estimated to be $15 million per day [43]. Still, their strong online presence helped them to respond to and reassure their customers in real time and this had a positive effect on their crisis management.

2.2.1.3 Business-customer relationship

Social networks have improved the communication between businesses and their customers. Businesses can now track comments on their products and services and respond to them [16]. This has promoted two-way communication and creates a sense of involvement among the customers, which can pose an advantage but also a disadvantage. It will be an advantage if companies can quickly and honestly respond to a customer's concerns and a disadvantage if they fail to do so [44]. A company which has developed a reputation for quick responses to its followers can manage the shift from conventional customer service through phone conversations to online interactions through Twitter and other social networks. Some companies have taken the initiative to solicit their customers' feedbacks in order to obtain objective and quantitative evaluation of the quality of products or services being rendered. This customer feedback can be mined and evaluated further. Their sentiments with regard to the products or services of the companies can be classified as positive, negative or neutral. The consequence of evaluating these customers' data is effective decision-making, such as the customization of products and services to suit various customers' needs.

2.2.1.4 New approach to marketing

Tweets reach millions of users quicker than conventional media messages do [45]. This particular attribute has been harnessed by many companies to launch campaigns for products or services, and to introduce new products or services. Customers' anticipation and expectations of a new product can be created by this means. Another advantage of social network marketing is its comparatively low cost and quicker results. An example is Blendtec which created a cheap YouTube campaign that had quickly increased its sales to five times what it had been before the campaign [46]. Automated sentiment analysis on social networks can help to monitor the

effect of a marketing campaign in real time, on Twitter in particular. This is consistent with the experiments performed in this research where users' classification and clustering improve the automated sentiment analysis.

2.2.2 The Impact of Cultural Differences

Differences among corporate and public posts on social networks indicate that culture plays a significant role in shaping the dialogue between organizations and the public in different countries. This implies that the perspectives of different individuals are influenced by their culture. By using a cross-cultural perspective, Men and Tsai [39] have advanced the understanding of relationship cultivation on social networks. Their study examines how companies use popular social network sites to facilitate dialogues with the public in two countries with diverse cultures: China and the United States. The study demonstrates that while the specific tactics - relationship cultivation strategies; having corporate posts and customers posts on corporate pages - vary across the two markets, the companies in both countries have recognized the importance of social networks in relationship development and have employed the appropriate online strategies (disclosure, information dissemination, interactivity and involvement). Furthermore, the ways of and reasons for using social network sites are different, depending upon the social and cultural milieu [47, 48]. These studies examine the mixed and complex nature of social influence to understand how the cultural context shapes the use of social networks by considering the weight of motivation in different countries. This study shows that a single type of dialogue does not apply across cultural boundaries. Companies' messages must adapt to different target cultures to produce efficient results. In this research, it is important to understand the cultural background of the companies' target community which reflects a deeper understanding of the Twitter users' sentiment and the user network structure.

2.3 Twitter and Data Analysis Techniques

It has been estimated that more than 316 million per month active Twitter users post approximately 500 million tweets per day in 35 languages [49]. This quantity of data contains useful information when gathered and analysed. Various techniques of Twitter data analysis, both manual and automatic, have been proposed by different authors, a few of which are discussed in this section.

2.3.1 Network Structure

Twitter can be classified as a user-centric platform because it possesses "*high degree assortativity, small shortest path lengths, large connected components, high clustering coefficients, and a high degree of reciprocity*" [50]. Degree assortativity measures the preference for similar vertices of a graph to attach to one another [51]. Each path length is the number of movements along the edges required from one user to reach another. Therefore, shortest path length measures the closeness of users' connection [50]. Clustering coefficients are a measure of the fraction of users whose friends are friends of one another [52]. In this case, users are followed by other users because of who they are or of their offline relationships such as following a cousin or a colleague. However, it tends to be content-centric when the *'follow'* attribute of Twitter is considered. A user follows a company primarily to receive news about its products and services.

A basic attribute that can be used to differentiate user-centric platforms and content-centric platforms is the degree assortativity a measure of the closeness of a graph's virtues [50]. In most cases, the assortative measure for user-centric platforms is typically between 0.1 and 0.4 [53]. For example, a measure of Facebook's assortativity is 0.226. In their analysis of the Twitter follow graph, Myers et al. [50] observe that Twitter exhibits the characteristics of being user-centric in some ways and the opposite characteristics in some other ways. Twitter contains two user-centric attributes: The more users one follows, the higher the probability of these users following other users. As the number of one's followers increases, the higher the probability grows of the users being followed, following other users. The opposing attributes to a user-centric platform exhibited by Twitter include: The greater the number of users being followed, the lower the probability of growing the number of their followers; and as the number of one's followers increases, the number of users being followed, following other users' followers, decreases.

Myers et al. [50] observe that the behaviour of Twitter as both user-centric and content-centric is with respect to the users. They argue that when a new Twitter user is more interested in content, this influences the people that the user follows, such as users with many followers, companies or popular periodic events. The tendency of the user to follow this set of users decreases with time as the user begins to consider more than content, but factors such as common interests will further influence the choice of whom to follow. The user is said to be homophily - driven - users with similar interests have a higher probability of following one another than users with different interests [54].

The visibility of a tweet posted by a user is in proportion to the number of the user's followers. Unlike most user-centric platforms, Twitter users cannot invite other users to be their followers. Other than posting personal information, users can also join important and ad hoc conversations, generally preceded by the hashtag symbol. Rossi and Magnani investigate the factors that affect

the acquisition of new followers [55]. Logically, one may think that the level of activity of a user in a conversation will strongly influence the increase in the user's number of followers. Although this affects the followers' acquisition, it is not as strong as the number of mentions of a user, or the number of retweets of a user's original tweet. They conclude that in order to acquire new followers, a user not only has to be active, but s/he must pay attention to the content of this tweet in order to increase the probability of replies or retweets. This is consistent with the observation of Myers et al. [50] that Twitter is both a user-centric and content-centric platform.

2.3.2 Sentiment Analysis

Sentiment analysis aims at determining opinions, emotions and attitudes reported in source materials such as documents, short texts and sentences from reviews, blogs and news among others [56, 57, 58]. Opinion mining and sentiment analysis have witnessed significant developments in research over the past few years. Consumers and users have enthusiastically raised their voices and expressed their sentiments - which can be positive, negative or neutral - in the form of textual posts on social media for virtually anything they care about. Sentiment analysis has been applied to predict best travel destinations [59], public opinion in political debates and the subsequent elections [60], online reviews and product sales [61]. The large volume of opinionated data poses severe challenges for data processing and related sentiment extraction. Contemporary solutions such as machine learning have been proposed for the sentiment analysis of online textual data [62]. Existing machine learning approaches have given promising results [63].

2.3.2.1 Importance of sentiment analysis in Twitter

The importance of sentiment analysis in Twitter grows by the day, as evident from the following instances. The recent terrorist events in Boston, USA; Woolwich, London, the UK and Borno, Nigeria sparked widespread reaction and news reporting via social media. In each case information pertaining to the events had both positive and negative impacts in the immediate aftermath. In relation to positive impacts, Twitter was used by law enforcement officials and journalists to request information and relay assurances to the public. On the negative side, in Boston the vast amount of information posted by the public on Twitter led to law enforcement becoming overwhelmed with multiple lines of enquiry. In the UK a number of arrests were made following the event, due to the allegedly religiously offensive comments being posted on Twitter [64]. In Nigeria the then incumbent government was labelled weak due to the antagonistic commentaries posted on Twitter. In each case, the tension expressed in each tweet affects its propagation by retweets. The potential damage of these tweets can be minimized by direct engagement with the users to provide the required update and clarify any pressing issues.

2.3.2.2 Sentiment analysis approaches

Sentiment analysis is a process of determining the orientation of public opinion, which can be positive, negative or neutral. It is a method used to extract subjectivity and polarity from public opinion, usually in textual form. The two main sentiment analysis approaches are the lexicon-based method and machine learning.

1. Lexicon-based approach

   The lexicon-based approach involves the calculation of a document's orientation from the polarity and strength of words, phrases, or texts [65]. It is often assumed by researchers that words have prior polarity - that is, that words have an inherent polarity independent of their use in context. The classification involves building classifiers from predefined instances of texts and sentences [56]. These instances make up the dictionaries which can be created manually or automatically through the use of seed words - sets of words with strong positive or negative associations - to expand the existing list of words [66]. The popular dictionaries used by various researchers are based on adjectives, adverbs, nouns and verbs. The polarity of a document based on these dictionaries involves three steps: construction of word/value pairs; replacement of the words in a document with their values; and aggregating the values. Generally, the orientation of a document is determined by the aggregated effects of the values of the words as ranked in their dictionaries. The values of these words can be constructed on weighted values by modifying them through intensifiers. Some of the popular opinion lexicons in use are WordNet and General Enquirer. But the general assumption that a word or phrase has an inherent polarity is a major drawback of this method. Most of the posts on social network are informal and words can be of different polarity, depending on the context in which they are used.

2. Machine learning approach

   The other approach of sentiment analysis is machine learning, machine learning involves parameter-value pairs. The parameter represents the existing words; the value represents their frequencies if the position of the words in a document is ignored [67]. In this research, tweets were collected to produce their representatives' bag-of-words, discarding all information about the order of words. The text in each tweet and user profile was used to form the bag-of-words excluding the words in the stop-word list. Each word was stemmed using the built-in capabilities of the software AlchemyAPI and the frequencies of each stem were computed for each tweet. The machine learning classifiers are described in the eponymous subsection 2.3.2.3. AlchemyAPI has been used in this work as a principal automated sentiment analysis method because of its robust and simple interface. AlchemyAPI was also able to provide a discounted subscription package for

research purposes that enabled the required dataset to be processed without restriction and at no extra cost. AlchemyAPI have also been used successfully in previous related research [68, 69, 70]. However, there are a number of different automated sentiment analysis tools such as Sentistrength [71],SentiWordNet [72] and SenticNet [73] but all of them suffer from similar problems as highlighted in the following section.

2.3.2.3 Problems of sentiment analysis in Twitter

As with most Natural Language Processing (NLP) tasks, sentiment analysis is faced with challenges that can negatively impact on its accuracy. Some of the notable challenges are:

- **Context**: When a text expresses an opinion, context plays a huge role in determining its true polarity. In order to boost the accuracy of sentiment analysis systems, domain specific knowledge and context awareness often play an important part. A particular opinion word that is considered positive in one situation or domain may be considered negative in another scenario [74]. For example, "stubborn" can be classified as positive in "The Prime Minister is stubborn, he will not give in to the pressure of smugglers", whereas the same "stubborn" can be classified in the negative domain in "What a stubborn Prime Minister, he will not listen to wise counsel". However, generic sentiment analysis systems often fail to account for these idiosyncrasies, thus lowering their accuracy. In addition, the sparse nature of tweets (140 characters at most) makes it difficult to effectively convey context - which may refer to local tweets in the timeline.

- **Presence of ambiguous terms**: updates shared on Twitter contains many ambiguous terms such as misspelt words, abbreviations, ad hoc short forms, emoji, social media slang and many other irregularities that are not within the confines of a standard language. The presence of these terms can significantly affect the accuracy of traditional NLP tools for Twitter sentiment analysis. For instance, when carrying out parts of speech tagging, the models in use are often trained on lexically correct grammar, which makes them ineffective on the incorrect ones frequently expressed on Twitter. Furthermore, misspelt words and abbreviations result in inconsistencies in word count when using statistical techniques [75].

- **Diversity in user expression and validity of opinion lexicons**: in techniques using dictionary-based methods such as a lexicon of opinion words, the effectiveness of the setup heavily depends on the accuracy of the lexicon. This introduces additional ambiguity because many words are neither positive nor negative in absolute terms but rather depend heavily on the context in which they are used. In addition, some individuals tend

to use unusual words to express emotions [56]. For instance the use of negative adjectives to express emotion is common on social media, for example, "the movie is **insanely** awesome". While this tweet can be easily interpreted by humans and classified as positive, the use of 'insanely' in describing the movie introduces ambiguities into knowledge-based methods.

- **The use of slang and sarcastic words**: Twitter is a casual form of social media where people frequently use slang and sarcasm, giving words several conflicting meanings. This can also negatively impact the efficiency of generic sentiment analysis systems [76]. The operation of most automated sentiment analysis tools depends on predefined algorithms. Such settings dictate which tweet content will be viewed as relevant and which content will be identified as irrelevant. If the automated programme is too general, there is a possibility that a great deal of irrelevant content will be included in the brand, product or service evaluation. Similarly, if the programme or tool is very sensitive, a great amount of relevant content may be excluded from the evaluation process.

2.3.3 Supervised and Unsupervised Machine Learning

The term machine learning includes algorithms, which operate by building a model from input data and then using this model to make predictions for decision making [77]. Machine learning is primarily divided into supervised [78] and unsupervised learning [79] paradigms. Supervised learning is termed target-based learning [78], whereas unsupervised learning is a target free learning criterion [79]. Feature selection [80] and feature extraction [81] are two major techniques used in machine learning for dimensionality reduction. Feature extraction transforms the existing features into a lower dimensional space whereas feature selection selects a subset of the existing features without transformation [82].

2.3.3.1 Text categorization (supervised machine learning)

Text Categorization (TC) has been shown to be a powerful building block in several information frameworks and methods of data management. Automated text categorization is attractive because it liberates the classifier from manually curating document databases, which, as the number of documents increases, can be time-consuming and inefficient. In addition, automated text classification is applied using information retrieval (IR) technology and machine learning (ML) technology, which are more accurate than manual optimization. Published approaches mainly assign text to a specific category by comparing them with a bag-of-words model of documents. However, during this process the linguistic features such as micro-text [83], semantics [84] and

syntax recognitiion [85] are still ignored in the automated learning. Spam recognition and filter are widely used examples of applying text categorization whereby received emails are automatically categorized as spam/junk or non-spam [86]. Applying text classification in practice is an interesting topic for research since so much text-based data is generated every day. A deep understanding of text classification gives researchers the chance to develop new applications, for they can easily obtain data which require classification, including emails and micro-text. Early techniques used in text categorization were built up from linear classifiers, which focused on efficiency. Other aspects of text categorization include, for example, leveraging cross-category dependencies, ways of "borrowing" training examples surrounded by mutually dependent categories and ways of discovering latent structures in a functional space for the joint modeling of dependent categories [87, 88]. Current research focuses on classifying data according to topic; other types of classification are also interesting, for example, classifying data by sentiment: or determining whether a review is positive or negative [76] or, when texts are being classified, whether a text is misleading or not. Nevertheless, the models and procedure for topic categorization are also significant in these problems and some remarkable deliberations over the qualities of the categorization seem to be the best guides for improving performance. Notable current applications include the use of text mining techniques to predict the rise and fall of the stock market. For instance, [89] used the classifier ANN (Artificial Neural Network) on Twitter data to understand users' moods in relation to the stock market and on this basis to predict its fluctuations. [88] predicted the results of stock market indicators such as the Dow Jones, NASDAQ and S&P 500 by analyzing Twitter posts. [90] is an exploratory study of Twitter user attribute detection which uses simple features such as n-gram models, simple sociolinguistic features such as the presence of emoticons, statistics about the user's immediate network such as the number of followers/friends and retweet frequency as communication behaviour. Commonsense knowledge-based approaches for text analysis have also gathered plenty of buzz in this field of research. Some of the notable machine learning classifiers include:

- **Support Vector Machines (SVM):** Support Vector Machines are discriminate classifiers formally defined by a separating hyper-plane which is efficient for text categorisation. SVMs were developed from statistical learning theory by Vapnik & Vapnik [91] on the basis of the structural risk minimisation principle. The algorithm classifies opinionated text vectors by separating it into positive and negatives classes with a hyper-plane, which can be further extended to non-linear decision boundaries using the kernel trick [63].

- **Naïve Bayes (NB):** The Naïve Bayes classifier is a probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naïve) independent assumptions. A more descriptive term for the underlying probability model would be the "*independent feature model*". NB classifiers have worked unexpectedly well in many complex

situations in practice [92].

- **Decision Tree (DT):** A decision tree is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision [93]. Decision trees are easily interpretable, because the tree structure can be represented graphically and we can follow branches down the tree according to the input variables, requiring less training time.

- **K-Nearest Neighbour Classifier**: The K-nearest (KNN) classifier is an instance based classifier that relies on the class labels of training documents which are similar to the test document. Thus, it does not build an explicit declarative model for the class. KNN classifications proceed in two stages; the first determines the nearest neighbours and the second determines the class using those neighbours [94].

2.3.3.2 Clustering (unsupervised machine learning)

Data clustering is the process of identifying natural groupings or clusters within multidimensional data based on some similarity measures [95]. Clustering can be defined as "*the organisation of a collection of patterns (usually represented as a vector of measurements, or a point in a multi-dimensional space) into clusters based on similarity*" [95]. While classification is supervised learning in which the categories are known beforehand and given in advance for each training document, clustering is the unsupervised version of in which the goal is to discover the natural grouping (clustering) of patterns [96]. In clustering, there is no predefined class but groups of cognate documents are sought. Clustering algorithms are used in many applications, such as image segmentation [97], vector and colour image quantization, data mining, compression and machine learning [98]. The first studies of cluster analysis were conducted in the field of analytical psychology. According to Bailey [99], it was Robert C. Tryon who originally conceived clustering and first applied it to psychological data in 1930. It was not until three decades later, in 1965, that the method was implemented as part of a software package, following the introduction of the first modern computer, at the University of California. Tryon referred to this particular practice of clustering as *variable analysis* or *V-analysis*. The main purpose was to identify composites of abilities that could serve as more "general" and relevant descriptors than the whole set of scores, to provide a more accurate analysis of human differences. This clustering method, called "key-cluster factor analysis", was proposed as an alternative to the factor analysis generally adopted at the time. In recent years, researchers have focused on various problems such as the summarization and detection of topics for Twitter messages, as well as the mass clustering of tweets. For example, TweetMotif [100] (an exploratory search application for Twitter) uses an unsupervised approach to message clustering. The weakness of this method

as presented in [100] is that it does not report metrics on the system performance, nor provide comments on the generalizability of the approach. This happens due to the lack of applicable performance metrics and gold standard labels. Another application of unsupervised techniques on Twitter data, given by [101], focuses on predicting the geo-location of a tweet based on the text in the tweet, which made use of the geo-tagged information in the tweets as the gold standard label for measurement. In [102] the authors propose a novel clustering hashtag criteria for tweets. The authors argue that two hashtags are similar if they co-occur in a tweet. As in [103], the authors expand this concept by introducing a novel method for measuring the similarity between two hashtags. The authors use a larger set of hashtags and test several clustering methods instead of focusing on only one. Another work on clustering text-related entries typically, to make clustering computationally feasible, focuses on a bag-of-words model that takes all the words of the entity followed by dimensionality reduction [104, 105]. Since Twitter provides a constant stream of real-time updates from around the globe, much research has focused on detecting noteworthy, unexpected events as they rise to prominence in the public feed. These include the detection of influenza outbreaks [106], seismic events [107] and the identification of breaking news stories [108]. These applications are similar to the efforts to stream data mining focused on other media outlets such as [109]. Clustering has a number of algorithms; one of the most popular, which is used in this study, is K-Means.

K-Means

Assuming that some elements are drawn from a certain probability distribution, objective of k-means is to find the values of the prototype vectors to minimize the error [110]. In the k-means algorithm, the search for the optimum weights of the prototypes is performed iteratively through a stochastic gradient descent on the error surface. For example, n data points are divided into k clusters such as to minimize the distance between each data point and the centroid of the nearest cluster. Two major methods have been used to measure the distance between any data point and the centroid of the nearest cluster: Euclidean distance and Cosine distance. In this study, Euclidean distance was used to measure the distance between the centroid of k clusters and the data points. The main advantages of k-means are its fast convergence and thus the low computational cost of the algorithm [111]. It is one of the simplest partitional algorithms [112]. K-means clustering is one of the most widely used clustering techniques in the commercial field and also works efficiently on high dimensional data [113]. The k-means algorithm is very easy to implement and its linear time complexity makes it suitable for very large amounts of data [112]. However, the main limitation associated with k-means algorithm is its proneness to converge to a local optimum, depending on the choice of the initial prototypes. The k-means clustering algorithm attempts to split a given dataset into a fixed number (k) of clusters. Initially

k number called centroids are chosen. A centroid is a data point at the centre of a cluster. In each centroid is an existing data point in the given dataset, picked at random, such that all centroids are unique (c_i and c_j, c_i neq c_j). The resulting classifier is used to classify the data (using $k = 1$) and thereby produce an initial randomized set of clusters. Each centroid is thereafter set to the arithmetic mean of the cluster it defines. The process of classification and centroid adjustment is repeated until the values of the centroids stabilize and converge. The final centroids will be used to produce the final clustering of the input data, effectively turning the set of initially anonymous data points into a set of data points, each with identified class.

2.4 Evaluation Measures

Some of the common systems of measurement to evaluate information retrieval methods are precision and recall and F-measure. These metrics are used for validating accuracy in different ways, yet they can be applied to other purposes also and are useful in describing how Twitter classification methods are successful. In data mining, precision is generally defined as the ratio of true positive documents to the documents that are correctly real positives, that is, the summation of true positive and false positive documents [114] as shown in Equation 2.1. Recall is the ratio of true positive documents to those that are correctly predicted, that is, the summation of false negatives and true positives [114], as shown in Equation 2.2.

$$precision = \frac{true\ \ positives}{true\ \ positives + false\ \ positives} \tag{2.1}$$

$$Recall = \frac{true\ \ positives}{false\ \ negatives + true\ \ positives} \tag{2.2}$$

In their study to assess retrieved tweets, Castilo et al. [115] define precision as the ratio of the number of correct credible classifications to the number of total classifications made and recall as a ratio of the correct classifications to potential classifications (see Equations 2.3 and 2.4).

$$precision = \frac{Number\ \ of\ \ correct\ \ classifications}{Total\ \ classifications\ \ made} \tag{2.3}$$

$$Recall = \frac{Number\ \ of\ \ correct\ \ classifications}{Total\ \ number\ \ of\ \ potential\ \ classifications} \tag{2.4}$$

In another study carried out by Hong et al. [116] they observed that the goal is to maximize both precision and recall. In their efforts to determine the interestingness of tweets, they defined

precision as a ratio of the number of interesting tweets available to a user to the total number of available tweets to a user and recall as a ratio of the number of interesting tweets available to a user to the number of interesting tweets available to all users. They conclude that recall increases with the increasing number of followers while precision decreases with the increasing number of followers. This form of measurement was also employed by Pak and Paroubek [117] in their sentiment analysis method in order to determine if a tweet is positive, negative or neutral. The authors used accuracy and decision instead of precision and recall respectively. They defined accuracy as precision was defined by Castillo et al. [115]. On the other hand, they defined decision as:

$$Decision = \frac{Number\ \ of\ \ retrieved\ \ tweets}{Number\ \ of\ \ all\ \ tweets} \tag{2.5}$$

In the present study however, precision is the ratio of the number of tweets that correctly categorized by automated sentiment tool to the number of total categorizations made automatically; and recall as the number of correct automated sentiment categorizations to the total number of manual sentiment categorizations of tweets as mathematically shown in Equations 2.1 and 2.2.

The categorization of a tweet into positive, negative or neutral is correct if decisions such as predicting the market trend for a business is consistent with the categorization. Another widely used quality metrics are the F-measure and the accuracy; these are also used in this research. They essentially measure the quality of sentiment methods. This statistical tool is the harmonized mean of precision and recall while the accuracy is the total correctly classified tweets normalized by the total number of tweets [118]. They can be mathematically described as follows:

$$F - measure = 2 \times \left(\frac{precision \times recall}{precision + recall}\right) \tag{2.6}$$

$$Accuracy = \frac{true\ \ positive + true\ \ negative}{true\ \ positive + false\ \ positive + true\ \ negative + false\ \ negative} \tag{2.7}$$

Additionally, this study uses sensitivity and specificity measures to verify the clustering algorithm. Sensitivity also known recall or true positive rate while specificity called true negative rate [114] that defined as:

$$Specificity = \frac{true\ \ negative}{true\ \ negative + false\ \ negative} \tag{2.8}$$

2.5 Conclusion

Social media platforms such as Twitter and Facebook have proven to be a powerful communication media for companies in recent years. This work aim to use sentiment analysis techniques to analyse company communication from Twitter. It is worthy of mention that Twitter seems to be the preferred option because of its more compact data size compared to others. Facebook for instance allows users to write up to 5000 characters while Twitter restricts tweets to a maximum of 140 characters. Although analysis done on Twitter data could easily be extended to data from other social media platforms with some modifications. This chapter explored the two major sentiment analysis approaches, lexicon-based and machine learning. Majority of existing sentiment analysis tools are based on these two approaches. Any of the existing tools could have been successfully used for analysis of the datasets in this research with similar result. However, only one is used as the overall aim is to compare automated sentiment results with manual method. In addition, both automated and manual sentiment methods can be combined in some cases. Automated sentiment performed well with the tweets issued by some user groups and failed with others. To reduce this problem the indirect tweets issued by specific groups that difficult to evaluate automatically or give inaccurate accuracy both techniques should be used. Supervised machine learning is applied to predict the user groups and investigate the role of user type in the sentiment accuracy. Various classifiers with different algorithms are implemented in this work to reach the accurate results. In addition, unsupervised machine learning is also used to investigate the sentiment accuracy within similar group of tweets by applying the k-means as a popular clustering algorithm.

Chapter 3

Twitter Usage by Oil Companies

Introduction

This chapter gives an overview of the data used in this research. It explains how the data was extracted and sets collected. Since tweets may contain other objects, apart from simply words and phrases, such as *hashtags* (#), hyperlinks, *mentions* (@), and *retweet* (RT) among others; this chapter also gives brief details of each of these objects as well as comparison of their usage between the two principal oil companies adopted as a case study.

This study considers two of the biggest oil companies in the world as ranked by the Petroleum Intelligence Weekly [119]. British Petroleum (BP) based in America represents western, and Saudi Aramco based in Saudi Arabia, represents Middle Eastern cultural background. Both companies have active Twitter accounts; and for the remainder of this thesis BP will be referred to as *BP_America* and Aramco as *Saudi_Aramco*. The rationale for choosing these companies is based on factors such as: similar business size, industry, objectives and the fact that they are in culturally distinct countries. BP_America is a global oil company located in the U.S. The company reported revenue was $358 billion in 2014, with 84,500 people employed by the company worldwide and a production of 3.2 million barrels per day [120]. BP_America joined Twitter in August 2008 with total 14,800 tweets, 115,000 followers and follow 568 users till 2015. Further, Saudi_Aramco is located in the kingdom of Saudi Arabia, with revenue of $378 billion, 60,000 employees, and production of around 12.5 million barrels per day [121]. In June 2009 Saudi_Aramco joined Twitter. They have, until 2015, generated 2403 tweets and were followed by 240,000 followers, and follow 70 users. Another reason why oil companies were chosen is that they face significant challenges to overcome in their public relations. This is particularly in light of recent negative environmental issues such as the oil spill in Gulf of Mexico incident caused by an explosion on a platform operated by British Petroleum (BP) in 2010 [17]. Since their followers on Twitter belong to a wide variety of groups and as such express many different reactions in their user mentions.

Section 3.1 provides an overview of datasets used in this thesis. The subsection 3.1.1 presents

the tweet rate over time by the two oil companies. The utilization percentage of Twitter features (hashtags, hyperlinks, retweets and mentions) and the differences between both companies are presented in subsections 3.1.2, 3.1.3, 3.1.4, 3.1.5 respectively.

3.1 Overview of Primary Analysis Dataset Collection

A Python script was developed to consume the Twitter API and fetch the tweets from both accounts. The tweets retrieved were stored in a SQLite database for further processing. Two types of datasets were used in this research: (1) companies tweets (tweets generated by companies) and (2) companies mentions (tweets generated by users mentioning the companies' name). The companies' tweets datasets were continuously collected from April 2011 to August 2014, while the "mentions" datasets were collected on two different time periods from November 2012 to August 2014 (see Table 3.1).

Dataset I (A, B) is the total set of tweets and mentions collected within the given time frame. Dataset II (A, B) is a subset of Dataset I and represents the tweets and mentions that undergoes sentiment analysis in chapter 4. Further, Datasets II (B) is used in the classification and clustering experiments in chapters 5 and 6 respectively.

Moreover, Dataset I was further reduced to Dataset II by modifying the extraction script to exclude tweets and mentions that are not relevant for sentiment analysis. Tweets and mentions that are:

- Arabic or any language other than English as the sentiment analysis tool used in this work is restricted to specific languages and does not support Arabic
- Retweets; as they will amount to duplication of original tweets

Due to the type of companies chosen as case study, the volume of their tweets and mentions is low compared to companies in other sectors. In addition, the search by Twitter API is focused

Table 3.1: Datasets details

	Tweets (A) April 2011 - August 2014		**Mentions (B) November 2012 - August 2014**	
	Saudi_Aramco	**BP_America**	**Saudi_Aramco**	**BP_America**
Dataset I	1550	5951	4567	5662
Dataset II	1000	1000	3000	3000

on relevant tweets and not completeness. This means that some tweets and users may be missing from search results.

3.1.1 Tweet Rate over Time

As shown in Table 3.2, BP_America have a high daily average tweet rate of approximately 5 while Saudi_Aramco have a meagre 1. This difference may be attributed to the cultural and social differences between the two environments where these companies operate. The information in Table 3.2 reveals that BP_America daily tweet rate is three times more than Saudi_Aramco.

Table 3.2: The rate of companies' tweets over time (01/04/2011- 14/08/2014)

	Total tweets	**Average daily tweets**	**Average weekly tweets**	**Average monthly tweets**
BP_America	5951	4.83	33.60	140.31
Saudi_Aramco	1550	1.30	8.82	37.60

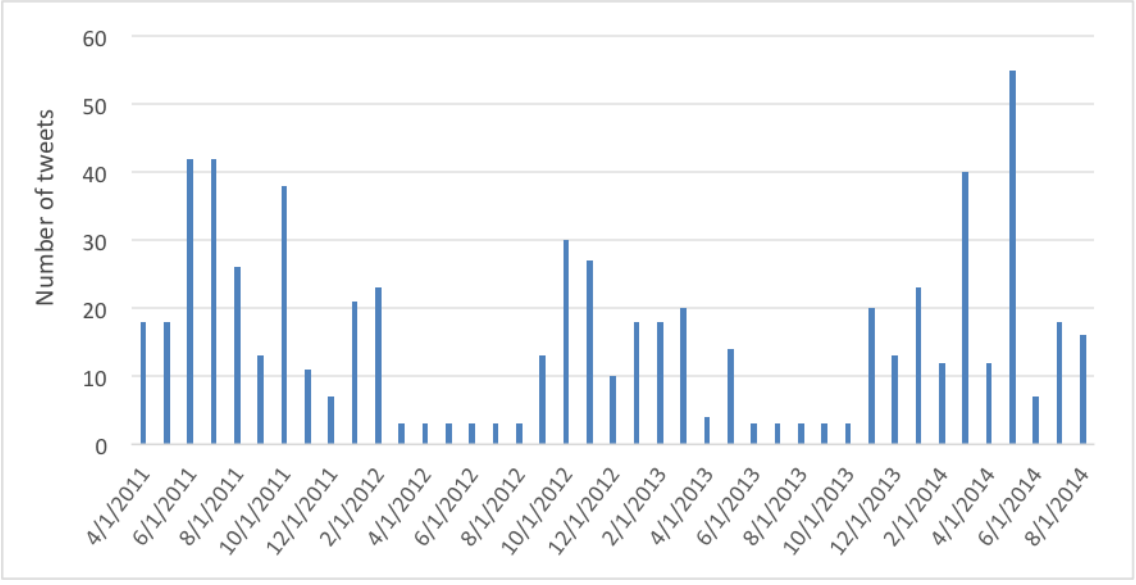


Figure 3.1: BP_America tweets rate over time

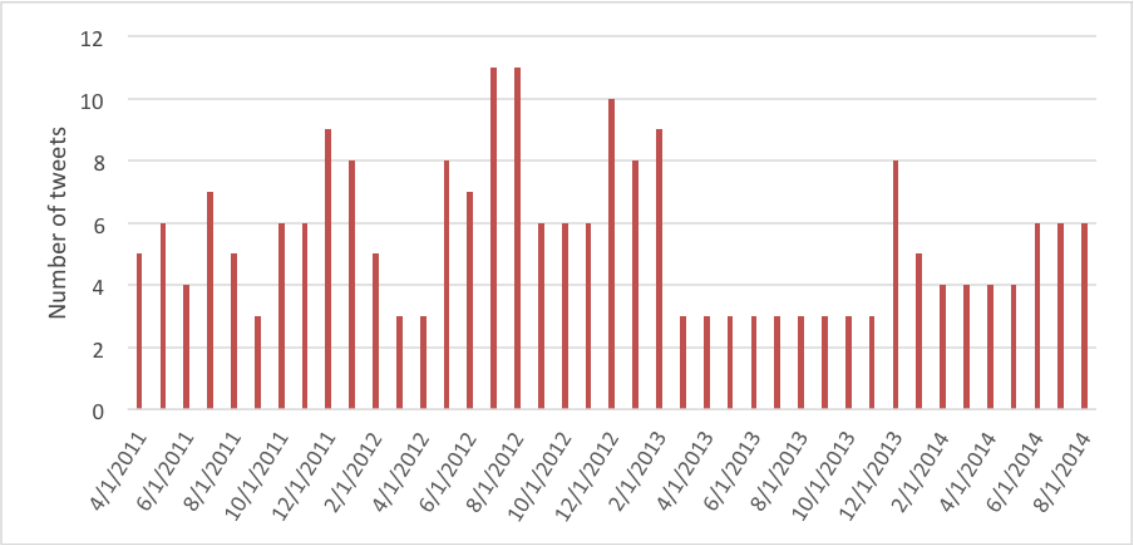
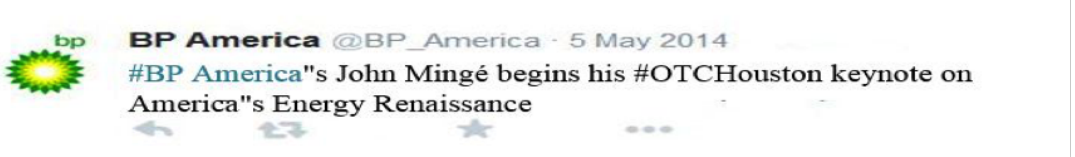


Figure 3.2: Saudi_Aramco tweets rate over time.

Figures 3.1 and 3.2 show the trend of activities on the Twitter pages of BP_America and Saudi_Aramco respectively. The spikes seen in these figures are related to some hot topics or events such as publication of reports, release of products and important talks. For example, on 08/06/2011 BP tweeted (42) tweets because of statistical review of world energy was released.

Another example on 05/05/2014, BP tweeted (55) tweets about America's Energy Renaissance event.



Similarly, figure 3.2 shows the tweet rate of Saudi Aramco is higher on some dates. For example, on 24/07/2012, they tweeted (11) tweets because of the publication of a new edition of “Aramco Journal of Technology”:

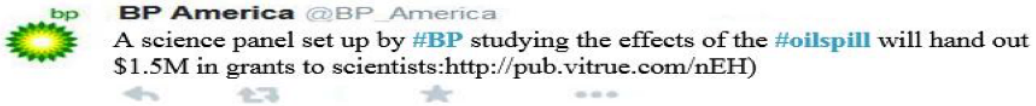
The tweet rate on 30/12/2012 is (10) tweets; mainly related to the "Traffic and Safety Research Chair" established by Aramco at the university of Dammam which was considered as a big research achievement for the company. For example:

3.1.2 Hashtags

A hashtag is a word or unspaced phrase preceded by a hash symbol (#). Hashtags are usually used as a label or a tag to mark keywords or topics in a tweet [122]. Companies use Hashtags to categorize their tweets and attract users that may not be followers of their pages, brands, products, services or events. When users search by hashtags they will find tweets on specific subject easily than searching for full text of specific tweets [123]. Hashtags also allow trending topics to be readily identified. The extracted data used in this thesis shows that BP often use

the hashtag (#BP) in each tweet–a brand hashtag unique to its business. Brand hashtags can be company name or a tagline (a slogan or phrase convey company theme) that people associate with a company. They are often short, easy to spell, and memorable. When companies use brand hashtags consistently potential followers are easily attracted and the company name gets extended reach. The tweet below is an example of using BP a brand hashtag:

Additionally, companies can use content hashtags which are related to their post such as products, events or news hashtags. For example:



As shown in Table 3.3 the average number of hashtags used by BP is 1.62 per tweet while Aramco used 0.79 hashtags per tweet. After investigating Aramco tweets, the brand hashtag (#SaudiAramco) seems not to be used consistently whereas content hashtags are used more in special events or when important news is released. A good example of a hashtag in Aramco tweets is:

Table 3.3: BP_America and Saudi_Aramco hashtags usage (01/04/2011- 14/08/2014)

	Total tweets	**Average daily tweets**	**Average weekly tweets**
BP_America	5951	9670	1.62
Saudi_Aramco	1550	1231	0.79

3.1.3 Hyperlinks

Hyperlinks in social network sites reference other data or webpages. In Twitter, all hyperlinks (URLs) posted in tweets are automatically shortened to a maximum of 22 characters by Twitter's links shortener service (t.co) in this format: `http://t.co`. This service allows users to share long URL without exceeding the character limit and also check links against a list of potentially dangerous sites to protect users [124]. Companies can use hyperlinks in their tweets to drive users to their websites, blogs or any related sites. Hyperlinks can be internal which lead to companies' websites or external which lead to different websites. It is clear that both BP and Aramco post links in most of their tweets, however, the main difference between them is the type of links that they share. Table 3.4 shows that BP frequently post external links to other websites such as YouTube to clarify projects or issues, while Aramco often shares internal links referring to information on its website.

Table 3.4: BP_America and Saudi_Aramco hyperlinks usage (01/04/2011- 14/08/2014)

	Total tweets	**Total hyperlinks**	**Hyperlinks per tweet**	**Total internal Hyperlinks**	**Total external Hyperlinks**
BP_America	5951	4635	0.78	878 (18.94%)	3757 (81%)
Saudi_Aramco	1550	1307	0.84	821 (62.81%)	486 (37.18%)

3.1.4 Retweets

A retweet (RT) is an original tweet that has been re-posted by another user and appears as (RT@ user name). Twitter users can use the retweet feature to share and spread any tweet to their followers [125]. Users can retweet any kind of information from other users even if they are not followers and such tweets will be visible on their timeline as a normal tweet. Companies use the retweet feature to distribute interesting and positive tweets published by other users about their company. This can be a good way to attract more followers. From Table 3.5 it is obvious that both BP and Aramco have a relatively low percentage of retweeting of other users' tweets. Less than 1% of their tweets were retweeted and these tweets are often related to other company's department Twitter accounts, for example:

Table 3.5: RT rate of BP_America and Saudi_Aramco (01/04/2011- 14/08/2014)

	Total tweets	**Total RTs**	**Percentage RTs**
BP_America	5951	4635	0.09%
Saudi_Aramco	1550	56	0.04%

3.1.5 Mentions

A mention is a tweet which contains another Twitter username, preceded by the "@" symbol (example: @username) [126]. Mentions appear in the recipient's timeline if they are following the sender but can also be sent from followers or non-followers. Companies can engage directly with users who mention their Twitter user name. Monitoring mentions helps companies to work out issues before they escalate to preserve their reputation and publicly respond to complaints. Mentions also give the company a good idea about the categories of their online followers and their trends. Companies usually have high mention rate when they are involved in events that affect fellow Twitter users. Table 3.6 and 3.7 shows the mention rate of BP_America and Saudi_Aramco respectively. The unusual spike seen in Figure 3.3 of BP_America mentions was the period when an oil spillage court case was decided against BP_America, hence the high rate of mentions from Twitter users hailing the court decision [127]. Below are a couple of mention examples:

Table 3.6: The rate of BP_America Mentions over two time period

Time period	Date	Total tweets	Average daily mentions	Average weekly mentions	Average monthly mentions
I (3 months)	09/11/2012 - 06/02/2013	1656	32	226	574
II (2 months)	15/06/2014 - 22/08/2014	4006	114	696	1501

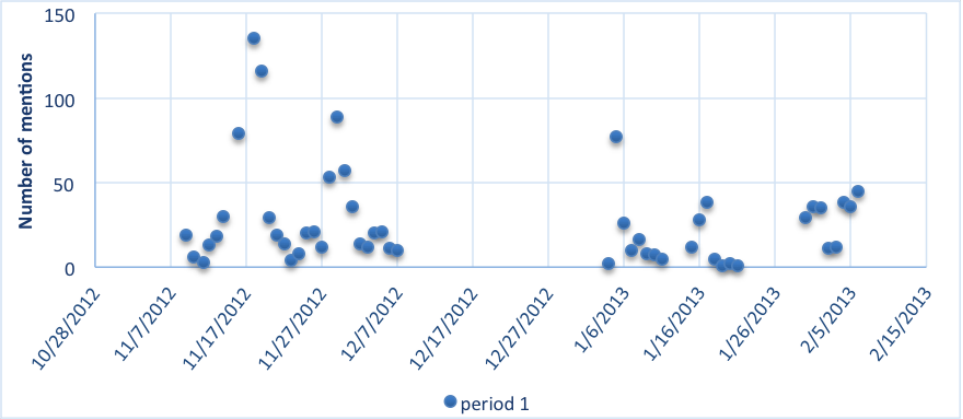


Figure 3.3: BP_America mentions rate for period 1 between 09/11/2012 and 06/02/2013

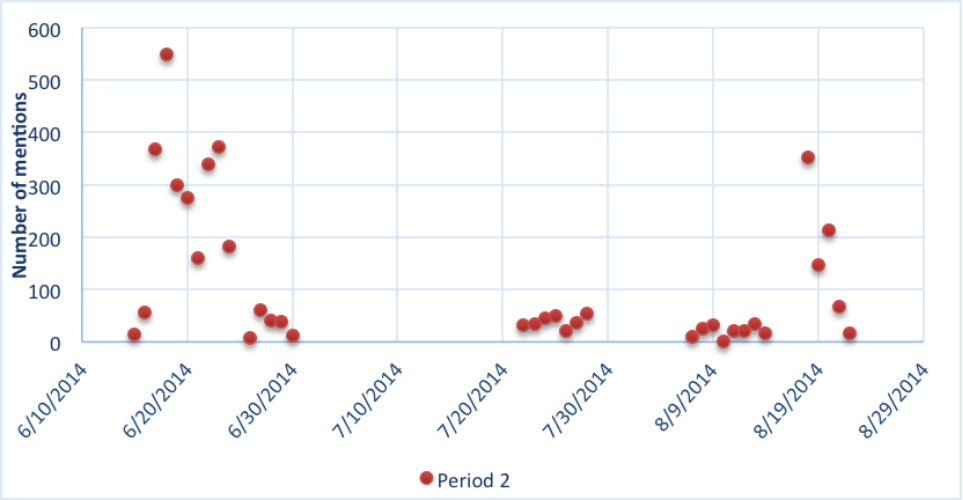


Figure 3.4: BP_America mentions rate for period 2 between 15/06/2014 and 22/08/2014

It is clear that the average rate of Saudi_Aramco mentions is lower than BP_America as indicated in Table 3.7. Some information and news released from Aramco resulted in increased mention rate; as seen in Figure 3.5, the spike shows large amount of mention in that date, a couple of mention examples from this date include:

Table 3.7: The rate of Saudi_Aramco mentions over two time period

Time period	Date	Total tweets	Average daily mentions	Average weekly mentions	Average monthly mentions
I (3 months)	05/11/2012 - 06/02/2013	1327	28	197	800
II (2 months)	20/06/2014 - 28/08/2014	1672	84	495	708

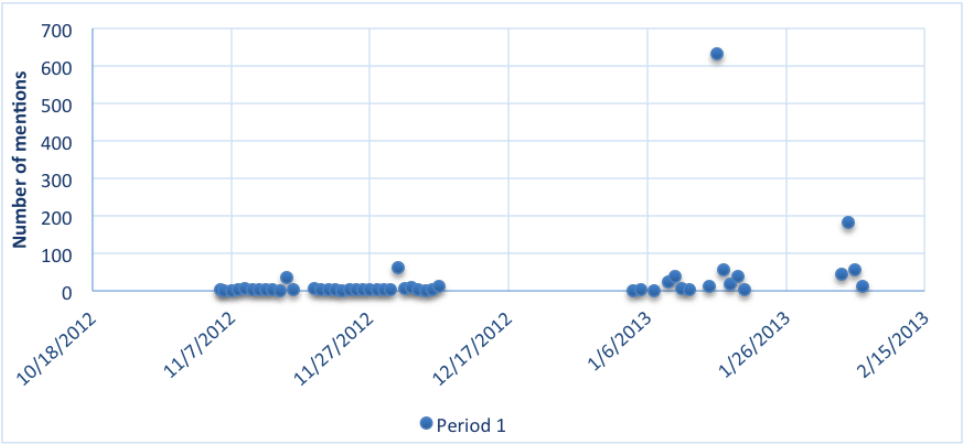


Figure 3.5: Saudi_Aramco mentions rate for period 1 between 05/11/2012 and 06/02/2013

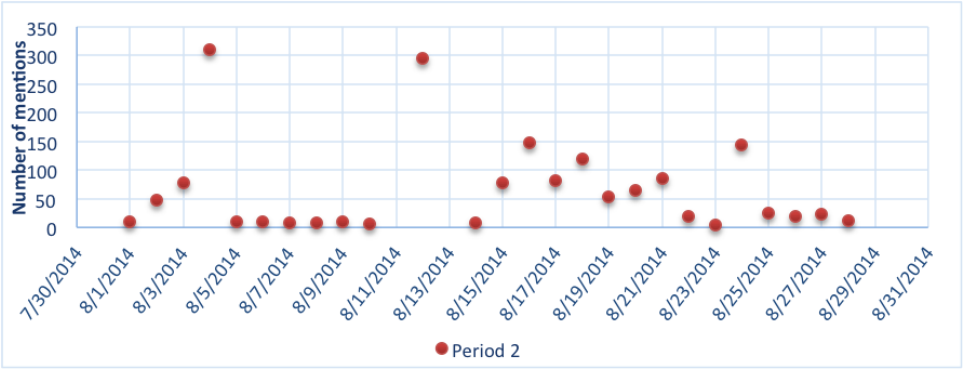


Figure 3.6: Saudi_Aramco mentions rate for period 2 between 20/06/2014 and 28/08/2014

3.2 Conclusion

A review of tweets extracted from the Twitter pages of BP_America and Saudi_Aramco over a given period of time shows that BP_America have a more pronounced online presence than Saudi_Aramco. This is also evident in their use of other Twitter features such as hashtags (#), hyperlinks, mentions (@), and retweet (RT) among others. Saudi_Aramco have limited use of Twitter and other social media tools may be attributed to the more reserved culture of individuals from the Middle Eastern part of the world. Further features can be used in Twitter is "favourite" feature with star icon which changed lately in 2015 to "like" feature with heart icon (after the data collection period of this theses). Users usually favourite interesting tweets even positive or negative. That means this feature does not support the sentiment polarities decision. In future work, using "like" feature by users when they like company's tweets will be examined to evaluate its role in the sentiment analysis and user categories classification.

Chapter 4

Sentiment Analysis

Introduction

Sentiment Analysis or (Opinion Mining) refers to the research area that deals with extracting the opinion or emotion conveyed in text [128], and is commonly handled as a Natural Language Processing (NLP) task. Sentiment information is found in text across diverse domains such as product reviews, tweets, blog comments, movie reviews, instant messaging and many other sources of textual information.

Social media platforms such as Twitter allow people to share experiences and user opinion across a wide range of different sectors. The public information found on Twitter is therefore a valuable source of data to be mined to investigate sentiment in different domains. At the time of this research, no application of sentiment analysis to the oil sector was found neither was there any of similar approach to this work. Oil companies were chosen because they have significant challenges to overcome in their public relations. This is particularly true in the light of negative environmental issues usually blamed on this sector. In addition, Twitter users who follow or mention these companies belong to a wide variety of groups and as such express many different opinions in their tweets.

The research work presented in this chapter focuses on investigating the sentiment expressed about oil companies across different cultures. The tweets collected and described in chapter 3 are analysed and investigated in order to identify any differences between the public and stakeholders opinions, and if any cultural factors can be discerned. The hypothesis is that the way people express themselves in different cultures will be evident in their tweets. The analysis performed in this work is referred to as *sentiment polarity classification*, which involves classifying tweets as either positive, negative or neutral. Two techniques are used for this task: (1) manual human classification through online crowdsourcing platforms, and (2) an automated method through the AlchemyAPI. The manual sentiment analysis technique represents the gold standard or ground truth as it is conducted by humans. While it is hard for machines to correctly identify the different types of emotions expressed through the complexities of human

language, human judgment can detect this easily. This chapter aims to rigorously evaluate the effectiveness of existing automated technique compared to the manual method, across different user groups and cultures.

This chapter is organized as follows: Section 4.1 presents sentiment analysis techniques including manual and automated approaches used in this research. Some experimentation details regarding results and an initial comparison in two phases are explained. Firstly, between two techniques of sentiment analysis and secondly, between the sentiments (positive, negative or neutral) expressed by the tweets of BP_America and saudi_Aramco are discussed in Section 4.2. Finally, the companies' mentions are classified into different groups and analyse the sentiment for each group by both manual and automated techniques and then combine them to suit each category in section 4.3.

4.1 Sentiment Analysis Techniques Used in This Work

This section looks to the manual sentiment to know what the real sentiment is then investigate how close the automated sentiment is to that. To achieve that two distinct methods were used for sentiment analysis in this research. The first is a manual technique which involves humans annotating tweets with the sentiment polarity conveyed. The second technique used is an automated technique using the AlchemyAPI. The aim of using automated and manual sentiment techniques is to investigate the effectiveness of automated tools compared to the manual method. The intricacies of both methods and the motivation for choosing AlchemyAPI as a representative of automated tools are described in subsection 4.1.2.

4.1.1 Manual Sentiment Analysis

The manual approach to sentiment analysis involves asking humans to read a document (or tweet in this case) and decide the polarity of sentiment. The manual method to sentiment analysis can be seen as the most accurate because it represents how human perceive sentiment polarity in text. For example, using indirect expression could be miss-interpreted by automated methods such as "*It's difficult to take a bad picture with this camera*" it is a positive statement but automated method can failed to evaluate it correctly because of the word (bad) is negative. Although manual approaches are considered as accurate, they are very time consuming. Therefore, large-scale manual sentiment analysis, in particular in systems requiring real-time sentiment information would be impractical. Another grey area in manual sentiment analysis is that humans are known to differ in opinion about sentiment expressed in text. This often results in different humans' judgment conflicting sentiment polarity of the same statement. This

research aims to investigate such conflicts by estimating the extent to which humans agree on the sentiment expressed in text.

Manually labelling a large corpus with sentiment information is often performed via crowdsourcing platforms like Amazon Mechanical Turk (AMT) or CrowdFlower [129, 130, 131]. These platforms provide an interface to outsource the annotation task at a fee for each participant. AMT and CrowdFlower were used in this thesis.

4.1.2 Automated Sentiment Analysis

There are a number of different automated sentiment analysis tools such as AlchemyAPI [68], Sentistrength [71], SentiWordNet [72] and SenticNet [73] but all of them suffer from similar problems as highlighted in chapter 2.3.2. In this work, AlchemyAPI was used as the principal automated sentiment analysis tool because of its robust and simple interface. AlchemyAPI was also able to provide a discounted subscription package for research purposes that enabled us to process the required dataset without restriction and at no extra cost. AlchemyAPI has also been used successfully in previous related research [68, 69, 70].
AlchemyAPI offers a very simple and easy-to-use method of identifying the positive and negative sentiments within any web page or document [68]. It has the capability of computing sentiment for a user-specified target, quotation-level sentiment, document-level sentiment, entity-level sentiment, keyword-level sentiment and directional-sentiment. These several types of sentiment analysis offer a variety of useful cases ranging from social media monitoring to trend analysis. The underlying algorithm of AlchemyAPI sentiment analysis looks for words in a text that contain either a positive or negative connotation then it finds out which place, person or thing they are referring to. The algorithm also understands negations and modifiers. This tool works well on both large and small documents including product reviews, tweets, comments, news articles, and blog posts [69].

4.2 Experiments, Findings and Discussion

The sentiment analysis task was carried out using both manual and automated methods. The following subsections describe the sentiment classification process followed for both methods. It should be noted that three distinct classification tasks were carried out, each of them are outlined below:

- In the first phase the tweets and mentions were classified as positive, negative or neutral using both the manual and automated methods.

- In the second phase, the role of neutral sentiment was assessed.
- In the third phase the Twitter account of users and organisations mentioning the oil companies were categorized into eight main groups and analysed, then investigate the role of user group in the accuracy of both sentiment methods.

4.2.1 Experiments and Analysis

The dataset used in this chapter is a subset of the data described in the previous chapter 3. Manual and automated sentiment analysis have been carried out on the tweets dataset (Dataset II-A) and the mentions dataset (Dataset II-B).

4.2.1.1 Manual sentiment analysis techniques

Manual sentiment analysis was done using two crowdsourcing platforms, namely Amazon Mechanical Turk and CrowdFlower. Since humans are known to differ in their opinion about sentiment in text, each tweet was classified by different annotators (people that perform tasks on crowdsourcing platforms) to get a balanced view. In previous works, researchers tend to evaluate their data by 3-5 annotators to get the average agreement for best results [130]. In this research, a trial of 100 tweets was carried out with 3 annotators for each tweet, another 100 tweets for 5 annotators. Analysis of the outcomes from both trials revealed that where annotators differ, they only do so between neutral and positive, or between neutral and negative. None of the 200 tweets in both trails have annotators split between positive and negative sentiments. Given this outcome, 3 annotators were considered sufficient for this experiment to draw a good human judgement of each tweet. Hence, the overall sentiment assigned to a tweet was defined as the average of the sentiment assigned by all annotators. As a requester (people that add tasks to crowdsourcing platforms) the HITs (Human Intelligent Tasks) were added and provided a description for the annotators to help them perform the task. Additionally, golden questions were added to the tasks which include the ideal answers and evaluation for different tweets belong to different sentiment polarities and user groups. These golden questions can guide the annotators if they have any confusion.

Amazon Mechanical Turk (AMT)

Amazon Mechanical Turk[1] (AMT) is a platform that allows humans to perform tasks that are very difficult for automation tools to execute intelligently, such as extracting data from images,

[1]Amazon mechanical Turk can be accessed at: `https://www.mturk.com/`

audio transcription, and filtering adult content. The platform makes it easy to distribute the task across multiple individuals and generate high quality labelled datasets for artificial intelligence related tasks. The service is ideal for sentiment analysis because humans can effectively identify the sentiment conveyed in a 140-character tweet. Humans are better at accurately detecting sarcasm and identifying the context in which a word when compared to machines. Figure 4.1 represents a screenshot of the task page as seen on AMT with the instruction and a sample task.

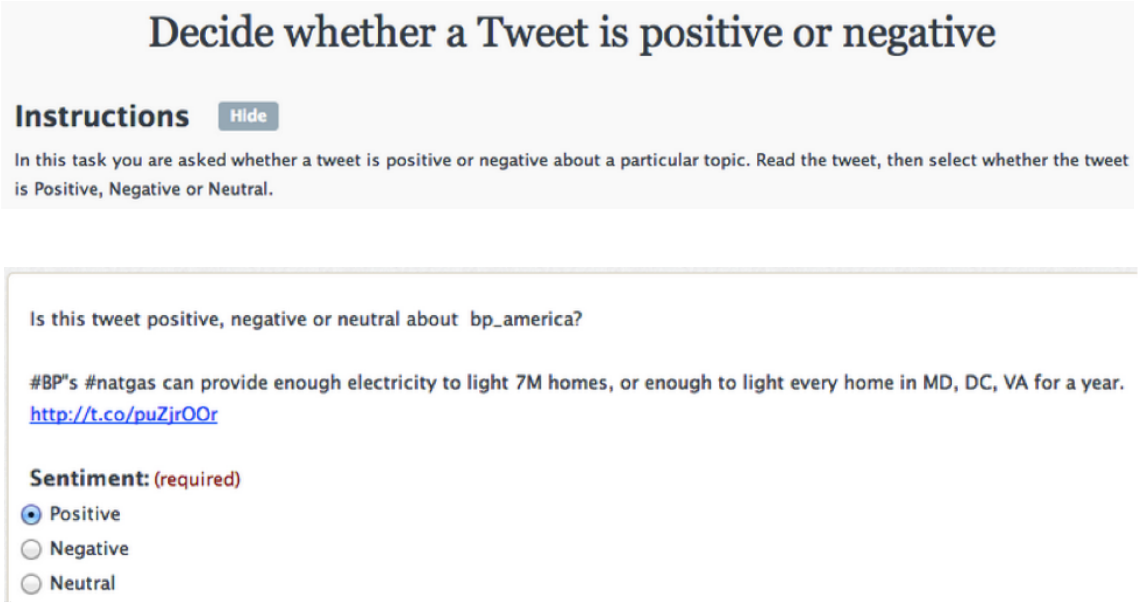


Figure 4.1: Example of AMT task page

The cost assigned to each task was $0.10 per tweet. In the literature [129, 132, 133], there were various experiments have been made to identify suitable payment for AMT with rates as low as $0.02 and as high as $0.50. Further, in some cases the cost could be increased based on the task scope and expected completion time. AMT allows verification of the qualification of AMT annotators for tasks assigned to them. For these experiments, annotators who are native English speakers based in the US and familiar with Twitter conventions were selected. This was verified using the results sheets that contained annotator id, country, region, city and their IP address. The manual classification of the data was carried out using AMT. However, just before all classification was completed, a restriction was placed on the use of AMT outside the US in 2013. Hence the use of another crowdsourcing platform called CrowdFlower was explored.

CrowdFlower

CrowdFlower[2] is a crowdsourcing platform which functions in a similar way to AMT. To use CrowdFlower, a job is posted with instructions on how to complete it and then assigned to

[2] http://crowdflower.com

annotators. The classification task was assigned on CrowdFlower in a similar way. Each tweet was allocated as a task to three annotators to classify as positive, negative or neutral, then decide which category a Twitter user belongs to based on defined groups options. Since categories may overlap, each Twitter user is classified in one group that considered the dominant category in the user description. Initially, single category is adequate to understand the main users' categories who are tweeting about oil companies. The actual polarity assigned to the tweet was the agreement of the three annotators' judgment. The cost of classifying a single tweet by three annotators was $0.30. Figure 4.2 shows a screenshot of the task on CrowdFlower; with the instructions and questions.

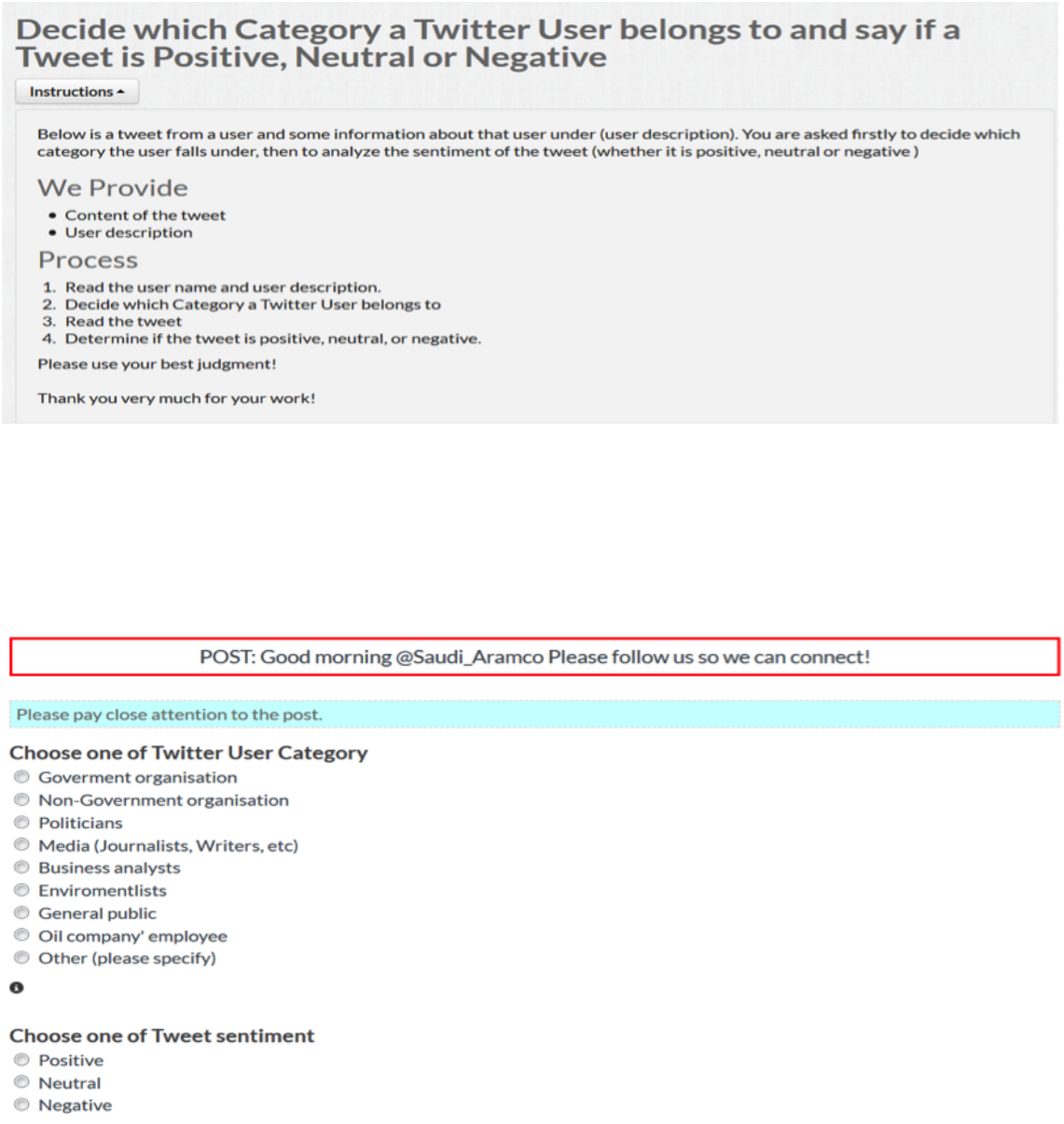


Figure 4.2: Example of CrowdFlower task page

4.2.1.2 Automated sentiment analysis technique (AlchemyAPI)

AlchemyAPI is a cloud platform that attempts to make Machine Learning algorithms available for general consumption. AlchemyAPI claims to have over 40,000 users which makes it easy for developers to integrate specific API tasks into their applications without the need for training data. Some of their APIs use billions of data for training that makes AlchemyAPI reliable for the purpose of research. AlchemyAPI has components for several machine learning tasks including sentiment analysis. In this research, AlchemyAPI was used classify sentiment expressed in the tweets of companies we investigated. The sentiment function exposed via AlchemyAPI returns results as a floating point number between -1 and 1, that represents the evaluation of the emotional content in the tweet (>0 for positive sentiment; <0 for negative sentiment and 0 for neutral). The function can process approximately 30 to 100 tweets per second. For approved academic users, Alchemy allows 30,000 sentiment analysis API calls per day. This member of API calls are appropriate for current sentiment analysis due to the low scale of oil companies' tweets and mentions.

4.2.2 Findings and Discussion

As mentioned earlier, existing experimentation involves analysis of the sentiment expressed in the tweets from Twitter accounts of two oil companies. The results from manual and automated sentiment analysis methods are compared to identify the efficiency and accuracy of both methods. To achieve that, the following research questions will be answered:

1. How close are the results of the automated method when compared to the manual gold standard?
2. What percentage accuracy is achievable, and in what application (manual or automated) is this sufficient?
3. Can the result of the automated method be considered accurate enough to be used as an approximation to the manual one?

4.2.2.1 Manual sentiment analysis finding

Summary results obtained from manual sentiment analysis can be found in Figure 4.3 and 4.4. As expected, companies usually do not publish negative tweets about themselves, hence the reason for 0% negative tweet in Figure 4.3. This shows some accuracy of manual process.

However, for the mentions classifications, the companies have a fair share of neutral and negative mentions according to Figure 4.4. This is most likely a result of these mentions generated by their customers and other stakeholders directly or indirectly affected by the activities of the company.

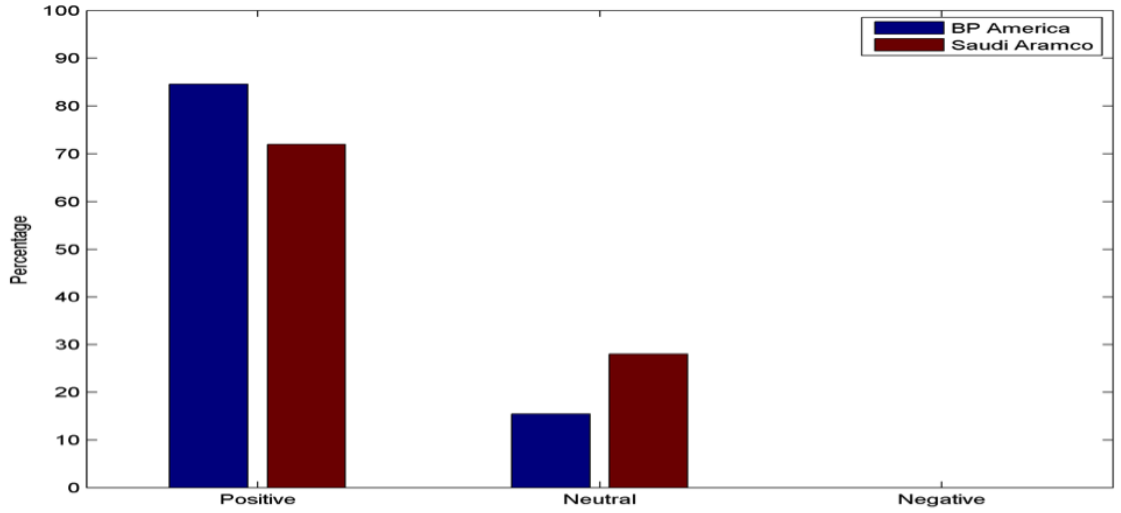


Figure 4.3: Manual sentiment analysis for companies' tweets (Dataset II-A)

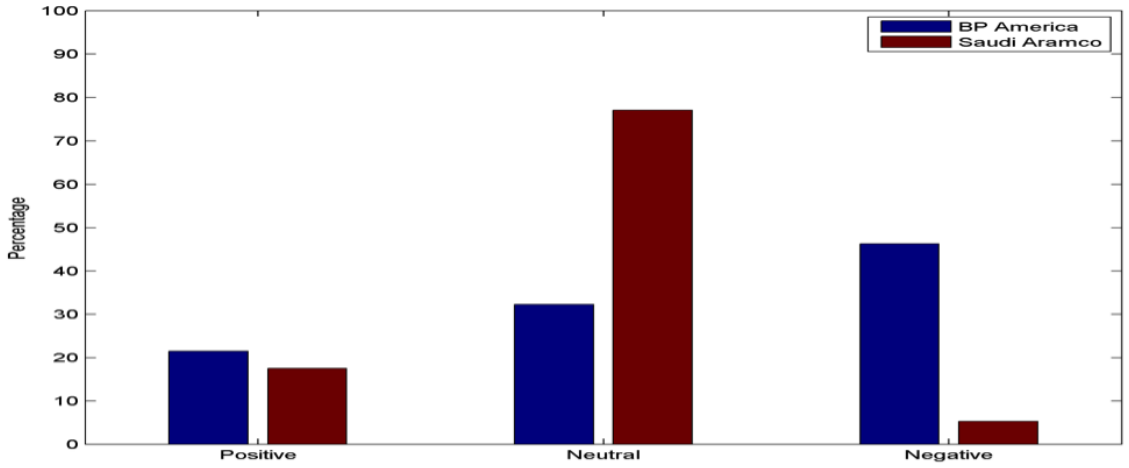


Figure 4.4: Manual sentiment analysis for companies' mentions (Dataset II-B)

As mentioned earlier, each annotator was allowed to categorize a tweet into either positive, neutral or negative. Tables 4.1 and 4.2 show a breakdown of the agreements among annotators for tweets and mentions. The agreements of the annotators are mainly positive as expected of tweets from any company (Figure 4.3). There were no tweets or mentions that had no agreement across three annotators. In all cases, at least two annotators agreed for a given tweet as seen in the tables.

The tweets that have two annotators agreeing account for more than 60% while the third disagreeing annotators were never incompatible completely. For instance, cases where two annotators agreed on positive, the third one choose neutral; and cases where negative was agreed, the third annotator settled for neutral as well. Therefore, we can conclude that manual classification is very effective.

Table 4.1: Manual sentiment agreement for BP_America

Manual sentiment agreement	**Number of Tweets**	**Percentage Agreement**	**Number of Mentions**	**Percentage Agreement**
Three annotators	373	37.3%	1034	34.5%
Two annotators	627	62.7%	1966	65.5%
No agreement	0	0%	0	0%
Total tweets and mentions	**1000**	**100%**	**3000**	**100%**

Table 4.2: Manual sentiment agreement for Saudi_Aramco

Manual sentiment agreement	**Number of Tweets**	**Percentage Agreement**	**Number of Mentions**	**Percentage Agreement**
Three annotators	341	34.1%	1186	39.5%
Two annotators	659	65.9%	1814	60.5%
No agreement	0	0%	0	0%
Total tweets and mentions	**1000**	**100%**	**3000**	**100%**

4.2.2.2 Automated sentiment analysis results finding

AlchemyAPI was applied to both companies' datasets that have been analysed via the manual platforms to identify the sentiment of tweets and mentions automatically. Figure 4.5 and 4.6 shows the polarity distribution of tweets and mentions for both companies using the AlchemyAPI for automatic classification.

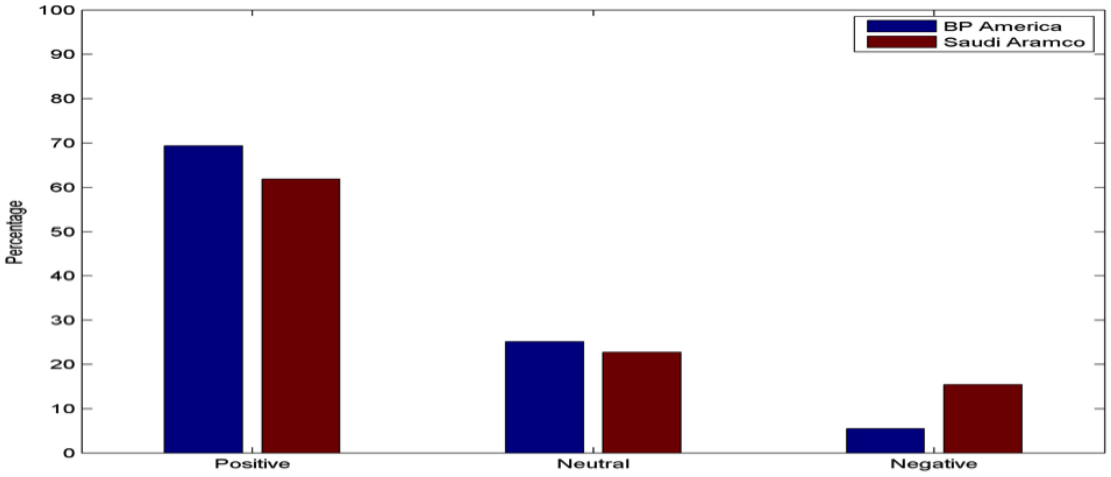


Figure 4.5: Automated sentiment analysis for companies' tweets (Dataset II-A)

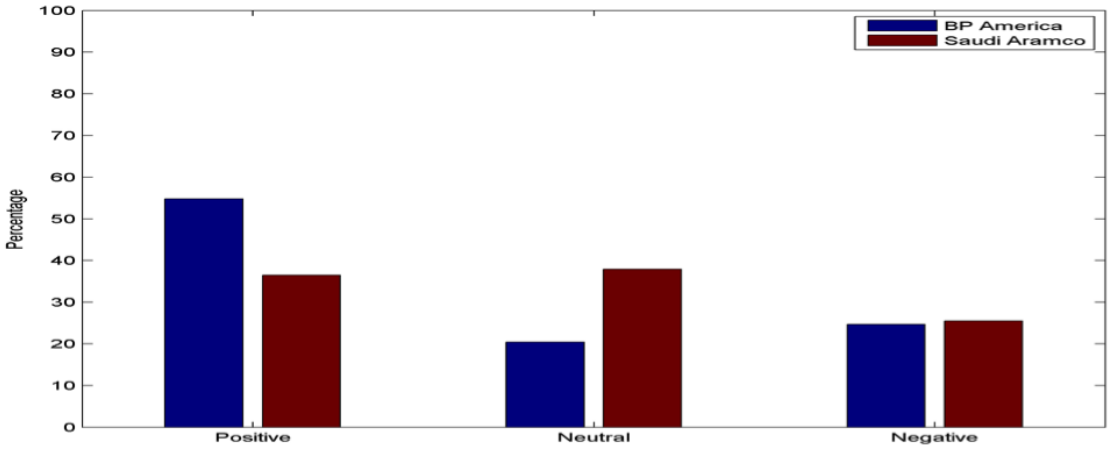


Figure 4.6: Automated sentiment analysis for companies' mentions (Dataset II-B)

From these results, it can be observed that BP_America still had more positive tweets when compared to Saudi_Aramco which indicates the automated approach is correctly identified the positive sentiment in tweets. On the other hand, there are more negative sentiment conveyed in the mentions when compared to the company tweets. This may be explained by the reasoning that, while companies are more likely to tweet positive things, regular users will express their opinions on the activities of the oil companies without constraint. However, the actual variation between manual and automated is quite high.

Tables 4.3 and 4.4 present the comparisons between manual and automated techniques from the results obtained for both companies' tweets and mentions. Unlike the manual method, the automated classification detected some of the positive tweets as either neutral or negative.

Table 4.3: Agreement between automated and manual sentiment for companies' tweets

Methods	**BP_America**			**Saudi_Aramco**		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Manual sentiment	84.6%	15.4%	0%	72%	28%	0%
Automated sentiment	69.4%	25.1%	5.5%	61.9%	22.7%	15.4%
Total sentiment polarities agreement between methods	**68.1%**			**52.2%**		

As seen in the Table 4.3, the total sentiment polarities agreement between both automated and manual methods were calculated generally. After that, sentiment polarities percentage of both methods were calculated separately by counting how many positive, neutral and negative were marked by each method. The automated technique gives comparable sentiment classification to the manual approach in general: 68.1% of BP_America and 52.2% of Saudi_Aramco. Furthermore, it can be observed that the automated sentiment analysis technique is not as accurate as the manual technique so that the difference is 32% and 48% for each company respectively. The most evident error can be found in the classification of negative tweets where 5.5% of the tweets were wrongly classified as negative for BP_America and 15.4% were incorrectly classified as negative for saudi_Aramco.

Table 4.4: Agreement of automated and manual sentiment for companies' mentions

Methods	**BP_America**			**Saudi_Aramco**		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Manual sentiment	21.46%	32.23%	46.26%	17.5%	77.1%	5.3%
Automated sentiment	54.8 %	20.4%	24.76%	36.5%	37.9%	25.5%
Total sentiment polarities agreement between methods	**43.86%**			**43.7%**		

Table 4.4 shows that both methods give generally similar sentiment classification: 43.86% and 43.7% for BP_America and Saudi_Aramco respectively, which is less than 50% of the total 3000 mentions for each company. Since mentions are from external sources to the companies, they tend to contain more sarcasm than tweets originating from the companies directly. Below are a couple of examples of sarcasm in mentions from external users directed at BP_America.

Hence, this is the reason why the automated method has erroneously classified most of the negative and neutral mentions as positive tweets. This implies that the automated technique is less reliable with mentions due to the higher degree of sarcasm introduced by the peculiarities of mentions from different sources.

4.2.3 Results Evaluation

4.2.3.1 Statistical Analysis

To comparatively evaluate manual and automated sentiment analysis techniques, statistical analysis was carried out to compare the results of both techniques. All tests (Chi-square and Fisher exact) were carried out using IBM SPSS.

The null and alternative hypotheses tested are:

H_0: **There is no significant difference between manual and automated sentiment analysis tools.**

H_a: **There is significant difference between manual and automated sentiment analysis tools.**

The level of statistical significance determines whether to reject the null hypothesis and accept the alternative one or fail to reject it if there is no evidence to prove it. Chi-square test is appropriate to examine the significance of relationships between two nominal (categorical) variables with large data size [134]. It has been used since both sentiment analysis methods were utilised on the same data. When Chi-square test is run on the data, a p-value is obtained which indicates the degree of agreement or difference between variables of interest. However for a small data size below 10 in any cell of data table, the Fisher exact test was used to compute the p-value [135]. The Fisher exact test has the same concept as Chi-square test but is more effective and accurate when the analysis involves a smaller dataset. The implications of different p-value ranges are shown below:

P > 0.05 No significant difference
P < 0.05 (0.01 to 0.05) Minimum significant difference
P < 0.01 (0.001 to 0.01) Higher significant difference
P < 0.001 (0.0001 to 0.001) Strong significant difference

AlchemyAPI renders its output as floating-point numbers between -1 and 1 that represents its evaluation of the emotional content in the text (> 0 = positive sentiment; < 0 = negative sentiment; 0 = neutral). The dataset was normalised to represent negative, neutral and positive sentiment as -1, 0 and 1 respectively. The results of our statistical analysis can be found in Tables 4.5 and 4.6

Table 4.5: Statistical test for companies' tweets

Automated sentiment	**BP_America tweets**			**Saudi_Aramco tweets**		
	Manual sentiment		**Total**	**Manual sentiment**		**Total**
	.00	1.00		.00	1.00	
-1.00	9	46	55	57	97	154
.00	66	185	251	63	164	227
1.00	79	615	694	160	459	619
Total	154	846	1000	280	720	1000
Chi-square p-value	**.000**			**.022**		

Table 4.6: Statistical test for companies' mentions

Automated sentiment	**BP_America mentions**				**Saudi_Aramco mentions**			
	Manual sentiment			**Total**	**Manual sentiment**			**Total**
	-1.00	.00	1.00		-1.00	.00	1.00	
-1.00	533	149	62	744	32	678	55	765
.00	169	322	121	612	51	948	140	1139
1.00	687	496	461	1644	78	687	331	1096
Total	1389	967	644	3000	161	2313	526	3000
Chi-square p-value	**.000**				**.000**			

From the Chi-Square test results above, a p-value of (.000) was obtained for @BP_America tweets and (.000) for mentions. This implies that the results from both techniques have a high

significance difference relationship. The same goes for the @saudi_Aramco tweets where the p-value is (.022) and (.000) for mentions. The p-values indicate that there are significant differences between sentiment methods which mean the null hypothesis was rejected. This implies that results from the automated technique requires improvement or cannot be relied on as an accurate measure.

4.2.3.2 Evaluation metrics

To further investigate the nature of differences in the sentiment analysis results from both techniques, we adopt the standard performance measures for text categorization, such as precision, recall, F-measure and accuracy. In the context of information retrieval, precision and recall are defined in terms of a set of retrieved and relevant documents. Buckland and Gey [136] describe them as following: "Recall is a measure of effectiveness in retrieving (or selecting) performance and can be viewed as a measure of effectiveness in including relevant items in the retrieved set. Precision is a measure of purity in retrieval performance, a measure of effectiveness in excluding non-relevant items from the retrieved set".

The F-measure is the harmonized mean of precision and recall, while the accuracy is the total correctly classified tweets normalized by the total number of tweets [118].

In this work we define them as follow:

Let N_A^+ and N_M^+ denote the sets of tweets classified as having positive sentiment by automated and manual techniques respectively. Similarly, N_A^0, N_M^0 and N_A^-, N_M^- represents neutral and negative tweets respectively. Considering the human classification from crowdsourcing platforms as the ground truth, the precision and recall can be defined for the automated classification of positive tweets as:

$$P^+ = \frac{|N_A^+ \cap N_M^+|}{|N_A^+|} \tag{4.1}$$

$$R^+ = \frac{|N_A^+ \cap N_M^+|}{|N_M^+|} \tag{4.2}$$

$$F^+ = 2 * \left(\frac{P^+ * R^+}{P^+ + R^+}\right) \tag{4.3}$$

P^0, R^0, F^0, P^-, R^- and F^- are defined similarly.

$$Accuracy = \frac{\text{Total correctly classified tweets (CCT)}}{\text{Total number of tweets (NT)}} \tag{4.4}$$

where CCT is the total tweets accurately classified and NT is the total number of tweets.

Table 4.7: BP_America and Saudi_Aramco tweets performance measures

	BP_America tweets			**Saudi_Aramco tweets**		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.886	0.262	0	0.742	0.278	0
Recall	0.726	0.428	0	0.638	0.225	0
F-Measure	0.798	0.325	0	0.686	0.249	0
Accuracy	**0.681**			**0.522**		

Table 4.8: BP_America and Saudi_Aramco mentions performance measures

	BP_America mentions			**Saudi_Aramco mentions**		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.280	0.526	0.716	0.302	0.832	0.042
Recall	0.716	0.333	0.384	0.629	0.410	0.199
F-Measure	0.403	0.408	0.500	0.408	0.550	0.070
Accuracy	**0.438**			**0.437**		

These relevancy measure metrics obtained are in agreement with the results obtained through the statistical tests and further prove that the automated classifier is unsuitable for the classification of mentions. This is because mentions contain more sarcasm that the automated tool wrongly classifies in most cases. Table 4.7 confirms that because both companies do not publish negative or sarcastic tweets about themselves, it is easier to measure their sentiment (0.79 and 0.68 F-measure). Table 4.7 also shows that the positive and negative classes (the classifier does not correctly classify any tweets as negative as there were none) can be easily discriminated while the neutral class was the most challenging (0.32 and 0.24 F-measure). The poor reliability of the classifiers is more evident in Table 4.8. For example, while similar result to Table 4.7 is expected, i.e., reliable (or high precision-recall statistics) for the positive and negative classes is found, this is evidently not the case. For example, the F-measure is very low (0.40) for the positive class of BP_America and less than (0.1) for the negative class of Saudi_Aramco in mentions. A possible explanation may be the use of layman terms and expressions such as slang

and sarcasm in mentions. For example, tweet using sarcasm such as "Thanks to @BP_America we now have to teach shrimp braille before we eat them" is miscategorised as positive. Similarly, AlchemyAPI performs poorly in identifying positive mentions.

4.2.4 Sentiment Analysis as Binary Classification Task

Further investigation of the automated technique gives more insight into the nature of misclassification. Motivated by the observation that manual classifiers only differ in interpreting neutral tweets, the classification problem was reduced to a binary classification task such that the problem can be handled in two distinct ways, namely identifying each tweet simply as positive or non-positive by merging neutral with the negative class, and as negative or non-negative by merging neutral with the positive class. This approach was used on purpose to investigate if the classification errors can be reduced for the automated technique by simplifying the problem. In addition, analysing sentiment as positive or negative is of greater importance when evaluating public perception of a brand. This merging helps to eliminate the noise added to the dataset by the neutral class. In this section, the performance measure was defined similarly to subsection 4.2.3.2.

Let N_A^1 and N_M^1 denote the sets of tweets classified as having non-positive sentiment (neutral + negative) by automated and manual techniques respectively.

$$P^1 = \frac{|N_A^1 \cap N_M^1|}{|N_A^1|} \tag{4.5}$$

$$R^1 = \frac{|N_A^1 \cap N_M^1|}{|N_M^1|} \tag{4.6}$$

$$F^1 = 2 * \left(\frac{P^1 * R^1}{P^1 + R^1}\right) \tag{4.7}$$

Non-negative sentiment (neutral + positive) is described by N_A^2 for automated and N_M^2 for manual technique and P^2,R^2 and F^2 are defined similar to non-positive sentiment class.

$$Accuracy = \frac{\text{Total correctly classified tweets (CCT)}}{\text{Total number of tweets (NT)}} \tag{4.8}$$

The results obtained for each of the binary classification tasks are discussed below.

4.2.4.1 Positive vs. non-positive tweets

The classification of tweets was considered as positive or non-positive (i.e. either negative or neutral), the aim to make it easier to retrieve positive tweets from all company tweets. Reducing the problem to this form makes it easy to identify tweets belonging to the class of interest. The Chi-square statistical test results are presented in Table 4.9 for both companies.

Table 4.9: Positive vs. non-positive tweets

Automated sentiment	**BP_America tweets**			**Saudi_Aramco tweets**		
	Manual sentiment		**Total**	**Manual sentiment**		**Total**
	Positive	**Non-positive**		**Positive**	**Non-positive**	
Positive	615	79	694	459	160	619
Non-positive	231	75	306	261	120	381
Total	846	154	1000	720	280	
Chi-square p-value	**.000**			**.053**		

From the results obtained, it can be observed that the p-values indicate that the classification of the two methods differ significantly in both companies' tweets. The null hypothesis was rejected with BP_America while accepted with Saudi_Aramco. This implies that by reducing the problem to a binary classification task of positive vs. non-positive, both techniques produce results that are comparable and usable. Table 4.10 shows the precision and recall metrics obtained.

Table 4.10: Companies (positive vs. non positive) tweets performance measures

	BP_America tweets		**Saudi_Aramco tweets**	
	Positive	**Non-positive**	**Positive**	**Non-positive**
Precision	0.886	0.245	0.741	0.314
Recall	0.726	0.487	0.637	0.428
F^1-Measure	0.798	0.326	0.685	0.363
Accuracy	**0.644**		**0.579**	

The classifier shows a precision of (88%) in identifying the positive tweets for BP_America while a precision of (74%) was obtained for saudi_Aramco. The precision relevancy measure looks good but it does not tell the complete story due to class imbalance - a case where a

classification category is under represented and other majorly represented. Using BP_America as a case study, 84.6% of the tweets are positive as presented in table 4.3. This implies that a classifier that always predicts the positive class will have a precision of 84.6% which is 3.4% less than the automated sentiment classifier (88%). A relevancy measure that combines the precision with the recall (F1-measure) is therefore the ideal measure of the effectiveness of the algorithm. For example, considering the F1-measure, the classifier was able to reliably discriminate the positive classes: BP_America (0.80) and saudi_Aramco (0.69), while the non-positive classes were notably more challenging: 0.22 and 0.26 respectively.

4.2.4.2 Negative vs. non-negative tweets

The second binary classification task makes it easy and more efficient to retrieve negative tweets and non-negative (positive and neutral) by merging the neutral and positive classes together. Negative tweets are often the focus of brand monitoring tools because they help give an opinion of the areas that need improvement in an organisation. Hence, a more efficient way of identifying them will prove useful. The results obtained can be seen in Table 4.11.

Table 4.11: Statistical test for companies' (negative vs. non-negative) tweets

Automated sentiment	**BP_America tweets**			**Saudi_Aramco tweets**		
	Manual sentiment		**Total**	**Manual sentiment**		**Total**
	Negative	**Non-negative**		**Negative**	**Non-negative**	
Negative	0	55	55	0	154	154
Non-negative	1	944	945	1	845	846
Total	1	999	1000	1	999	1000
Fisher's Exact Test p-value	**.809**			**.669**		

During statistical analysis, the Fisher exact test was also introduced since it is more effective than the Chi-square when the number of data points in one of the results being compared is less than 10, and especially when it is ≤ 1 [135]. The p-value obtained shows that the results for the automated and manual technique do not differ significantly which accept the null hypothesis. That indicates the neutral sentiment has an effective role on the accuracy. Since human classification represents the ground truth, the opinions on the company tweets were formed based on the manual classification results.

Table 4.12: Companies' (negative vs. non negative) tweets performance measures

	BP_America Tweets		**Saudi_Aramco Tweets**	
	Negative	**Non-negative**	**Negative**	**Non-negative**
Precision	0	0.998	0	0.998
Recall	0	0.944	0	0.845
F^1 - Measure	0	0.971	0	0.915
Accuracy	**0.944**		**0.845**	

Table 4.12 shows that reducing the problem to a binary classification task for tweets: negative vs. non-negative, gave a very reliable results. Notably, 0.97 and 0.92 $F^2 - measure$ resulted for the non-negative classes for BP_America and Saudi_Aramco tweets respectively. The high rate of $F^2 - measure$ indicates that merging neutral with positive class is effective way to increase the accuracy.

As expected, it can be observed that all tweets from the company accounts were positive and neutral which shows that the companies are keen on painting a good picture of themselves on social media and they understandably start conversations that enhance their brand. To get a better picture of how the companies are perceived on social media, mentions were analysed as well. The following section discusses the results obtained after analysing the mentions of both companies.

4.2.4.3 Positive vs. non-positive mentions

Companies' mentions are binary classified as the tweets. The results obtained for both companies can be seen in Table 4.13. The Chi-square statistical test results are also presented in the table.

Table 4.13: Statistical test for companies' (positive vs. non positive) mentions

Automated sentiment	BP_America mentions			Saudi_Aramco mentions		
	Manual sentiment		Total	Manual sentiment		Total
	Positive	Non-positive		Positive	Non-positive	
Positive	461	1183	1644	331	765	1096
Non-positive	183	1173	1356	195	1709	1904
Total	644	2356	3000	526	2474	3000
Chi-square p-value	**.000**			**.000**		

From the result obtained, it can be observed that the p-values are (.000) for both companies which indicates the variables differ significantly and reject the null hypothesis. Mentions classification are affected severely by layman terms and expressions (e.g. sarcasm); this was unlike the company' tweets results, where the language used was not ambiguous. Table 4.14 presents the performance metrics obtained which show that reducing the classification problem from a multi-class problem seen in Table 4.8 to a binary class problem can result in a more reliable results. However, the positive class seems to still be a challenge for the automated method.

Table 4.14: Companies' (positive vs. non positive) mentions precision and recall

	BP_America Mentions		Saudi_Aramco Mentions	
	Positive	Non-positive	Positive	Non-positive
Precision	0.280	0.865	0.302	0.898
Recall	0.716	0.498	0.629	0.691
F^1 - Measure	0.403	0.632	0.408	0.781
Accuracy	**0.545**		**0.680**	

4.2.4.4 Negative vs. non-negative mentions

As in the case of company tweets, the second binary classification task of negative vs. non-negative for companies' mentions. The statistical results obtained can be seen in Table 4.15.

Table 4.15: Negative vs. non-negative mentions

Automated sentiment	BP_America mentions			Saudi_Aramco mentions		
	Manual sentiment		Total	Manual sentiment		Total
	Negative	Non-negative		Negative	Non-negative	
Negative	533	211	744	32	733	765
Non-negative	856	1400	2256	129	2106	2235
Total	1389	1611	3000	161	2839	3000
Chi-square p-value	**.000**			**.092**		

From the Chi-square test results above, a p-value of (.000) was obtained for BP_America. This implies that the results from both manual and automated techniques have a high significant difference which reject the null hypothesis. Otherwise, Saudi_Aramco have accepted the null hypothesis with a p-value of (.092) which indicates there is no significant difference when the neutral is merged with the positive class between both techniques. This difference can be attributed to the use of sarcasm and vulgar words, which is more common in America where BP_America is based, rather than in the more conservative host region of Saudi_Aramco.

Table 4.16: Companies' (negative vs. non negative) mentions precision and recall

	BP_America Mentions		Saudi_Aramco Mentions	
	Negative	Non-negative	Negative	Non-negative
Precision	0.716	0.621	0.041	0.942
Recall	0.383	0.869	0.198	0.741
F[1]-Measure	0.499	0.724	0.069	0.830
Accuracy	**0.644**		**0.712**	

Similarly, Table 4.16 shows that reliable results can be obtained by reducing the problem to a binary classification problem. However, while the classification of the non-negative class shows good performance, the negative class shows unreliable performance with 0.50 and 0.07 $F^2 - measure$ for BP_America and saudi_Aramco respectively. It can be concluded from this section that the sentiment methods performed better with binary classification when tweets and mentions are classified as negative and non-negative (positive + neutral). This indicates that misclassification occurred more with positive and neutral classes.

4.3 Sentiment Analysis for Different User Groups

Section 4.2 identified the use of sarcasm as a potential problem for automated sentiment analysis techniques. This section hypothesizes that, since different types of user may be less disposed to using sarcasm, that some groups will show more accuracy than others. To further understand the nature of the sentiment expressed in the mentions and identify the nature of the source, the Twitter account of users mentioning the oil companies were classified into one of eight main groups which are general public, media, environmentalist, business analysts, oil company employee, politicians, government and non-government organisations.

The aim of the classification is to establish a relationship between the mentions and the source of the tweet. The classification was carried out using the manual technique and the features exposed to the annotators on CrowdFlower were the Twitter handle (@username) and Twitter account biography (profile description). The Twitter user profiles often provide useful information especially if the user belong to professional body. In contrast, users from general public sometimes can use nicknames and provide ambiguous description in their profiles which affect the classification accuracy. The classification details are described previously in subsection 4.2.1.1. The groups are listed in Table 4.17.

Table 4.17: BP_America and Saudi_Aramco users' groups

No.	Category	BP users number	Aramco users number
1	General public	784	2354
2	Media	732	112
3	Environmentalist	380	25
4	Business analysts	134	63
5	Oil company employee	34	190
6	Politicians	29	42
7	Government organisation	303	23
8	Non-government organisation	604	191
	Total	30000	30000

Users classified as part of 'General public' are owners of Twitter accounts with an empty biography or ambiguous description. The 'Media group' is used to identify accounts belonging to journalists, writers or news agencies. Accounts classified under the 'Environmentalists group' are organisations that condemn pollution activities of oil companies e.g. @Greenpeace. Employees of oil companies that indicate their employer information in their account description

were all classified into one group called ‘oil company employee’. Since oil companies are known to have a significant contribution on the economy of host countries, a group was created to identify ‘Business Analysts’ who are often employees of companies like McKinsey and provide opinions on the economic implication of oil company activities. Politicians, government organisations and non-government organisation were also classified into individual groups based on their description data and prior public knowledge.

4.3.1 BP_America User Groups

The efficiency of both sentiment analysis methods was investigated on each of the user groups for tweets mentioning BP_America. We used Chi-square test and Fisher’s exact tests were used. General public and media groups showed a difference between manual and automated with p-value (.000) as did government and non-government organisations. On the other hand, environmentalists group represent p-value (.743), business analysts (.109), oil company employees (.007) and politicians (.113). From the p-values obtained, there was a significant difference in the result obtained from both classification techniques for four of the user groups, namely, general public, media, government organisations and non-government organisation. Table 4.18a and 4.18b shows the performance measures of these user groups with a p-value indicating a significant difference.

Table 4.18: BP_America user groups with significant difference in using sentiment tool
(a)

	General Public			**Media**		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.177	0.404	0.894	0.253	0.680	0.575
Recall	0.784	0.342	0.498	0.640	0.356	0.426
F[1]-Measure	0.289	0.371	0.640	0.363	0.467	0.489
Accuracy	**0.498**			**0.434**		

(b)

	Government Organisation			Non-goverment Organisation		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.505	0.484	0.389	0.435	0.549	0.628
Recall	0.664	0.378	0.273	0.803	0.307	0.362
F[1]-Measure	0.574	0.424	0.321	0.564	0.394	0.459
Accuracy	**0.4785**			**0.491**		

Table 4.18 generally shows poor performance with the classifier achieving less than 0.50 in F-measure for the majority of classes across different user groups. Yet, across these groups it can be seen that the negative class was easier to discriminate for the General public and Media, while the classifier showed better performance on the positive class from Government and Non-government organisations. Conversely, classification results from the four other groups (environmentalists, business analysts, politicians and oil company employees) indicate that there is no significant difference. The precision, recall and F-Measure for each of these groups can be found in Table 4.19a and 4.19b .

Table 4.19: BP_America user groups with non-significant difference in using sentiment tool
(a)

	Environmentalist			Business analysts		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.053	0.121	0.844	0.254	0.622	0.205
Recall	0.736	0.142	0.199	0.608	0.318	0.304
F[1]-Measure	0.099	0.131	0.322	0.358	0.421	0.245
Accuracy	**0.221**			**0.365**		

(b)

	Oil Company Employee			Politicians		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.833	0.083	0.750	0.188	0.714	0.830
Recall	0.576	0.250	0.750	1.000	0.333	0.454
F^1 - Measure	0.681	0.125	0.750	0.315	0.454	0.588
Accuracy	**0.558**			**0.365**		

From the results, it is evident that most of the groups have good precision and recall values, which is comparable to the results from the manual method. This is because such accounts were handled by organisations who post tweets in plain terms without the use of ambiguous expressions or slangs that introduce noise into the automated classification system. Further, these scores are derived from the evaluation of a small number of labels which makes these results unreliable.

4.3.2 Saudi_Aramco User Groups

In the same way of analysing BP_America user groups, Saudi_Aramco user groups were also analysed to investigate the efficiency of both sentiment analysis techniques. In terms of statistical analysis results, different different p-values were obtained in the user groups. General public showed a difference between manual and automated with p-value (.000), as did non-government organisations (.003). The nature of inaccuracy is expected in these groups because they can use indirect expression about Aramco Company. On the other hand, media group represents(.497), environmentalists and business analysts' groups represent (.473), oil company employees (.007), politicians (.489) and government organisations group (.837). From the p-values obtained, there was a significant difference in the result obtained from both techniques for general public and non-government organisation while there was non-significant difference for media, government organisations, environmentalists, business analysts, oil company employees and politicians. For example, media in Saudi Arabia is more respectful and guided from the government. Table 4.20 shows the performance measures of user groups with a p-value indicating a significant difference.

Table 4.20: Saudi_Aramco user groups with significant difference in using sentiment tool

	General Public			**Non-government Organisation**		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.308	0.839	0.037	0.283	0.780	0.108
Recall	0.607	0.423	0.198	0.800	0.226	0.400
F[1]-Measure	0.409	0.563	0.063	0.418	0.351	0.170
Accuracy	**0.443**			**0.440**		

Table 4.20 shows poor classification performance with notably unreliable results (e.g., the negative class across both user groups). In addition, the few number of labels used for the evaluation, for the non-government organisation group in particular, makes it problematic to draw any conclusion.

Table 4.21a, 4.21b and 4.21c showed that the performance measures for each of the six other groups (media, government organisations, environmentalists, business analysts, oil company employees and politicians) indicate that there is no significant difference.

Table 4.21: Saudi_Aramco user groups with non-significant difference in using sentiment tool

(a)

	Media			**Environmentalist**		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.227	0.820	0.055	0.500	0.500	0
Recall	0.588	0.465	0.142	0.666	0.416	0
F[1]-Measure	0.327	0.594	0.080	0.571	0.454	0
Accuracy	**0.464**			**0.492**		

(b)

	Oil company employee			Business analysts		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.313	0.848	0.048	0.277	0.812	0
Recall	0.666	0.400	0.181	0.714	0.500	0
F^1-Measure	0.426	0.543	0.076	0.400	0.619	0
Accuracy	**0.442**			**0.492**		

(c)

	Government Organisation			Politicians		
	Positive	**Neutral**	**Negative**	**Positive**	**Neutral**	**Negative**
Precision	0.250	0.625	0	0.200	0.888	0.125
Recall	0.333	0.294	0	1.000	0.242	0.250
F^1-Measure	0.285	0.400	0	0.333	0.380	0.166
Accuracy	**0.304**			**0.333**		

Table 4.21a, 4.21b and 4.21c show notably poor classification performance as a result of these scores are derived from the evaluation of a small number of labels which makes these results unreliable.

4.4 Conclusion

Sentiment can be classified manually as well as automatically. Manual classification is however tedious and cumbersome when large datasets need to be classified. Automated classification can be achieved using commercially available tools such as the AlchemyAPI used in this project. Automated methods allow sentiment classification of large dataset within a short time frame. However, automated methods tend to not be accurate enough for all domains. In this chapter, it has been shown that automated tools can also be as effective as the manual method for sentiment analysis of tweets under certain conditions. Some of these cases are listed below as related to tweets from the Twitter page of the two oil companies analysed in this project:

- Tweets that are originated from the company

- Tweets originating from Governmental organisations and other professional bodies
- Tweets from different countries where more or less use of sarcasm and layman terms occurs

Two oil companies operating in different geographical regions were considered in this project, BP_America based in America and Saudi_Aramco based in Saudi Arabia. For the reasons listed above, the automated tool was more effective and accurate for companies' tweets and mentions from some users groups. The inferences were drawn from result produced by a variety of tests from IBM SPSS statistical tool. The outcome of this work will enable users to use automated sentiment analysis tools in efficient way.

In the next chapters machine learning classifiers and clustering algorithm are used to predict user groups and cluster these groups based on their similarities. More details are provided about the level of sentiment accuracy within the groups. Then determining which groups are appropriate to evaluate by automated sentiment method and which groups faced difficulties with it and should use manual method.

Chapter 5

User Categorization Using Machine Learning

Introduction

Text categorization is a well-known machine learning method for understanding text. It can be applied in many forms, such as authorship detection and text mining by extracting useful information from documents to assign one or more predefined categories to text documents automatically [84, 137]. In this research, the terms "documents" and "tweets" refer to a similar concept. Considering each tweet as a document, the text-categorization concepts, such as tokenization, stemming, term-frequency, and document-frequency [138] were used to encapsulate a flexible representation of the problem, making it easy for text categorization algorithms to be efficiently applied to this problem. The aim of this chapter is to categorize incoming tweet users automatically into a number of pre-defined classes. As was pointed out in the previous chapter, the automated tools can not predict the sentiment very well that was a motivation to investigate how easier to predict user groups. Classification of users into different groups helps to predict the level of accuracy in sentiment by understanding which of automated or manual methods suits each groups. Machine learning was investigated as a method to identify those who posted tweets into categories of user. The task is performed by extracting key features from tweets and subjecting them to a machine learning classifier. This work can be termed a *multi-class* categorization problem where multi-class refers to the problem where the input documents can be classified into more than two classes or categories [139]. Classification also can suffer from class imbalance, whereby a class that has more training data is more likely to be predicted as an output class. Multi-class categorization is typically more difficult than binary-class classification (with only two output classes or categories). Current problem specifically concerns users who tweet about oil companies, most of data are noisy enough to affect the accuracy, however the analytical techniques used here are still capable of providing structured and valuable information for oil companies.

Section 5.1 discusses the categorization experiment setup including pre-processing descriptions, applied methods and evaluation. Section 5.2 gives detailed classification findings and discussion. The chapter is concluded in section 5.3.

5.1 Experimental Methodology

This section gives an overview about the primary analysis and the methods that have been used to build categorization experiments. Figure 5.1 shows the process of the experiment in general.

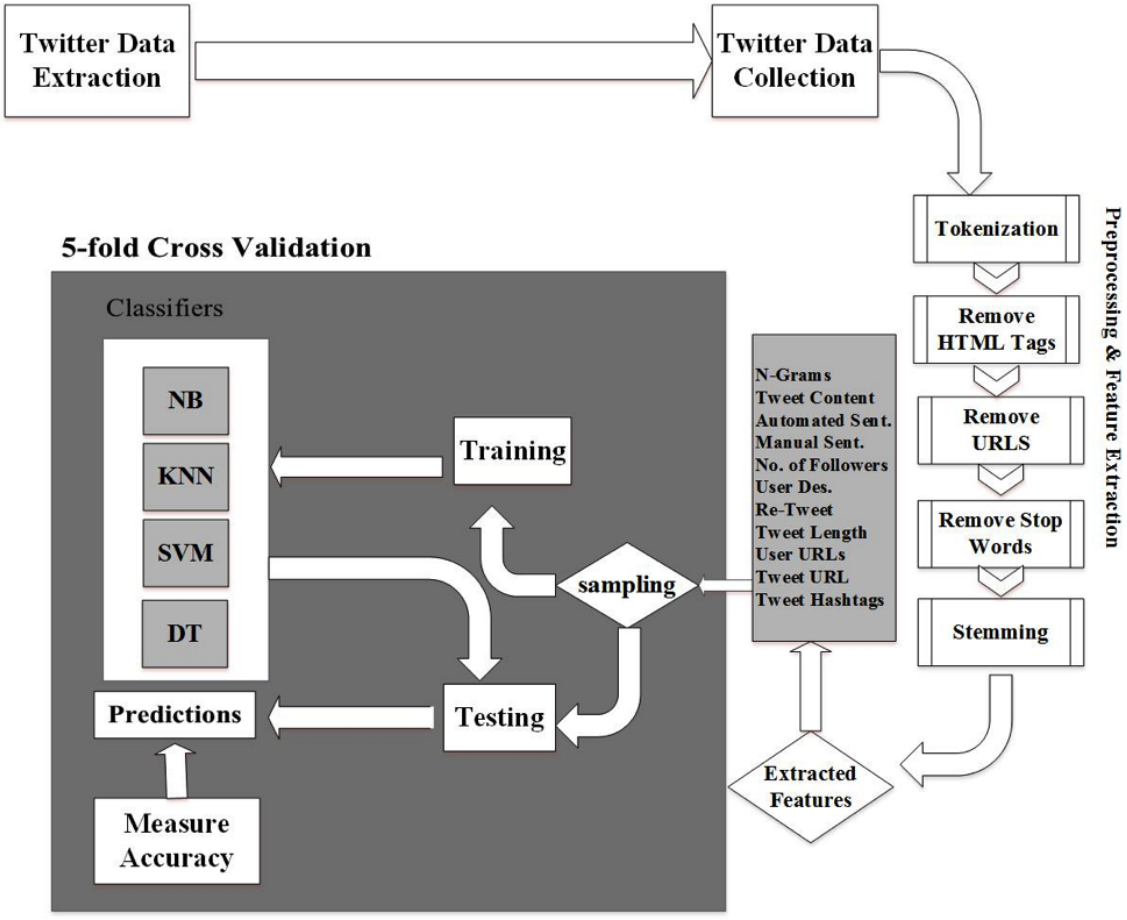


Figure 5.1: Categorization process

5.1.1 Primary Analysis

As described in Chapter 4 crowdsourcing platform was used to categorise the users who mentioned each company name in their tweets to identify the main user groups into eight categories: general public, media, environmentalists, politicians, business analysts, oil company employee,

government and non-government organisations. Then the sentiment within the categories by both automated and manual techniques was analysed, the details of these methods were discussed previously in chapter 4. As was pointed out in the previous findings, the sentiment analysis was not accurate within the role of types of users. These results show sarcasm and layman terms as a potential mechanism behind inaccurate classification, and it is reasonable to assume these affect some groups more than others. That was motivated to predict the types of users and investigate the groups' structure by using machine learning classifiers based on different features.

5.1.2 Pre-Processing

This section describes the pre-processing steps that are applied to the tweet datasets to enable classification.

5.1.2.1 Feature extraction

Feature extraction techniques aim to find the specific pieces of data in natural language documents [137], which are used for building (training) classifiers. Documents usually consist of string characters. Machine learning algorithms (e.g. Text Categorization algorithms) cannot work with these strings directly. They have been converted into a format which is suitable for the machine learning classifiers. The sequence of steps that has been performed to carry out this task is, as follows:

1. Convert the documents into tokens' sequences of letters and digits.
2. Perform the following modifications
    - Remove HTML and other tags, e.g. Author tag (@), hash tag (#)
    - Remove URLs
    - Remove stop words
    - Perform stemming

Stop words are frequently occurring words that carry no (or very little) information; it is usual in machine learning to remove them before feeding the data to any learning algorithms. Removing stop words will not affect the results because tweet semantic for sentiment analysis is not considered in this experiment. Hashtags and URLs should be removed, because they can confuse the classifier with irrelevant information. On the other hand, URLs and Hashtags have

been used as a boolean values. Stemming eliminates the case and inflection information from a word and maps them into the same stem. For example, the words *categorization*, *categorized* and *categories* all map into the same root stem 'category'. As an example, the concepts of feature extraction were applied to the following tweet (from the dataset) one by one.

"*RT @BP_America: Did you know the first service stations opened around 1910? Self-service stations did not become the norm until 1970s: http://t.co/4obuMaCS*"

The given tweet is processed and after each step yields the following:

After tokenization

"*RT @BP_America***:** *Did you know the first service stations opened around 1910***?** *Self-service stations did not become the norm until 1970s***:** *http://t.co/4obuMaCS*"

After removing HTML and other tags

"*RT* ~~**@BP_America**~~ *Did you know the first service stations opened around 1910 Self-service stations did not become the norm until 1970s http://t.co/4obuMaCS*"

After removing URLs

"*RT Did you know the first service stations opened around 1910 Self-service stations did not become the norm until 1970s* ~~**http://t.co/4obuMaCS**~~"

After removing stop words

"*RT Did you know* ~~**the**~~ *first service stations opened around 1910 Self-service stations* ~~**did not**~~ *become the norm until 1970s* "

After performing stemming

"*RT Did you know first service station*~~**s**~~ *opened around 1910 Self-service station*~~**s**~~ *become the norm until 1970*~~**s**~~ "

5.1.3 Description of Pre-Processing

This section presents the detailed description of pre-processing phase.

5.1.3.1 Bag-of-words representation

The indexing step gives us a *bag-of-words* (where the word order is not taken into account) representation of documents, which is also called *attribute-value representation* in machine learning. Tokens were characterised as attributes and weights corresponding to these tokens' importance as values. First stop words were removed, and then assign a weight to each word based on their occurrence in each document / tweet (in this example using *Term-frequency*

approach). The words having more weight are considered more important *features* compared with the words having less weight. The order of the input words does not matter and hence this representation is also called *bags-of-word representation*.

5.1.3.2 Indexing

Each document is typically characterized by a vector of n weighted index terms. A Vector Space [140] is the most widely used document representation technique. In this model each document is represented by a vector of words. A collection of documents (in this case collection of tweets) was represented by a word-by-document matrix 'a', where every element a_{ij} of a matrix represents the occurrence of the word i in the document j. DF i be the number of times word i occurs in the number of documents containing i, and TF (i, z) be the frequency of word i in each document z. As an example of Term Frequency of each word is shown in the given sample tweet:

"**RT** *(TF=1)* **Did** *(TF=1)* **know** *(TF=1)* **first** *(TF=1)* **service** *(TF=1)* **station** *(TF=2)* **opened** *(TF=1)* **around** *(TF=1)* **1910** *(TF=1)* **Self-service** *(TF=1)* **station** *(TF=2)* **become** *(TF=1)* **norm** *(TF=1)* **until** *(TF=1)* **1970** *(TF=1)*"

There are different heuristics to define the weight of word i in document z. In this work, the Term-Frequency Inverse Document Frequency (TF-IDF) technique has been used. The main idea of (TF-IDF) is if a word appears in the document in the high frequency (TF) and rarely appears in other documents, this word has a good ability to distinguish between categories. It employs the frequency of a word in a given tweet as well as in the collection of tweets for computing weight. The weight w of a word i in document z is computed as a combination of TF_{iz} and IDF_i, i.e.

$$w_{iz} = TF_{iz} \times IDF_{iz} \tag{5.1}$$

where IDF is calculated from DF as follows:

$$IDF_i = \log\left(\frac{DF_i}{N}\right) \tag{5.2}$$

Where N is total number of documents in the corpus. Intuitively, the IDF of a word is high if it occurs in one document and is low if it occurs in many documents. This scheme assigns weight to word i in document z that is:

- Highest when it occurs many times within a small number of documents
- Lower when it occurs few times in one document or many documents

- Lowest when it occurs in almost all documents

TF-IDF approach does not take the length of each document into account, which could be a potential problem in certain situations where documents have different length. This problem can be eliminated, by normalizing the weight *(Normalized TF- IDF Weighting)* [141, 142]. In this work normalized TF-IDF approach was used for assigning the weight to individual words in a tweet. Instead of using the term count the relative frequencies have been used to normalize by document size. There are other approaches as well for finding the features weight in a document, such as, Boolean weighting, word frequency weighting, logarithmic TF-IDF weighting, and entropy [137]; however, normalized TF-IDF has been used as it is the simplest. Furthermore, the TF-IDF approach has successfully been used in many text categorization applications; for example [143, 144] used this approach for building a text-categorization based movie recommender system, and reported good results. Further details of using text features for classification purposes can be found in [145, 146].

5.1.4 Features and Categorization Labels

This section presents the features used in these experiments. There are different reasons to choose these features such as: they are easy to extract, simple but salient and intuitive and any machine learning classifier can be trained over them. The list of features extracted from the applied datasets in this chapter are shown in Table 5.1. Firstly tweet content information is used to support user classification because it includes the users' lexical usage and the main interested topics to the users. In addition, profissional classes use more formal expressions than others. N-grams are collected from the demonstrated corpus which can be considered as a contiguous sequence of n terms extracted from a given sequence of each textual tweet. Secondly the TF-IDF of pre-processed tweet content is considered as one of the extracted feature; which is then followed by inclusion of automated sentiments of tweets or manual sentiment, number of followers, pre-processed user description and features concerning re-tweet, tweet length, user URLs, tweet URL and hashtags as shown in Table 5.1.

The value of in-depth features such as n-gram models, sociolinguistic feature (e.g., tweet sentiment), statistics about the user's immediate network (e.g., number of followers) and communication behavior (e.g., retweet frequency) can reflect a deeper understanding of the Twitter user stream and the user network structure. Furthermore,user profiles may significantly help in estimating the authority of a user on a topic for example, a user primarily interested in politics is more likely to be authoritative on political issues than a user who has an attention among other interests.

Table 5.1: Features details

No.	**Feature name**	**Type**	**Feature details**
1	Tweet content	N-grams	Sequence of the words in the tweet
2	Tweet content	String	Content of the tweet itself
3	Automated sentiment	Positive, Negative, Neutral	Sentiment of the tweet marked automatically
4	Manual sentiment	Positive, Negative, Neutral	Sentiment of the tweet marked manually
5	Number of followers	Any non-negative Integer value	Number of followers of the user who tweeted
6	User description	String	Description of the user who tweeted
7	Re-Tweet (RT)	Boolean (Yes, No)	If the tweet is original or has been re-tweeted
8	Tweet length	Discrete	Length of the tweet
9	User URLs	Boolean (Yes, No)	Does the user description have a URL?
10	Tweet URL	Boolean (Yes, No)	Does the tweet content have a URL?
11	Tweet hashtags	Boolean (Yes, No)	Does the tweet have a Hashtag?

These features are used for categorizing the tweets into eight categories as mentioned earlier. The classification model trained on the following categories:

- General public
- Media
- Environmentalists
- Politicians
- Business analysts
- Oil company employees
- Government organisations
- Non-Government organisations

5.1.5 Text Categorization Methods

The following algorithms were used extensively for recommender/expert system and in particular for solving text categorization problems [138] - [140], [147]. These classifiers have different characteristic which can classify the data in different way to get good prediction accuracy, their features were discussed previously in Chapter 2.

- K-Nearest Neighbours (KNN) classification based on similarity.
- Naïve Bayes (NB) classification based on probability.
- Support Vector Machines (SVM) classification based on statistics.
- Decision Tree (DT) classification based on graphs.

5.1.6 Partitioning the Data into Testing and Training Sets

The 5-fold cross validation scheme was used to partition the given data files into a testing and training sets. The average accuracy of results that obtained over the 5-folds was reported. The 5-fold cross validation approach to partition the dataset has been the preferred approach in the machine learning literature for reporting results. Many researchers have used it, for example [138, 147]. For the same experiments, the dataset was randomly divided into 80% training set and 20% test set across 5-folds. It has been used in [144].

5.1.7 Evaluation Metric

The accuracy metric for measuring the performance of the classification approaches has been used. Formally, it is defined as [149, 96] :

$$Accuracy = \frac{\text{Number of correctly classified tweets}}{\text{Total number of classified tweets}} \tag{5.3}$$

The objective is to increase the accuracy score [96], which corresponds to lowering the rate of classification error.

5.2 Experiment Finding and Discussion

This section presents the experiment results obtained from different classifiers.

5.2.1 Classifiers Results

The following tables and figures present the user groups predicted by four different machine learning classifiers, namely, SVM, KNN, NB and DT for both companies' datasets.

5.2.1.1 BP_America dataset

The detailed results of predicted user categories on BP_America dataset are shown in Table 5.2, 5.3, 5.4 and 5.5 whereas the in-depth classifiers' visualizations are shown in Figure 5.2, 5.3, 5.4 and 5.5. The leftmost column in each table shows the list of all row categories and the columns on the right show the predicted categories against actual one. It is obvious that the dominant category is general public because users sometimes use nicknames and provide ambiguous description in their profiles which is hard to find their original affiliation, then they have been classified as a general public. However, Twitter user accounts that belong to professional organization or affiliation provide correct and clear description. It can be clearly seen from the descriptive results shown in the following 4 tables including all features that the Naïve Bayes algorithm has produced most of correct predictions with general public and media categories; which has produced the minimum misclassification compared to other classifier results. Due to that, Naïve Bayes algorithm outperforming other algorithms.

Table 5.2: SVM results for BP_America dataset

Row categories	**Predicted Categories**								**Grand Total**
	Business analysts	**Enviromentlists**	**General public**	**Government organisation**	**Media (Journalists, Writers, etc)**	**Non-Government organisation**	**Oil Company employee**	**Politicians**	
Business analysts	8		10	6	16		1		41
Enviromentlists	1		18	1	4				24
General public			104	28	25		9		166
Government organisation	2	1	17	12	10	1	16	2	61
Media (Journalists, Writers, etc)	12		38	18	54	1	19	1	143
Non-Government organisation	7	1	31	45	40	4	25	1	154
Oil Company Employee			3	3			4	1	11
Politicians			1						1
Grand Total	**30**	**2**	**222**	**113**	**149**	**6**	**74**	**5**	**601**

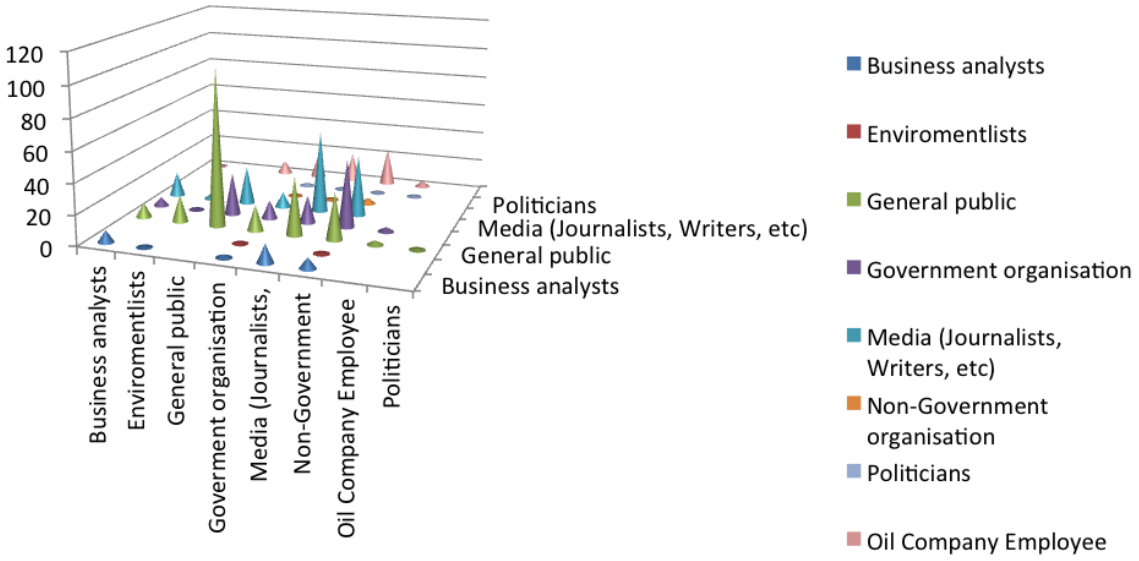


Figure 5.2: SVM results for BP_America dataset.

Table 5.3: KNN results for BP_America dataset

Row categories	**Predicted Categories**						**Grand Total**
	Business analysts	**Enviromentlists**	**General public**	**Government organisation**	**Media (Journalists, Writers, etc)**	**Non-Government organisation**	
Business analysts	7		9	1	20	4	41
Enviromentlists		1	17		5	1	24
General public		12	93	7	26	28	166
Government organisation	1	1	17	5	22	15	63
Media (Journalists, Writers, etc)	9		28	3	80	23	143
Non-Government organisation	3	1	26	8	59	57	154
Oil Company Employee			3		5	3	11
Politicians			1				1
Grand Total	**20**	**15**	**194**	**24**	**217**	**131**	**601**

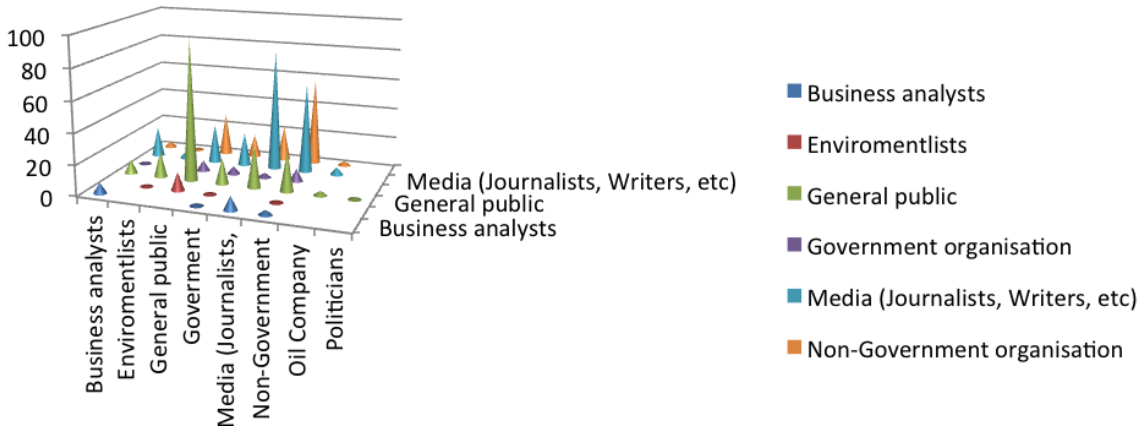


Figure 5.3: KNN results for BP_America dataset.

Table 5.4: NB results for BP_America dataset

Row categories	**Predicted Categories**					**Grand Total**
	Enviromentlists	**General public**	**Government organisation**	**Media (Journalists, Writers, etc)**	**Non-Government organisation**	
Business analysts		3		33	5	41
Enviromentlists		18		5	1	24
General public	2	121		38	5	166
Government organisation	1	19		26	15	61
Media (Journalists, Writers, etc)	1	24		100	18	143
Non-Government organisation	1	33	1	82	37	154
Oil Company Employee		1		6	4	11
Politicians		1				1
Grand Total	**5**	**220**	**1**	**290**	**85**	**601**

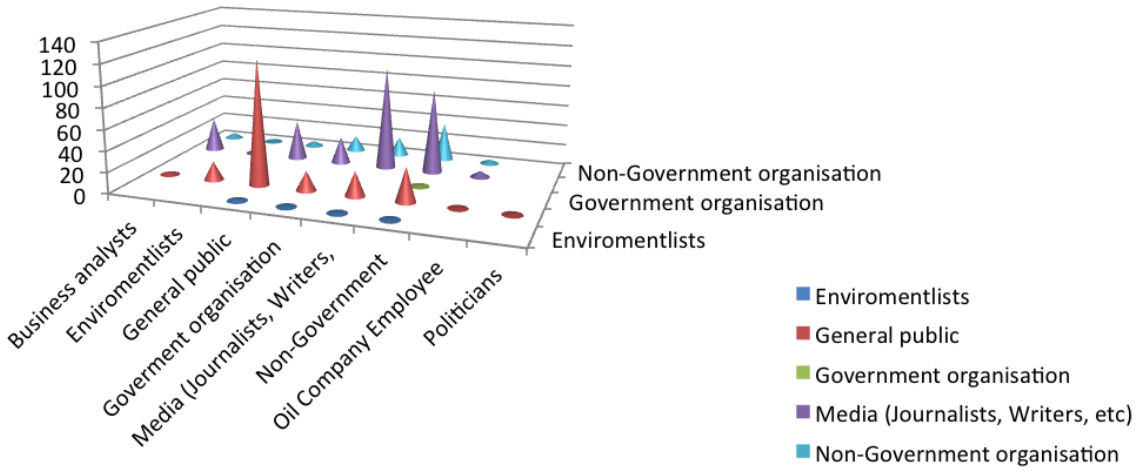


Figure 5.4: NB results for BP_America dataset.

Table 5.5: DT results for BP_America dataset

Row categories	Predicted Categories						Grand Total
	Business analysts	Enviromentlists	General public	Government organisation	Media (Journalists, Writers, etc)	Non-Government organisation	
Business analysts	7		9	1	20	4	41
Enviromentlists	1	1	17		4	1	24
General public	1	10	94	8	24	29	166
Government organisation	2	2	13	9	17	18	61
Media (Journalists, Writers, etc)	9	2	20	10	68	34	143
Non-Government organisation	3	5	25	16	51	54	154
Oil Company Employee			2		5	4	11
Politicians			1				1
Grand Total	**20**	**15**	**194**	**24**	**217**	**131**	**601**

5.2.1.2 Saudi_Aramco Dataset

Predicted user groups of Saudi_Aramco dataset are shown in Table 5.6, 5.7, 5.8 and 5.9 respectively. The leftmost column in each table shows the row categories and the right columns show the predicted categories against actual one. Similar to BP_America user categories the general public category is dominant in Saudi_Aramco for the same reasons explained earlier. From the descriptive results shown in the following tables that general public is the most frequent category of all tweets in Saudi_Aramco dataset. The minimal misclassification was produced by SVM algorithm. The best predictive category of SVM is general public category which is one of

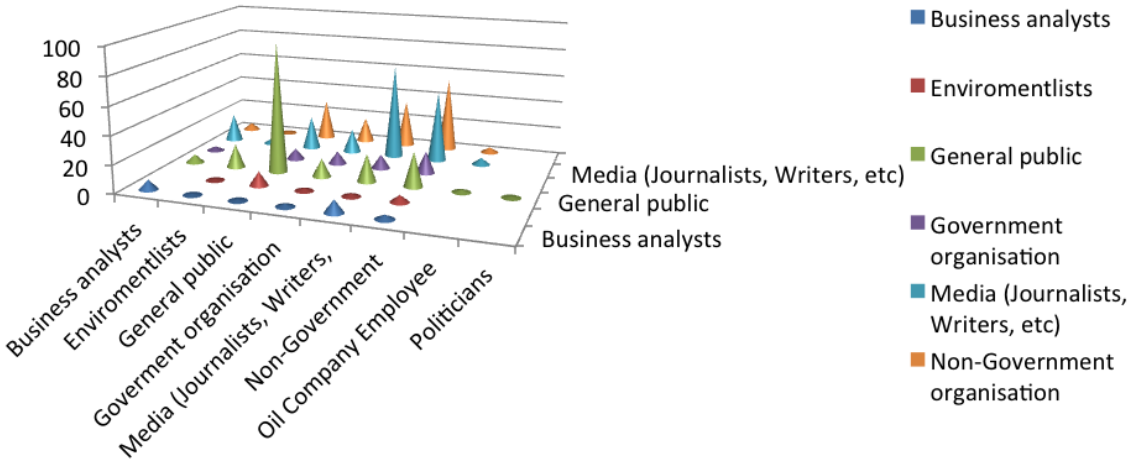


Figure 5.5: DT results for BP_America dataset.

the main reason of producing maximum accuracy compared to other classifiers. Visualizations of the four classifiers are shown in Figure 5.6, 5.7, 5.8 and 5.9.

Table 5.6: SVM results for Saudi_Aramco dataset

Row categories	**Predicted Categories**				**Grand Total**
	Business analysts	**General public**	**Non-Government organisation**	**Oil company employee**	
Business analysts		10	1	1	12
Enviromentlists		10			10
General public	7	417	33	17	474
Government organisation		4		1	5
Media (Journalists, Writers, etc)	1	22	3	3	29
Non-Government organisation	1	18	14	21	54
Oil Company Employee		8			8
Politicians	1	1	7		9
Grand Total	**10**	**490**	**58**	**290**	**601**

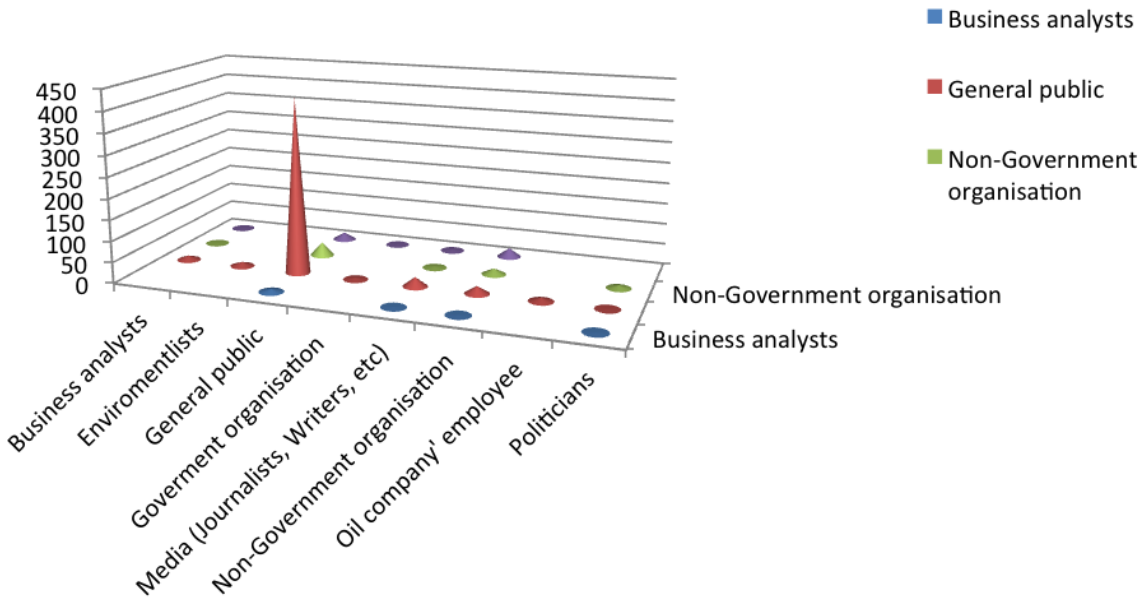


Figure 5.6: SVM results for Saudi_Aramco dataset.

Table 5.7: KNN results for Saudi_Aramco dataset

Row categories	**Predicted Categories**		**Grand Total**
	General public	**Non-Government organisation**	
Business analysts	12		12
Enviromentlists	10		10
General public	474		474
Government organisation	5		5
Media (Journalists, Writers, etc)	26	3	29
Non-Government organisation	54		54
Oil Company Employee	8		8
Politicians	9		9
Grand Total	**490**	**58**	**601**

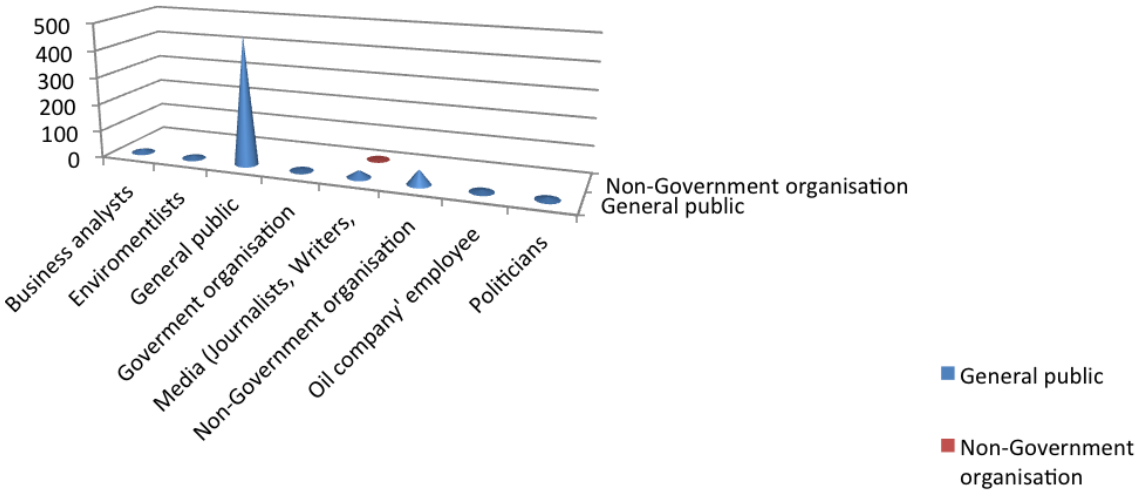


Figure 5.7: KNN results for Saudi_Aramco dataset.

Table 5.8: NB results for Saudi_Aramco dataset

Row categories	**Predicted Categories**			**Grand Total**
	General public	**Non-Government organisation**	**Oil Company employee**	
Business analysts	10	1	1	12
Enviromentlists	10			10
General public	379	67	28	474
Government organisation	5			5
Media (Journalists, Writers, etc)	22	6	1	29
Non-Government organisation	50	3	1	54
Oil Company Employee	7		1	8
Politicians	8		1	9
Grand Total	**491**	**77**	**33**	**601**

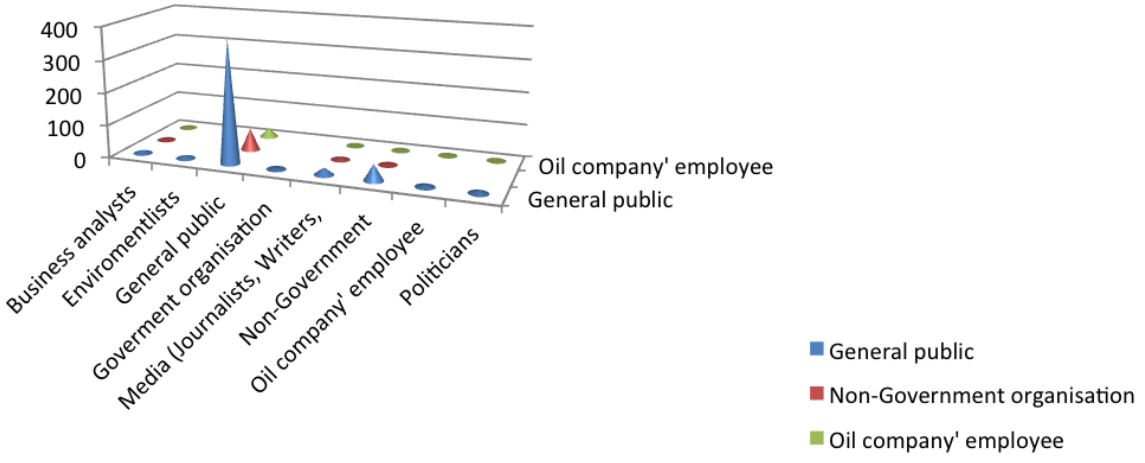


Figure 5.8: NB results for Saudi_Aramco dataset.

Table 5.9: DT results for Saudi_Aramco dataset

Row categories	Predicted Categories					Grand Total
	Business analysts	General public	Government organisation)	Media (Journalists, Writers, etc)	Non-Government organisation	
Business analysts		10	1	1		12
Enviromentlists		10				10
General public	7	417	33	17		474
Government organisation		4		1		5
Media (Journalists, Writers, etc)	1	22	3	3		29
Non-Government organisation	1	18	6	21	8	54
Oil Company Employee		8				8
Politicians	1	1	7			9
Grand Total	**10**	**490**	**50**	**43**	**8**	**601**

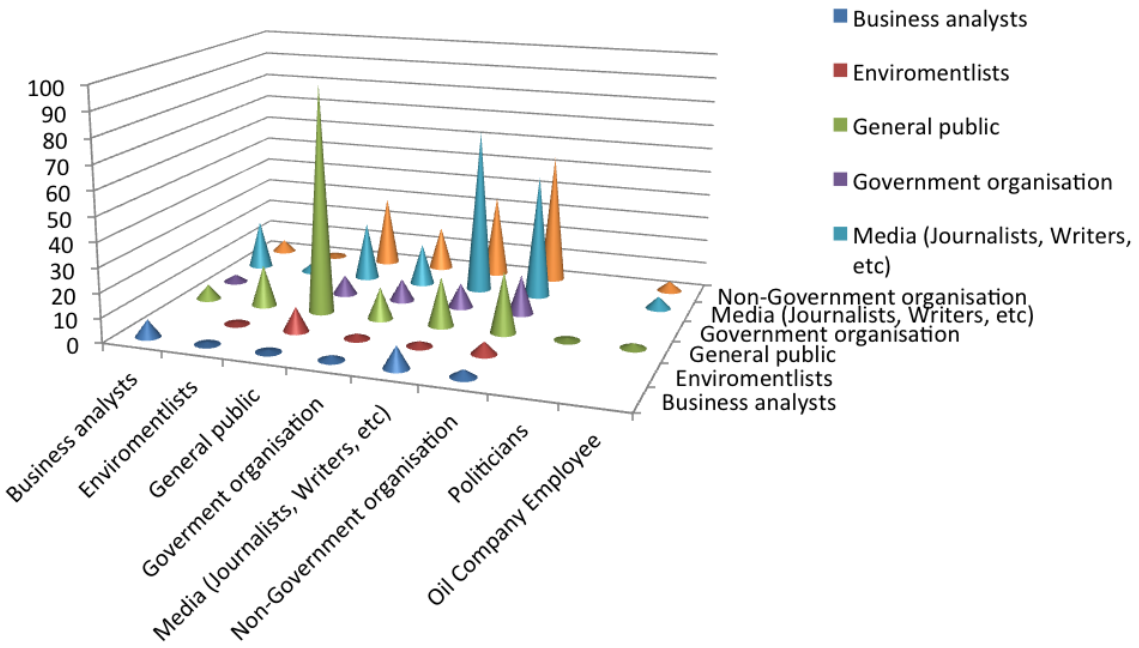


Figure 5.9: DT results for Saudi_Aramco dataset.

Table 5.10: Prediction accuracy on BP_America datasets

No. of Features	Features	SVM	KNN	NB	DT
1	**All features**	38.37	35.81	42.56	39.94
2	**Tweet content N-grams**	34.65	36.98	42.09	37.67
3	**Tweet content string**	36.98	35.17	42.79	38.84
4	**Automated sentiment**	38.84	35.81	42.56	38.14
5	**Manual sentiment**	36.74	34.19	42.56	38.37
6	**Number of followers**	40.23	37.44	42.09	39.07
7	**User description string**	39.07	36.28	42.33	39.30
8	**RT**	38.84	35.81	42.56	39.77
9	**Tweet length**	38.60	39.07	42.56	37.67
10	**URL in user description**	38.84	36.51	42.56	38.14
11	**URL in Tweet content**	37.44	34.42	40.47	37.44
12	**Hashtag in Tweet content**	39.07	37.21	42.33	38.84

Table 5.11: Prediction accuracy on Saudi_Aramco datasets

No. of Features	Features	SVM	KNN	NB	DT
1	**All features**	79.56	78.22	57.56	68.89
2	**Tweet content N-grams**	79.56	79.33	78.89	69.11
3	**Tweet content string**	79.56	78.89	57.56	67.78
4	**Automated sentiment**	79.56	78.89	57.56	70.22
5	**Manual sentiment**	79.56	79.11	57.56	68.67
6	**Number of followers**	79.56	79.11	57.56	68.22
7	**User description string**	78.44	78.89	57.56	78.22
8	**RT**	79.56	78.89	57.56	70.44
9	**Tweet length**	80.0	79.11	57.56	71.78
10	**URL in user description**	79.56	79.11	57.56	68.44
11	**URL in Tweet content**	79.56	78.67	57.56	69.56
12	**Hashtag in Tweet content**	79.56	78.89	57.56	76.89

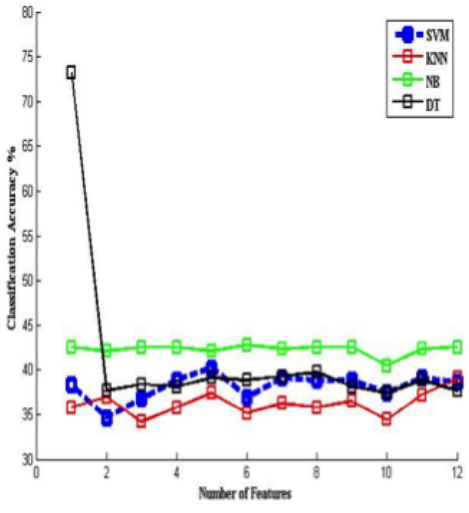
5.2.2 Prediction Accuracy

Tables 5.10 and 5.11 present the predictive accuracy percentage of 4 different machine learning classifiers based on 11 extracted features for both companies. For both of the datasets, the N-gram is the most important feature, as can be seen in Tables 5.10 and 5.11. The manual sentiment feature can increase the accuracy slightly but the general accuracy was quite low because user group prediction is a hard task. The accuracy was low for the BP_America dataset, because the tweets in this dataset were very noisy. The best overall predictive accuracy including all features on BP_America dataset was recorded with Naïve Bayes algorithm which is (42.56%). It can also be noted that inclusion of selective features with Naïve Bayes produced better results compared to including all features which is shown clearly in Table 5.10. The second best classifier recorded in BP_America dataset is DT with accuracy percentage of (39.94%) followed by SVM (38.37%) and KNN (35.81%) algorithm. However, each feature performed differently within classifiers. For example, number of followers feature performed well with NB (42.09%), SVM (40.23) and DT (39.07) while tweet length produced higher accuracy with KNN (39.07%) and NB (42.56%). Table 5.10 also shows that user description was very because of its high performance within all classifiers. On the other hand, URL in tweet content did not make big difference in the prediction accuracy with the four classifiers. On the Saudi_Aramco dataset, the accuracy obtained was satisfactory. SVM with percentage of accuracy (79.56%) followed by KNN classifier (78.22%), performed better than other algorithms on Saudi_Aramco dataset. Table 5.11 revealed that the features affected the prediction accuracy in different way within the classifiers. For instance, tweet length feature gave very high accuracy with SVM (80%), (79.11%) with KNN and (71.78%) with DT. All other features performed similarly with SVM and NB classifiers. By contrast, with DT classifier features such as user description, RT and automated sentiment produced higher accuracy while URL in user description and number of follower produced lower accuracy.

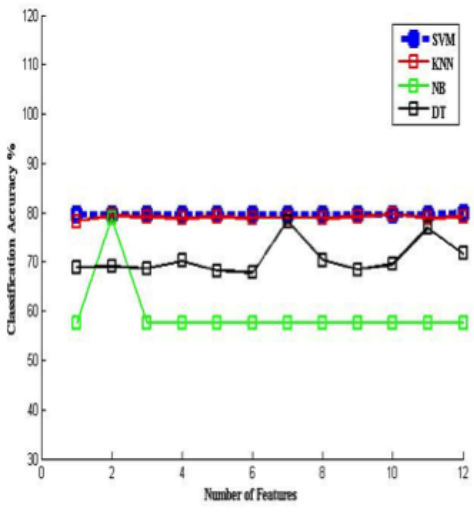
Figures 5.10a and 5.10b show the accuracy of the classifiers of both datasets. Table 5.11 revealed that the features affected the prediction accuracy in different way within the classifiers. For instance, tweet length feature gave very high accuracy with SVM (80%), (79.11%) with KNN and (71.78%) with DT. All other features performed similarly with SVM and NB classifiers. By contrast, with DT classifier features such as user description, RT and automated sentiment produced higher accuracy while URL in user description and number of follower produced lower accuracy. It can be seen that Naïve Bayes proved a better classifier than SVM on BP_America dataset, while SVM classifier produced more accurate results for Saudi_Aramco dataset. For other oil company datasets, or any controversial ones, Naïve Bayes or SVM classifiers are suggested to use, owing to their high rate of accuracy in classification.

5.2.3 Discussion

The results suggest that the Twitter data collected in this experiment are actually non-linear in nature, i.e. it is very hard to classify the data with a linear classifier. However, SVM in particular performed best, because of its non-linear poly kernel function. Other classifiers suffered in their lack of a non-linear function as an error function. It is true that multilayer perceptron (MLP) has a feature for classifying non-separable data linearly but the sigmoid function used in MLP cannot classify the data well. On the other hand, the nature of Twitter data is usual noisy as a result of including various labels, abbreviations and irregular form which can affect the accurecy of prediction. In the pre-processing data phase some of noisy data was removed to reduce the noise. All features were used to evaluate the importance of each feature. Tables 5.10 and 5.11 reported when the best performance was obtained. The classifiers showed poor results with the BP_America dataset but performed very well with the Saudi_Aramco dataset. After observing both datasets, the BP_America dataset was very noisy and its tweets contain mainly URL. To enhance the accuracy of classification much of the noise will be removed in the future work such as dataset sparsity which is an important factor that affects the overall performance of a typical machine learning classifiers. As showed in [69] Twitter data are sparser than other types of data due to the large number of infrequent words present within tweets.



(a) Classifiers accuracy on BP_ America dataset.



(b) Classifiers accuracy on Saudi_Aramco dataset.

Figure 5.10: Accuracy of the classifiers.

5.3 Conclusion

This chapter presents the users categorization framework based on 6000 Twitter post mentioning oil companies. The proposed approach is similar to predicting the authorship of textual data. The results clearly show SVM outperformed other classifiers while classifying Saudi_Aramco datasets, whereas Naïve Bayesian Classifier gives the highest classification accuracy for BP_America dataset. This concludes that the performance of classifiers for textual classification is fully dependent on the characteristic of data. Thus, understanding of group structure in both companies' network is required to cluster them to find the similarities and relationship between them, and this is addressed further as part of the clustering analyses in the following chapter.

Chapter 6

An Approach to Tweets Clustering

Introduction

Clustering belongs to unsupervised machine learning algorithms, which are used for identifying natural groupings or clusters within multidimensional data based on some similarity measure (e.g. Euclidean distance) [95]. Text mining is one of the common clustering applications, the analysis of large numbers of texts to find similarities between documents and extract underlying patterns in the data [150]. Since it is a quite hard to predict user groups as was mentioned in the previous chapter, this chapter aims to demonstrate the clustering of tweets derived from two oil companies, BP_America and Saudi_Aramco, as a different approach based on automated sentiment feature and other features. The number of tweets are predicted rather than an individual tweet and then investigate the accuracy of the sentiment and find the relationships between the user categories in each cluster. In this chapter three different experiments were performed using k-means, the most widely used clustering algorithm in Twitter data mining [150]. The first one considers the whole dataset as a training set to find the similarity in tweets based on user categories. In the second, predictive modelling is proposed based on automated and manual sentiment was proposed. A hybrid sentiment analysis technique is applied in the final experiment which is a better way of assessing sentiment by looking at users' behaviour in clusters.

In Section 6.1 a primary analysis including overview of pre-processing steps and feature labels is presented. The implicit unsupervised clustering of tweets by applying a hard clustering algorithm is discussed in Section 6.2. The demonstration of predictive modelling based in the automated and manual sentiment is described in Section 6.3. In Section 6.4 a novel method of hybrid sentiment analysis within predicted clusters is presented. The chapter concludes with Section 6.5.

6.1 Primary Analysis

In the clustering experiments, the same datasets were used as in previous experiments mentioned earlier in this research. Automated and manual sentiment analysis techniques are discussed in Chapter 4. Methods of categorizing Tweet users are outlined in Chapter 5.

6.1.1 Pre-Processing

To allow clustering, the pre-processing steps were applied to the tweet datasets. Pre-processing was needed to remove noise from the tweets, such as stop words, URLs, and HTML tags followed by tokenization and stemming. In the same way, this technique was used in text categorization described in detail in Chapter 5.

6.1.2 Feature Extraction Labels

A total of 11 features were used, namely, tweet content (N-grams), tweet content (string), automated sentiment, manual sentiment, number of followers, user description, RT, tweet length, user URLs, tweet URLs and tweet hashtags. In Chapter 5 more details of these features were described.

6.2 Unsupervised Learning by K-Means

In this section the demonstration of k-means by considering all the data as a training dataset for both Saudi_Aramco and BP_America was described. The rationale for doing this was to find the similarities in the tweets that belonged to different categories.

6.2.1 Experiments Methods/Algorithm

In order to correctly analyse the relationship between tweets belonging to different categories, the value of the k centres in k-means is considered to be far higher than the total number of categories. This helps to find the dissimilarities between tweets in each category that might result in being spread across multiple clusters. Similarly, the frequency of the same kind of tweet in each category have been seen. Additionally the relationship of different users belonging to both the same and different categories can also be highlighted. In fact that there is no

way to automatically set the right "k" or know how many clusters are needed before performing the algorithm. K-means is highly dependent on the initializations of the centroids. Since these initializations are random, multiple runs of k-means produce different results. Therefore, empirically setting "k" with different values $k = 10$ through 50. Then k-means was run a final time with the optimal value which suited the rationale that came out as $k = 20$. The Euclidean Distance approach was used in k-means as a distance measure in this experiment. Euclidean distance is the magnitude of the distance between the data points. For example, if two data points represented two sentences both containing three words in common, Euclidean distance takes into account the magnitude of similarity between the two sentences. Thus, a sentence containing the words "oil" and "media" 3 times and another containing those words 300 times are considered more dissimilar by Euclidean distance.

6.2.2 Results of Saudi_Aramco Dataset

Table 6.1 presents the row results that have been found which cluster the similar users together. General public category appeared in each cluster and came with media and non-government organisation in different clusters. On the other hand, environmentalist, government organisation, oil company employee and politicians clustered together in different clusters. That indicates these groups have similar behaviour. For example, general public, media and non-government organisation can focus more on the latest news about oil companies and crisis that faced them. In contrast, tweets that issued by other categories would be included more direct and logical criticism in both positive and negative sides. After presenting the main clusters, the variation of automated sentiment in these groups was investigated. Figure 6.1 shows the percentage of tweets (positive, negative and neutral automated sentiment) in each cluster. Positive automated sentiments are mostly clustered in clusters 0, 4, 9, 14 and 19, whereas neutral sentiments are generally clustered in clusters 2, 9, 16 and 18. Finally, negative sentiments are grouped most often in clusters 1 and 9. This shows the dissimilarities in the emotions of tweets which are predictable since positive tweets should obviously be different from negative tweets and both should be dissimilar from neutral tweets.

Similarly, Figure 6.2 shows the percentage of tweets (positive, negative and neutral manual sentiment) in each cluster. Positive manual sentiments are mostly clustered in clusters 0, 1 and 9, whereas neutral sentiments are most of the time clustered in clusters 0, 1, 9, 12, 14, 16, 18 and 19. Finally negative sentiments are grouped mostly in clusters 0 and 9. The clustering of manual sentiments shows somewhat abnormal behaviour but was expected due to the fact that they were labelled manually, which shows more closeness to each other than is shown in automated sentiments which are far different from each other in terms of their sentiment type.

Table 6.1: The row clusters in Saudi_Aramco dataset

Cluster	User Categories								Grand Total
	General public	Business analysts	Enviromentlists	Media (Journalists, Writers, etc)	Government organisation)	Non-Government organisation	Oil Company employee	Politicians	
0	255	2	4	8	1	8	24	3	305
1	520	8	2	11	5	30	28	6	610
2	198	7		9	1	7	18	3	244
3	1					1			2
4	24		1	1	1	5			33
5	14								14
6	4								4
7	24	2					2	1	29
8	1								1
9	1060	33	17	78	8	122	94	23	1435
10	17	1				4	2		24
11	12				1				13
12	17			2		3	4	2	28
13	3					1			4
14	96	3		2	1	3	5		110
15	1	1							2
16	53	2		3	1	5	4		68
17	4			1					5
18	48	2	1		1	1	4		57
19	33	1		1			4		39

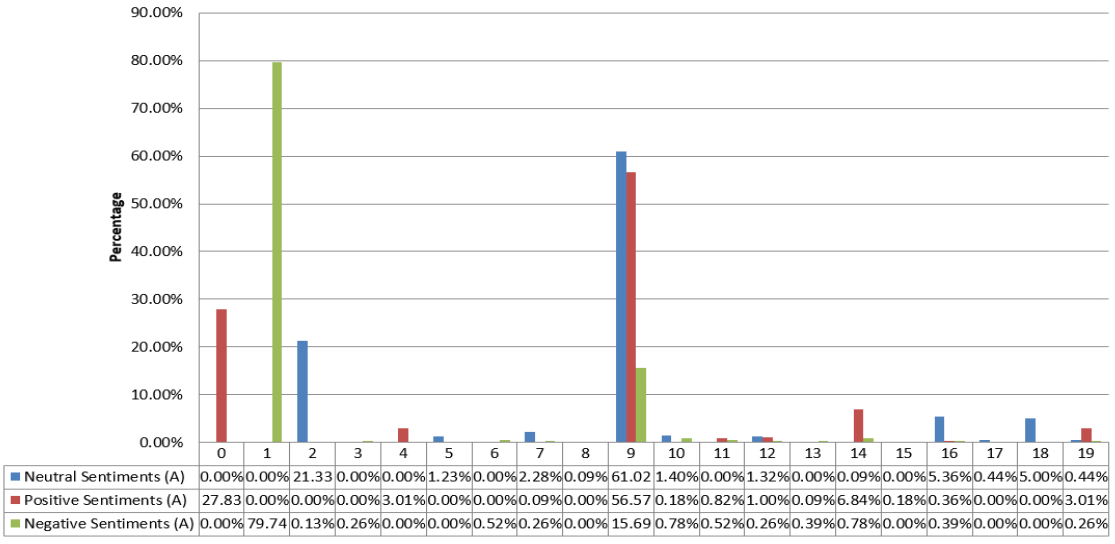
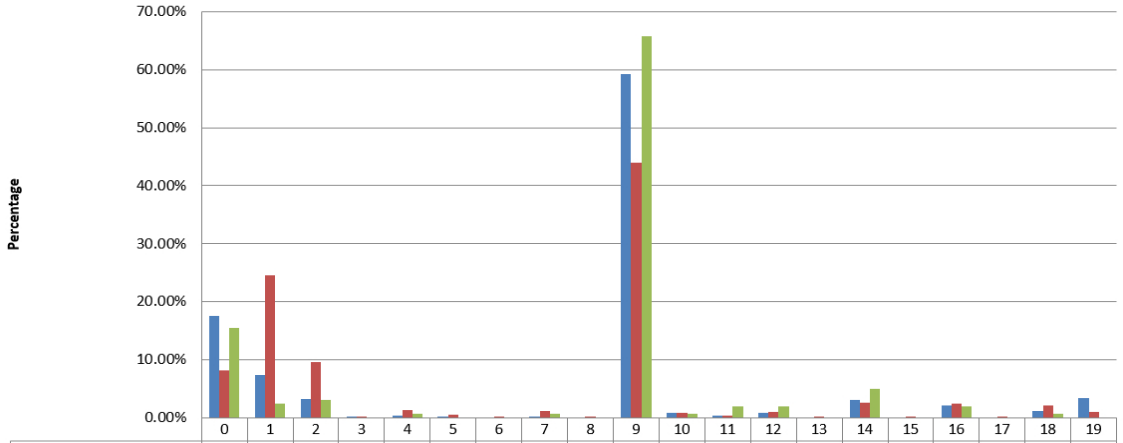


Figure 6.1: Percentage of automated sentiment in Saudi_Aramco clusters

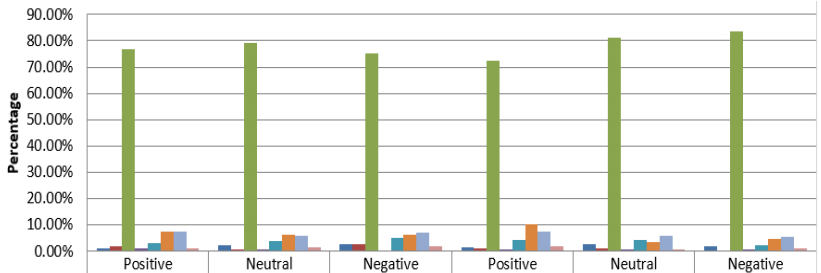
Figure 6.3 shows the combined chart counting both manual and automated sentiments category-wise. It can be clearly seen from this figure that most of the manual and automated sentiments belong to the general public category. Other categories have a lower count of tweets than this one has. That can show the clustering results of the counts of tweets belonging to different categories. Figure 6.4 illustrates the count of tweets belonging to different categories as shown in 20 clusters after k-means. It can be seen clearly from Figure 6.4 that business analysts',



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Positive Sentiments (M)	17.49%	7.41%	3.23%	0.19%	0.38%	0.19%	0.00%	0.19%	0.00%	59.32%	0.76%	0.38%	0.76%	0.00%	3.04%	0.00%	2.09%	0.00%	1.14%	3.42%
Neutral Sentiments (M)	8.13%	24.51%	9.60%	0.04%	1.30%	0.56%	0.17%	1.17%	0.04%	43.97%	0.82%	0.35%	0.91%	0.17%	2.51%	0.09%	2.33%	0.22%	2.16%	0.95%
Negative Sentiments (M)	15.53%	2.48%	3.11%	0.00%	0.62%	0.00%	0.00%	0.62%	0.00%	65.84%	0.62%	1.86%	1.86%	0.00%	4.97%	0.00%	1.86%	0.00%	0.62%	0.00%

Figure 6.2: Percentage of manual sentiment in Saudi_Aramco clusters

non-government organizations' and government organizations' tweets came out very close together, demonstrating a close relationship between the communications of those belonging to the three categories above. Similarly, the tweets of politicians and the general public emerged as obviously very close to one another, due to the fact that politicians are meant to interact with the general public on Web 2.0. The sentiments of environmentalists seemed self-contained. The tabular view of the detailed count of each tweet category-wise in each cluster with both automated and manual sentiment given in Appendix A.



	Positive Sentiments (M)	Neutral Sentiments (M)	Negative Sentiments (M)	Positive Sentiments (A)	Neutral Sentiments (A)	Negative Sentiments (A)
Business Analyst	1.14%	2.25%	2.48%	1.55%	2.81%	1.70%
Environmentalist	1.71%	0.52%	2.48%	1.09%	0.88%	0.39%
General Public	76.62%	79.33%	75.16%	72.45%	81.30%	83.53%
Governement Organisation	1.14%	0.61%	0.00%	0.64%	0.61%	0.78%
Media (Journalists, Writers, etc)	3.23%	3.93%	4.97%	4.38%	4.39%	2.35%
Non-Government Organisation	7.60%	6.10%	6.21%	10.31%	3.60%	4.84%
Oil company employee	7.41%	6.01%	6.83%	7.57%	5.71%	5.36%
Politicians	1.14%	1.25%	1.86%	2.01%	0.70%	1.05%

Figure 6.3: Percentage of manual and automated sentiment in Saudi_Aramco user categories

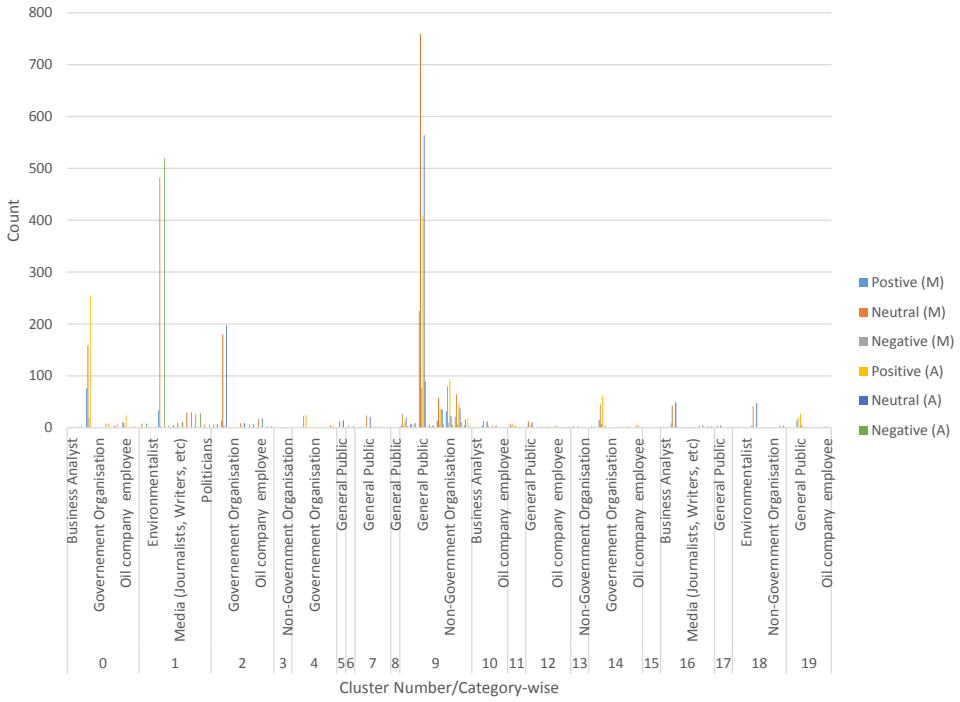


Figure 6.4: Count of tweets in Saudi_Aramco clusters

6.2.3 Results of BP_America Dataset

The row results of BP_America clusters reveals that some user groups usually cluster with each other such as (media and environmentalist) and (general public and non-government organisation). Table 6.1 presents more details about the row clusters results. Figure 6.5 shows the percentage of tweets (positive, negative and neutral automated sentiment) in each cluster. Positive automated sentiments are mostly clustered in cluster 0, 13, 14, 15 and 16, whereas neutral sentiments are most of the time clustered in 0, 8, 9, 13, 14, 15 and 16. Finally negative sentiments are grouped mostly in clusters 0, 13, 15, 16 and 19. This time there were many commonalities found in the attributes of emotions belonging to different categories. This clearly shows the closeness in tweets belonging to different emotions.

Similarly, Figure 6.6 shows the count of tweets (positive, negative and neutral manual sentiment) in each cluster. Positive manual sentiments are mostly clustered in clusters 0, 13, 14, 15 and 16, whereas neutral sentiments are most of the time clustered in clusters 0, 9, 13, 14, 15 and 16. Finally negative sentiments are grouped mostly in clusters 0, 1, 8, 9, 10, 15 and 17. The clustering of manual sentiments shows neutral and positive sentiments as very close to each other and most of the time clustered together, unlike negative sentiments for which the results were quite dissimilar.

Figure 6.7 shows the combined chart of the percentage of both manual and automated sentiments category-wise. A symmetrical distribution of tweets among all the categories can be seen

Table 6.2: The row clusters in BP_America dataset

Cluster	User Categories								Grand Total
	General public	Business analysts	Enviromentlists	Media (Journalists, Writers, etc)	Government organisation)	Non-Government organisation	Oil Company employee	Politicians	
0	111	39	45	223	88	169	12	11	698
1	95	2	19	11	9	26		1	163
2	2		6		1	1			10
3	5		20	9	8	6			48
4	1								1
5						1			1
6	4	1	26	7	2	4			44
7	1		1	1		1			4
8	5	1	12	1	1	5			25
9	82	8	16	37	13	40			196
10	10		56	17	12	9		2	108
11	4		50	14	12	10			90
12			4	1		2			7
13	1	16	11	22	8	12		1	71
14	14	21	8	85	25	56	5	3	341
15	299	34	61	222	86	231	11	10	954
16	15	12	6	64	20	34	5	1	157
17	10		38	16	15	6			85
18	1		1	1	1				4
19	15			1	2				18

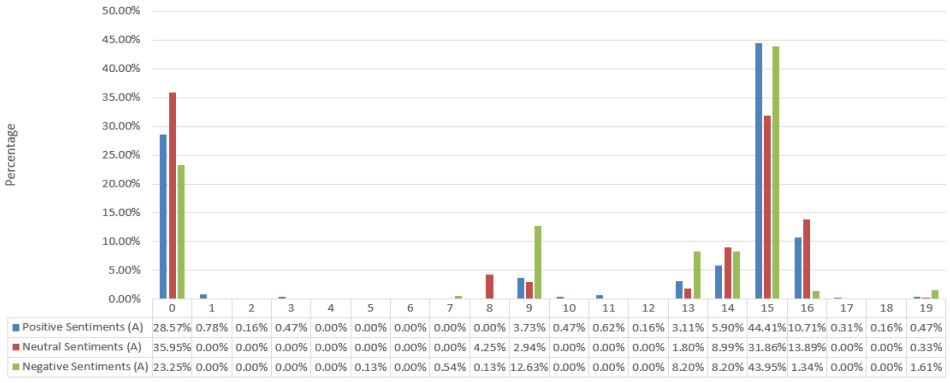
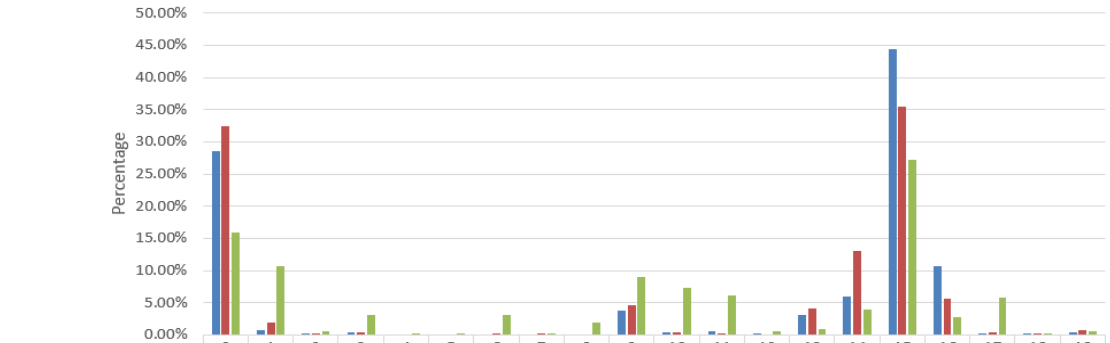


Figure 6.5: Percentage of automated sentiment in BP_America clusters

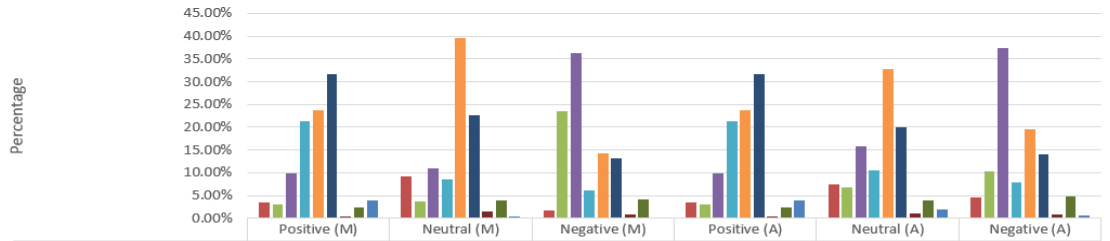
in Figure 6.7. That can show the clustering results of the count of tweets belonging to different categories. Figure 6.8 shows the count of tweets belonging to different categories shown in 20 clusters after k-means. First, a very close relationship between media workers (journalists, writers, etc.) and environmentalists resulted after clustering. This reveals that environmentalists have made some very new and interesting findings and communicated them to the media workers. Some very interesting discussions may have ensued between these two, for example their tweets about oil companies usually focus on the negative news, crisis and issues related to the impact of oil companies to damage the environment. Second, government organizations came out very self-contained. Third, a very close connection was discovered between the gen-



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Positive Sentiments (M)	28.57	0.78%	0.16%	0.47%	0.00%	0.00%	0.00%	0.00%	0.00%	3.73%	0.47%	0.62%	0.16%	3.11%	5.90%	44.41	10.71	0.31%	0.16%	0.47%
Neutral Sentiments (M)	32.47	1.86%	0.10%	0.41%	0.00%	0.00%	0.21%	0.10%	0.00%	4.55%	0.41%	0.31%	0.00%	4.14%	13.03	35.47	5.69%	0.41%	0.10%	0.72%
Negative Sentiments (M)	15.91	10.73	0.58%	3.02%	0.07%	0.07%	3.02%	0.29%	1.94%	8.93%	7.27%	6.19%	0.58%	0.94%	3.96%	27.21	2.74%	5.76%	0.14%	0.65%

Figure 6.6: Percentage of manual sentiment in BP_America clusters

eral public and non-government organizations. Last, politicians were mostly seen linked with business analysts and government organizations as they often tweet directly and logically about any issue. The tabular view of the detailed count of each tweet category-wise in each cluster is given in Appendix B.



	Positive (M)	Neutral (M)	Negative (M)	Positive (A)	Neutral (A)	Negative (A)
Business analysts	3.57%	9.10%	1.66%	3.57%	7.35%	4.57%
Enviromentlists	2.95%	3.62%	23.47%	2.95%	6.70%	10.35%
General public	9.94%	10.96%	36.21%	9.94%	15.85%	37.37%
Goverment organisation	21.27%	8.48%	6.05%	21.27%	10.46%	7.93%
Media (Journalists, Writers, etc)	23.76%	39.50%	14.18%	23.76%	32.68%	19.62%
Non-Government organisation	31.68%	22.54%	13.10%	31.68%	19.93%	14.11%
Politicians	0.47%	1.55%	0.79%	0.47%	1.14%	0.81%
FALSE	2.33%	3.83%	4.25%	2.33%	3.92%	4.70%
Oil Company Employee	4.04%	0.41%	0.29%	4.04%	1.96%	0.54%

Figure 6.7: Percentage of manual and automated sentiment in BP_America user categories

Finally, after running the experiment of both companies datasets 5 times randomly with different k values (10, 20, 30, 40 and 50), $k = 20$ produced better results. The main idea of considering the whole dataset as a training set to find the relationship between tweet authors based in the tweet similarity. This experiment results revealed which Twitter user categories come often together in one cluster. That can help to find out what the dominant sentiment of these groups in different clusters.

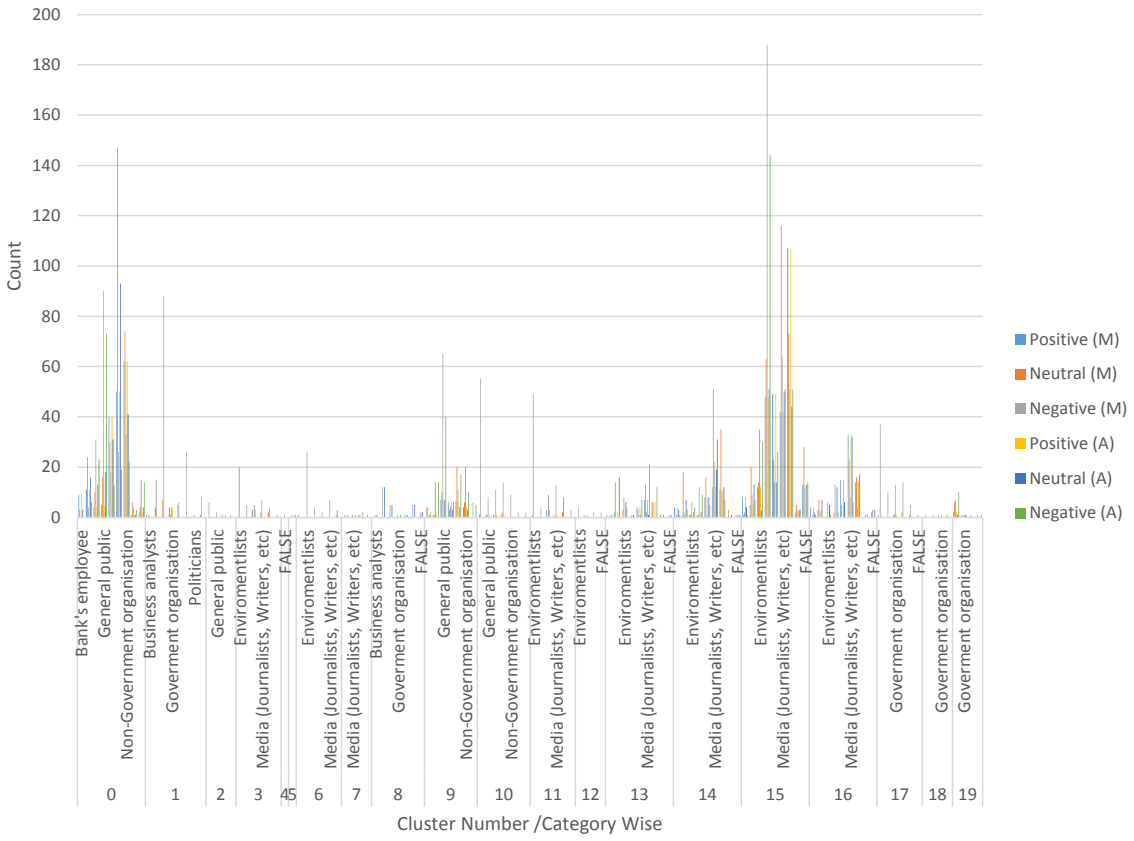


Figure 6.8: Count of tweets in BP_America clusters

6.3 Predictive Modelling

This section describes the application of predictive modelling for automated and manual sentiments in both the Saudi_Aramco and BP_America datasets. The aim of this model is to predict the sentiment of the testing tweets which could be positive, negative and neutral according to the given set of tweets as training data.

6.3.1 Experiment Method/Algorithm

The pre-processed features of Saudi_Aramco and BP_America datasets were split into training and testing, using the k-means hard clustering algorithm which labelled the training data of the tweets as positive, negative and neutral, using $k = 3$ (the number of classes). The labels of the automated sentiments was first tested followed by training and testing the labels of the manual sentiments. There were in total 3000 tweets extracted from each dataset, as described in Chapter 3. First the data was divided into training and testing subsets. K-fold cross validation was used by initializing $k = 10$ for the training and testing of datasets, which proved helpful in further maximizing the predictive accuracy. The distance measure used in k-means, as noted above,

is the Euclidean Distance approach. The False Positives (FP), True Positives (TP), True Negatives (TN), predictive accuracy, sensitivity and specificity of both the manual and automated sentiments of both datasets are shown in Tables 6.3 and 6.4.

6.3.2 Results of Saudi_Aramco Dataset

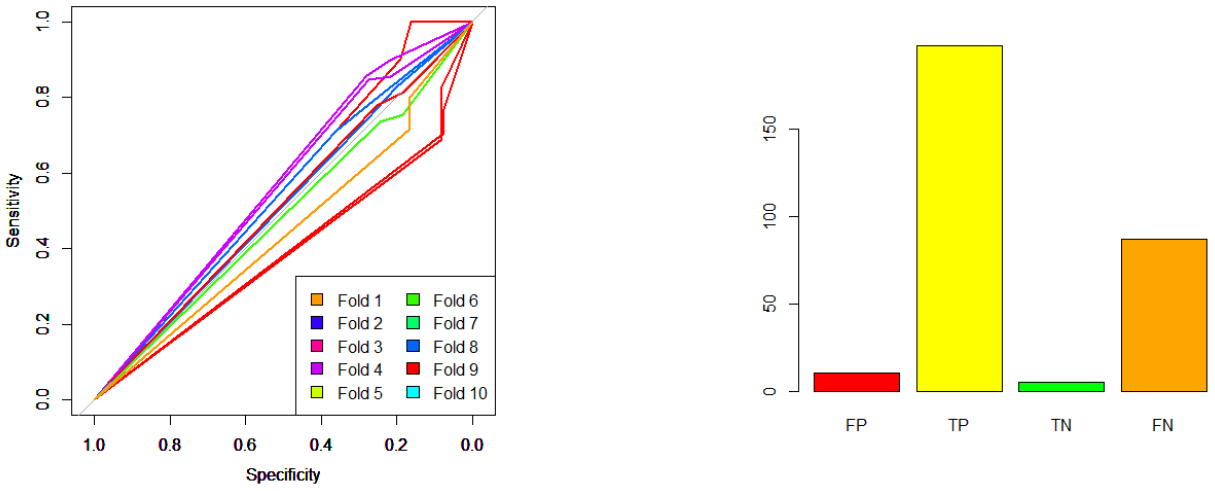
It can be seen from Table 6.3 which summarises the results of the Saudi_Aramco dataset that the predictive accuracy of manual sentiment came out a little higher than that of the automated sentiment. Table 6.3 shows the misclassification of 32.36 % of manual sentiment, which after training were wrongly predicted. Similarly the second row of Table 6.3 shows the misclassification accuracy of 33.83 % of automated sentiment wrongly predicted, compared with their original labels. To further demonstrate the ratio between sensitivity and specificity, each fold in terms of manual sentiment is tested by *receiver operating characteristic* (ROC) curve. It is a graphical plot that illustrates the performance of a binary classifier as its discrimination threshold is varied. The curve is created in this experiment by plotting (TP) against (FP) at 10 threshold as shown in Figure 6.9a; and Figure 6.9b, the barplot which shows a comparison of TP, TN, FP and FN is drawn. Similarly, to demonstrate the ratio between sensitivity and specificity, each fold in terms of automated sentiment is shown in Figure 6.10a; and Figure 6.10b, shows the barplot which shows a comparison of TP, TN, FP and FN is drawn.

Table 6.3: Predictive accuracy of manual and automated sentiment of Saudi_Aramco dataset

Sentiments	**FP**	**TP**	**TN**	**FN**	**Accuracy**	**Senstivity**	**Specificity**
Manual	10.4	197.2	5.7	86.7	0.676333	0.694610	0.35403727
Automated	49.2	171.2	27.3	52.3	0.661666	0.765995	0.356862

6.3.3 Results of BP_America Dataset

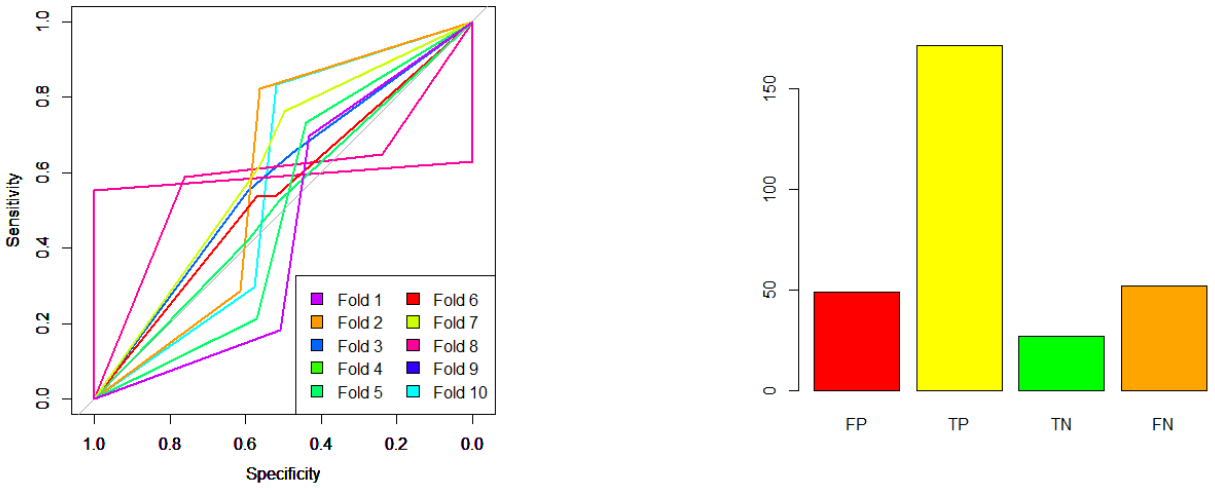
Likewise, Table 6.4 shows the results of the BP_America dataset. The application of BP_America dataset shows the predictive accuracy of the automated sentiments to be higher than the manual. Table 6.4 shows the misclassification of 44.26% of automated sentiments wrongly predicted, compared with their original labels. In the same way, the second row of Table 6.4 shows the misclassification of 48.26% of manual sentiment which after training were wrongly predicted. Again, to further demonstrate the ratio between sensitivity and specificity, each fold in terms of manual sentiment is tested by ROC curve as shown in Figure 6.11a; and Figure 6.11b, the barplot which shows a comparison of TP, TN, FP and FN is drawn. Similarly, to demonstrate the ratio between sensitivity and specificity, each fold in terms of automated sentiment is shown



(a) ROC curve.

(b) Barplot of FP, TP, TN and FN.

Figure 6.9: Predictive modelling results of Saudi_Aramco manual sentiment.



(a) ROC curve.

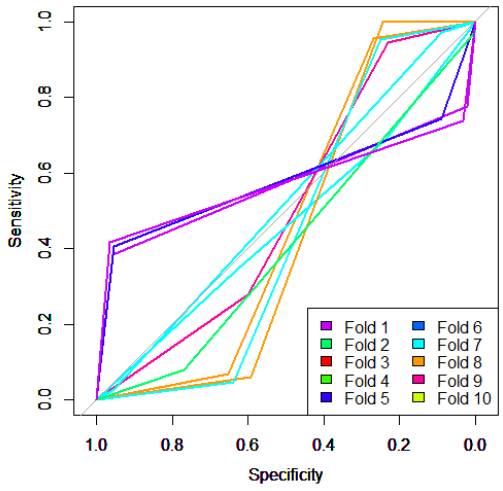
(b) Barplot of FP, TP, TN and FN.

Figure 6.10: Predictive modelling results of Saudi_Aramco automated sentiment.

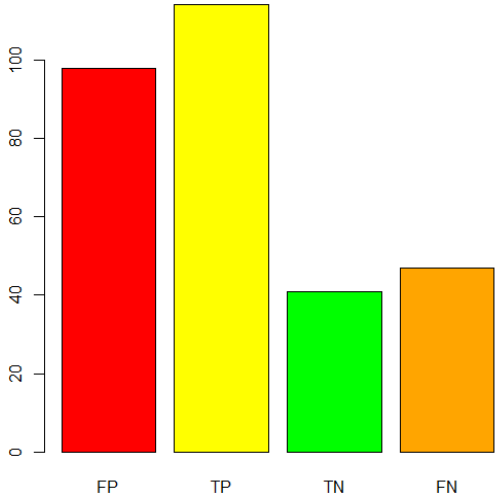
in Figure 6.12a; and Figure 6.12b, shows the barplot which shows a comparison of TP, TN, FP and FN.

Table 6.4: Predictive accuracy of manual and automated sentiment BP_America dataset

Sentiments	**FP**	**TP**	**TN**	**FN**	**Accuracy**	**Senstivity**	**Specificity**
Manual	97.9	114.2	41.0	46.9	0.517333	0.7088764	0.2951763
Automated	44.0	136.8	30.4	88.8	0.557333	0.6063829	0.4086021

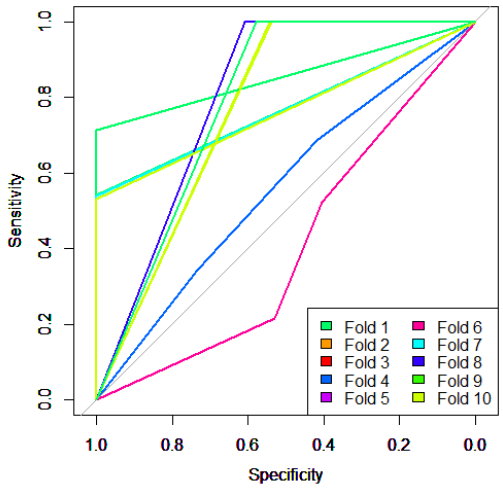


(a) ROC curve.

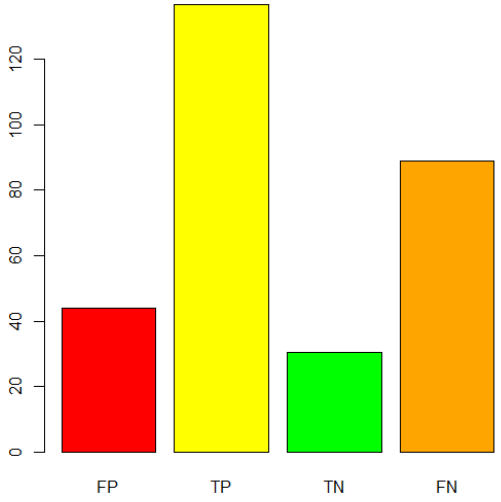


(b) Barplot of FP, TP, TN and FN.

Figure 6.11: Predictive modelling results of BP_America manual sentiment.



(a) ROC curve.



(b) Barplot of FP, TP, TN and FN.

Figure 6.12: Predictive modelling results of BP_America automated sentiment.

6.4 Hybrid Sentiment Analysis

This section describes a novel predictive modelling mechanism which uses automated sentiment as an adjustment factor to investigate the accuracy of manual sentiment in each cluster. In section 6.3 predictive modelling was based on the types of sentiments whereas this section demonstrates the percentage of error in predicting the actual sentiment of tweets by providing training on the basis of the number of tweets and the automatically computed sentiments. The reason for using automatically computed sentiments as a part of training is because of the previously computed results; where training and testing were done individually on the basis

of manual and automated sentiments. The results were mentioned previously are not giving accuracy on the basis of true sentiments which can only be possible by using automated sentiments as an adjustment factor which can further reduce the error while calculating the predictive sentimental accuracy of the tweet.

6.4.1 Experiment Method/Algorithm

The rationale of the experiment in this section was to demonstrate predictive modelling by training the k-means algorithm on the basis of pre-processed tweets and the automatically computed sentiments. The total length of each set was 3000. The first half of each dataset was selected for training and the remaining half for testing. This split was validated by repeating it five times across ten folds. The value of the k centres in k-means was empirically set to 20. The basic idea of applying adjustment is first to cluster training tweets and then look for true sentiment based on crowdsourcing, compared to automated sentiment. Next the clusters were predicted in testing tweets, followed by applying the adjustment in order to reduce the misclassification error. The objective was to use automated sentiment as an adjustment factor to investigate the accuracy of the manual sentiments in each cluster. Groups of similar tweets were expected to be mislabelled in the same way; this technique could be used to devote a small amount of crowdsourcing to improving the predictive accuracy.

To clarify this method, it was applied on cluster (3) as stated in Table 6.5. This cluster contains (26) tweets, all of which were marked as negative by AlchemyAPI. The ground truth from crowdsourcing indicates the correct values to be: (1) positive, (1) negative and (24) neutral. It is marked as negative, a score of 0.038 was added to the total for positive; 0.038 to negative and 0.923 to neutral. That is, 3.8% of the tweet were marked negative categorized correctly; 3.8% should have been marked as positive, and 92.3% should have been neutral. These values were used to define adjustment rules for any new addition to this cluster as follows: if tweet is marked positive or neutral by AlchemyAPI, it is left unchanged.

6.4.2 Results of Saudi_Aramco Dataset

The output of training and testing of the Saudi_Aramco dataset is shown in Tables 6.5 and 6.6. Table 6.5 shows the training outputs of the arrangement of tweets based on the predicted automated sentiments and actual manual sentiment in each cluster. Table 6.6 show the testing outputs of automated and manual sentiments. In order to calculate the error in each cluster, based on the training and testing results, first the difference between the actual and predicted sentiments was calculated by dividing the training count of manual sentiments in each cluster

by the sum of the predicted automated sentiments in each cluster as shown in Table 6.7. This gives the ratio of adjustments to predicted automated sentiments according to the count of actual positive, negative and neutral manual sentiments of the Saudi_Aramco dataset in each cluster. Further in Table 6.8 the adjustment predictions were calculated by taking the product of adjustment to the automated predictions to find the cluster-wise sentimental error by calculating its difference from the resulting actual manual sentiments. Last in Table 6.9 the percentage of misclassification of actual manual sentiments based on the automated adjustment predictions was shown, which leads to the calculation of the percentage of total misclassification error of each cluster. It can be seen in Table 6.9 that those clusters having high misclassification error such as clusters 3, 7, 8, 11, 17 and 18 can be considered misbehaving in comparison to the rest, which should be considered well-behaved clusters.

Table 6.5: Training results of automated and manual sentiment of Saudi_Aramco dataset

Count of clusters	**Count of prediction (automated)**			**Count of actual (manual)**			**Grand Total**
	1	-1	0	1	-1	0	
1		1				1	1
2		3		1		2	3
3		26		1	1	24	26
4			51	4	5	42	51
5		10		1	1	8	10
6	62	18		19	2	59	80
7	11	4		2		13	15
8	7			4		3	7
9			2	1		1	2
10		3		1		2	3
11		7		2		5	7
12			39	4	2	33	39
13	394			126	28	240	394
14			447	54	20	373	447
15		265		19	2	244	265
16		1		1			1
17	25			5	1	19	25
18		42		2	12	18	42
19			29	5	3	21	29
20	53			14	5	34	53

Additionally, to see the number of tweets clustered and categorized, a histogram chart is shown

in Figure 6.13, which diagrammatically gives the reader a view of the test results of the tweets clustered and labelled according to category. This chart shows the similarity between categories of tweets lying in the same cluster and also the opposite.

Table 6.6: Testing results of automated and manual sentiment of Saudi_Aramco dataset

Count of clusters	**Count of prediction (automated)**			**Count of actual (manual)**			**Grand Total**
	1	-1	0	1	-1	0	
1		2				2	2
2		5				5	5
3		23		4	2	17	23
4			40	8	1	31	40
5		8		1		7	8
6	51	19		14	6	50	70
7	6	9		2		13	15
8	5			2	2	1	5
9			5	2		3	5
10		3		1		2	3
11		7			2	5	7
12			34	4	1	29	34
13	400			119	33	248	400
14			479	58	16	405	479
15		261		16	1	244	261
16							0
17	30			7	2	21	30
18		48		4	7	37	48
19			13		3	10	13
20	52			18	3	31	52

Table 6.7: Adjustment to automated sentiment which is computed by calculating the ratio between actual manual and predicted automated sentiment in training dataset of Saudi_Aramco

Cluster	**Prediction (automated)**			**Actual (manual)**			**Adjustment to automated**		
	1	-1	0	1	-1	0	1	-1	0
1		1				1	0	0	1
2		3		1		2	0.3333	0	0.666667
3		26		1	1	24	0.038462	0.038462	0.923077
4			51	4	5	42	0.078431	0.098039	0.823529
5		10		1	1	8	0.1	0.1	0.8
6	62	18		19	2	59	0.3064	0.1111	0.7375
7	11	18		2		13	0.181818	0	0.8666
8	7			4		3	0.571429	0	0.428571
9			2	1		1	0.5	0	0.5
10		3		1		2	0.333333	0	0.666667
11		7		2		5	0.285714	0	0.714286
12			39	4	2	33	0.102564	0.051282	0.846154
13	394			126	28	240	0.319797	0.071066	0.609137
14			447	54	20	373	0.120805	0.044743	0.834452
15		265		19	2	244	0.071698	0.007547	0.920755
16		1		1			1	0	0
17	25			5	1	19	0.2	0.04	0.76
18		42		2	12	28	0.047619	0.285714	0.4285
19			29	5	3	21	0.172414	0.103448	0.724138
20	53			14	5	34	0.264151	0.09434	0.641509

Table 6.8: Adjustment to prediction (automated) sentiment and error calculation in testing dataset of Saudi_Aramco

Cluster	**Raw predictions (automated)**			**Adjusted predictions**			**Actual (manual)**			**Error**		
	1	-1	0	1	-1	0	1	-1	0	1	-1	0
1		2		0	0	2	0	0	2	0	0	0
2		5		1.67	0	3.33	0	0	5	0	0	0.33
3		23		0.88	0.88	21.23	4	2	17	0.78	0.56	0.25
4			40	3.14	3.92	32.94	8	1	31	0.61	2.92	0.06
5		8		0.8	0.8	6.4	1	0	7	0.2	0	0.09
6	51	19		15.63	2.11	51.63	14	6	50	0.15	0.64	0.03
7	6	9		1.09	0	13	2	0	13	0.45	0	0
8	5			2.86	0	2.14	2	2	1	0.43	1	1.14
9			5	2.5	0	2.5	2	0	3	1	0	0.17
10		3		1	0	2	1	0	2	0	0	0
11		7		2	0	5	0	2	5	0	1	0
12			34	3.49	1.74	28.77	4	1	29	0.13	0.74	0.01
13	400			127.92	28.43	243.65	119	33	248	0.08	0.14	0.02
14			479	57.87	21.43	399.7	58	16	405	0	0.34	0.01
15		261		18.71	1.97	240.32	16	1	244	0.17	0.97	0.02
16				0	0	0	0	0	0	0	0	0
17	30			6	1.2	22.8	7	2	21	0.14	0.4	0.09
18		48		2.29	13.71	20.57	4	7	37	0.42	0.96	0.44
19			13	2.24	1.34	9.41	0	3	10	0	0.55	0.06
20	52			13.74	4.91	33.36	18	3	31	0.24	0.64	0.08

Table 6.9: Calculation of misclassified sentiment classes and the total error percentage of Saudi_Aramco dataset

Cluster	**Calculation of Misclassification**			**Percentage of total error**
1	0	0	0	
2	0	0	1.67	33.4
3	3.12	1.12	4.23	30.22
4	4.86	2.92	1.94	24.3
5	0.2	0.8	0.6	20
6	1.63	3.89	1.63	10.20
7	0.91	0	0	6.06
8	0.86	2	1.14	80
9	2	0	0.5	50
10	0	0	0	0
11	2	2	0	133.33
12	0.51	0.74	0.23	4.37
13	8.92	4.57	4.35	4.46
14	0.13	5.43	5.30	2.27
15	2.71	0.97	3.68	2.82
16	0	0	0	0
17	1	0.8	1.8	12
18	1.71	6.71	16.43	51.79
19	2.24	1.66	0.59	34.48
20	4.26	1.91	2.36	16.40

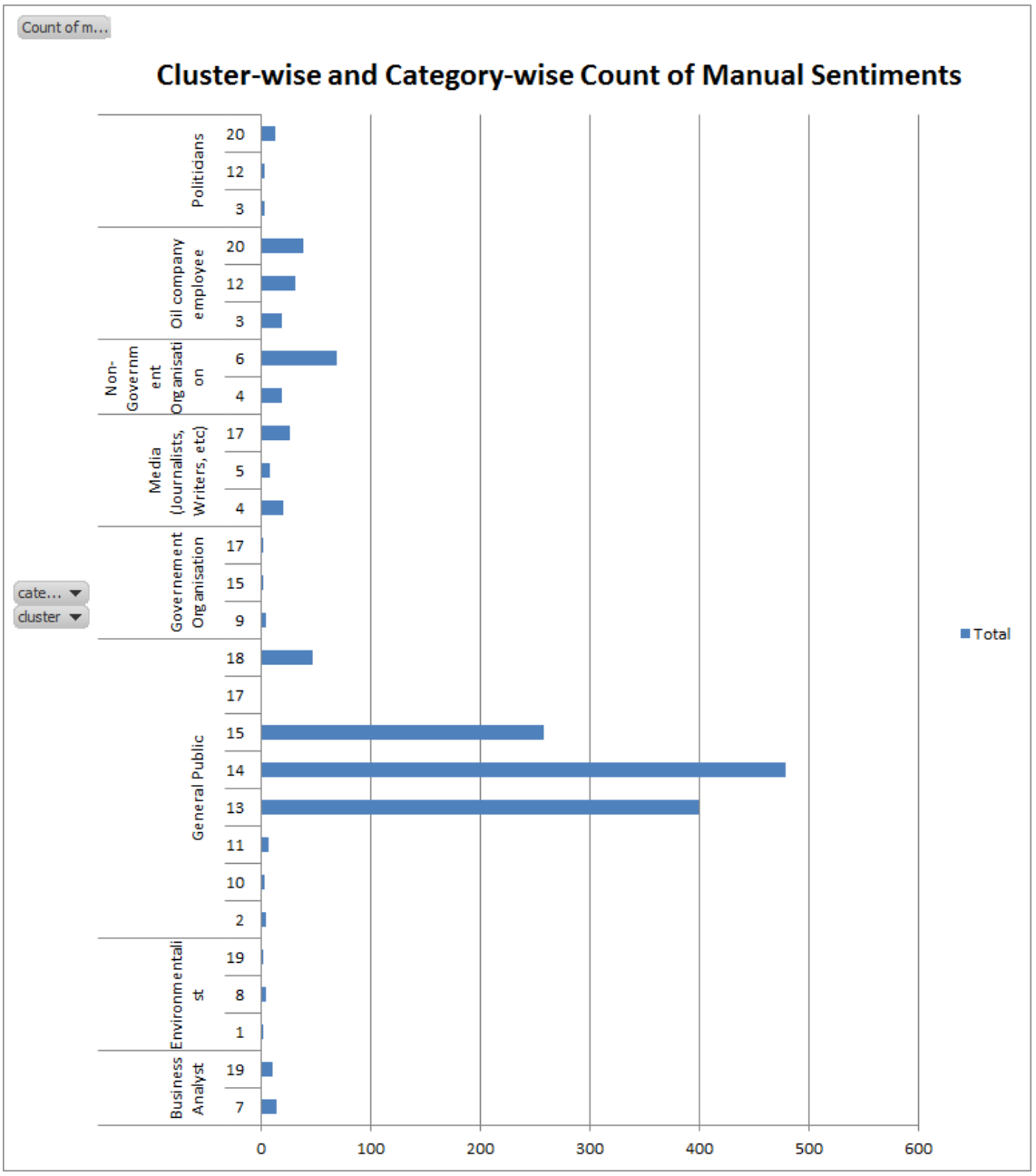


Figure 6.13: Count of categories in each cluster based on manual sentiment of Saudi_Aramco dataset.

6.4.3 Results of BP_America Dataset

The output of training and testing of BP_America tweets is shown in Tables 6.10 and 6.11. Table 6.10 shows the training output of the arrangement of tweets based on the predicted automated sentiments and actual manual sentiment in each cluster. Table 6.11 shows the testing outputs of automated and manual sentiments. In order to calculate the error in each cluster derived from the training and testing results, the difference between the actual and predicted sentiments was calculated by dividing the training count of manual sentiment in each cluster by the sum of the predicted automated sentiment in each cluster which shown in Table 6.12. This gave the ratio of adjustment to the predicted automated sentiment according to the count of actual positive, negative and neutral manual sentiments in each cluster. Second in Table 6.13 the adjustment predictions were calculated by taking the product of adjustment to the automated predictions and the testing results of the automated sentiments. The reason for calculating the adjustment to the automated predictions is to find the cluster-wise sentimental error by calculating its difference from the resulting actual manual sentiments. Last, in Table 6.14 the percentage of misclassification of actual manual sentiments based on the automated adjustment predictions was shown, which lead to the calculation of the total misclassification error of each cluster. It can be seen from Table 6.14 that those clusters having a high misclassification error such as clusters 20, 19 and 11, can be considered somewhat misbehaving in comparison to the others in which some of them, such as 3, 5, 6, 7 and 9, can be considered moderate whereas the remaining were identified as well-behaved predicted clusters.

Furthermore, to see the number of tweets clustered and categorized, a histogram chart is shown in Figure 6.12 which diagrammatically gives the reader a view of the test results of tweets clustered and labelled according to category. This chart clearly shows the similarity between the categories of tweets lying in the same cluster and also the opposite.

Table 6.10: Training results of automated and manual sentiment of BP_America dataset

Count of clusters	**Count of prediction (automated)**			**Count of actual (manual)**			**Grand Total**
	1	-1	0	1	-1	0	
1	95			7	1	87	95
2	75			8	1	66	75
3	16			3		13	16
4			81	15	3	63	81
5		39		3	4	32	39
6	32			9	2	21	32
7	80			16	3	61	80
8	95			13	9	73	95
9		156		38	8	110	156
10	19			1	3	15	19
11	155			31	6	118	155
12		35		8	5	22	35
13	32			4	1	27	32
14			224	43	12	169	224
15	87			15	9	63	87
16	89			18	5	66	89
17	43			12	2	29	43
18		107		24	4	79	107
19	15			1	3	11	15
20		25		7	1	17	25

Table 6.11: Testing results of automated and manual sentiment of BP_America dataset

Count of clusters	**Count of prediction (automated)**			**Count of actual (manual)**			**Grand Total**
	1	-1	0	1	-1	0	
1	102			8	3	91	102
2	62			8	1	53	62
3	12			3		9	12
4			74	12	3	59	74
5		28		5	2	21	28
6	36			9	2	25	36
7	72			8	3	61	72
8	78			15	4	59	78
9		176		27	11	138	176
10	18			1	2	15	18
11	138			16	6	116	138
12		38		7	3	28	38
13	37			3		34	37
14			233	50	12	171	233
15	98			21	9	68	98
16	93			22	5	66	93
17	50			10	2	38	50
18		117		21	6	90	117
19	15			2	1	12	15
20		23		2	4	17	23

Table 6.12: Adjustment to automated sentiment which is computed by calculating the ratio between actual manual and predicted automated sentiment in training dataset of BP_America

Training								
Prediction (automated)			**Actual (manual)**			**Adjustment to automated**		
1	-1	0	1	-1	0	1	-1	0
95			7	1	87	0.07	0.01	0.92
75			8	1	66	0.11	0.01	0.88
16			3		13	0.19	0	0.81
		81	15	3	63	0.19	0.04	0.78
	39		3	4	32	0.08	0.1	0.82
32			9	2	21	0.28	0.06	0.66
80			16	3	61	0.2	0.04	0.76
95			13	9	73	0.14	0.09	0.77
	156		38	8	110	0.24	0.05	0.71
19			1	3	15	0.05	0.16	0.79
155			31	6	118	0.2	0.04	0.76
	35		8	5	22	0.23	0.14	0.63
32			4	1	27	0.13	0.03	0.84
		224	43	12	169	0.19	0.05	0.75
87			15	9	63	0.17	0.1	0.72
89			18	5	66	0.2	0.06	0.74
43			12	2	29	0.28	0.05	0.67
	107		24	4	79	0.22	0.04	0.74
15			1	3	11	0.07	0.2	0.73
	25		7	1	17	0.28	0.04	0.68

Table 6.13: Adjustment to prediction (automated) sentiment and error calculation in testing dataset of BP_America

Cluster	**Raw predictions (automated)**			**Adjusted predictions**			**Actual (manual)**			**Error**		
	1	-1	0	1	-1	0	1	-1	0	1	-1	0
1	102			7.14	1.02	93.84	8	3	91	0.11	0.66	0.032
2	62			6.82	0.62	54.56	8	1	53	0.15	0.38	0.03
3	12			2.28	0	9.72	3	0	9	0.24	0	0.08
4			74	14.06	2.96	57.72	12	3	59	0.17	0.02	0.02
5		28		2.24	2.8	22.96	5	2	21	0.552	0.4	0.09
6	36			10.08	2.16	23.76	9	2	25	0.12	0.08	0.04
7	72			14.4	2.88	54.72	8	3	61	0.8	0.04	0.11
8	78			10.92	7.02	60.06	15	4	59	0.272	0.755	0.018
9		176		42.24	8.8	124.96	27	11	138	2.84	0.2	0.094
10	18			0.9	2.88	14.22	1	2	15	0.1	1.76	11.7
11	138			27.6	5.52	104.88	16	6	116	0.725	2.88	1289.92
12		38		8.74	5.32	23.94	7	3	28	0.25	0.77	0.145
13	37			4.81	1.11	31.08	3	0	34	0.61	0	0.08
14			233	44.27	11.65	174.75	50	12	171	0.12	0.03	0.02
15	98			16.66	9.8	70.56	21	9	68	0.21	0.08	0.04
16	93			18.6	5.58	68.82	22	5	66	0.15	0.116	0.04
17	50			14	2.5	33.5	10	2	38	0.4	0.25	0.12
18		117		25.74	4.68	86.58	21	6	90	0.23	0.22	0.038
19	15			1.05	3	10.95	2	1	12	0.475	2	0.08
20		23		6.44	0.92	15.64	2	4	17	2.22	0.77	0.08

Table 6.14: Calculation of misclassified sentiment classes and the total error percentage of BP_America dataset

Calculation of Misclassification			**Percentage of total error**
1	-1	0	
0.484210526	1.92631579	2.41052632	4.726522188
1.386666667	0.17333333	1.56	5.032258065
0.75	0	0.75	12.5
1.703703704	0.25925926	1.44444444	4.604604605
2.846153846	0.87179487	1.97435897	20.32967033
1.125	0.25	1.375	7.638888889
6.4	0.3	6.1	17.77777778
4.326315789	3.38947368	0.93684211	11.09311741
15.87179487	1.97435897	13.8974359	18.03613054
0.052631579	0.84210526	0.78947368	9.356725146
11.6	0.65806452	10.9419355	16.8115942
1.685714286	2.42857143	4.11428571	21.65413534
1.625	1.15625	2.78125	15.03378378
5.272321429	0.48214286	4.79017857	4.525597793
4.103448276	1.13793103	2.96551724	8.374384236
3.191011236	0.2247191	2.96629213	6.862389755
3.953488372	0.3255814	4.27906977	17.11627907
5.242990654	1.62616822	3.61682243	8.962377187
1	2	1	26.66666667
4.44	3.08	1.36	38.60869565

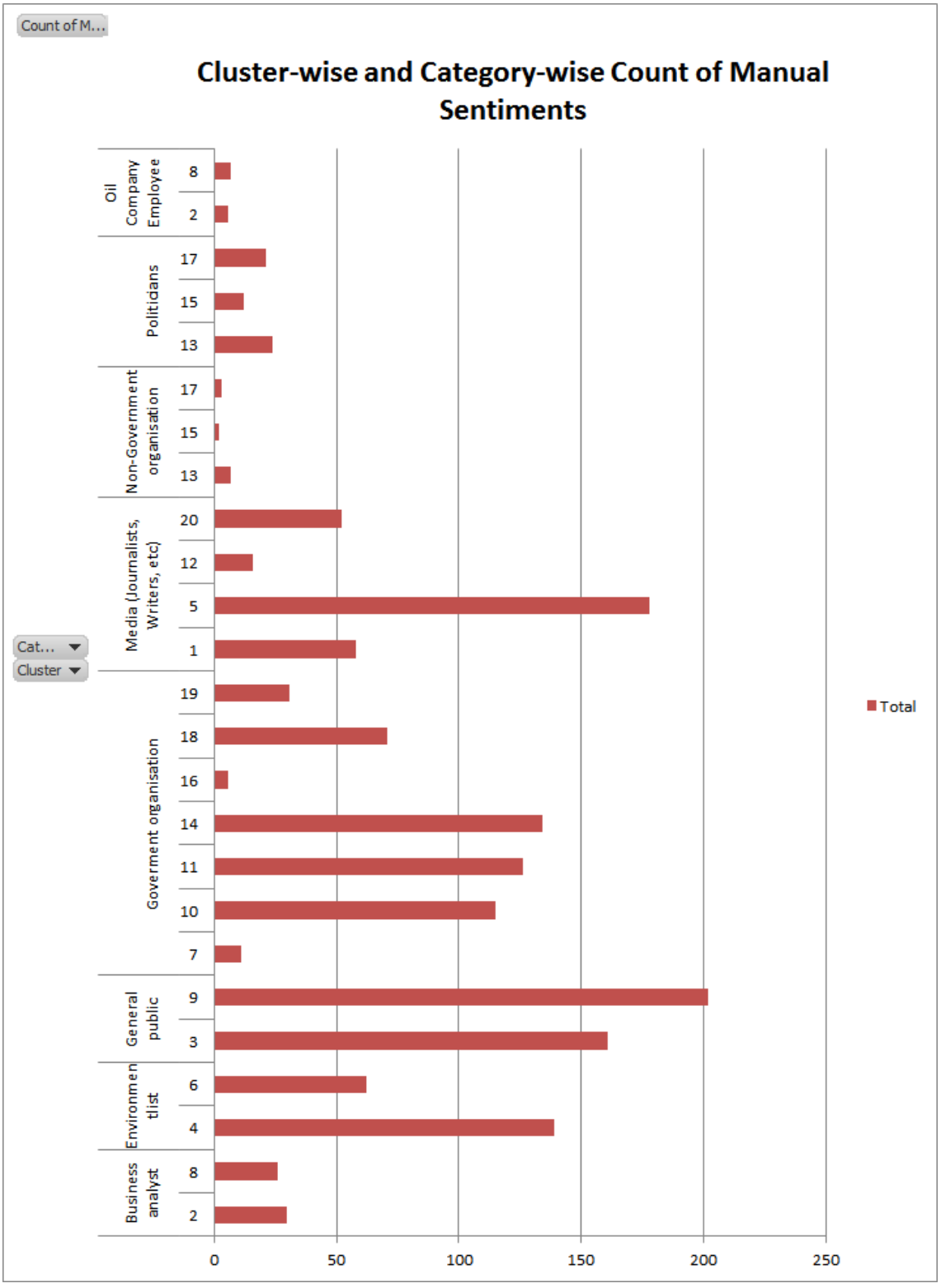


Figure 6.14: Count of categories in each cluster based on manual sentiment of BP_America.

6.5 Conclusion

This chapter has presented three major models of the implicit clustering of tweets in two extracted datasets of tweets from the companies Saudi_Aramco and BP_America:

1. Clustering on the basis of categories of users
2. Clustering on the basis of automated and manual sentiments
3. Clustering on the basis of tweets

First, the unsupervised prognosis model was described which clusters tweets based on the basis of the similarity between them and finds the relationships between the categories of user in each cluster. Second, the unsupervised prognosis model was described which clustered the tweets into three sentiment polarities: positive, negative and neutral. The results show the predictive accuracy of testing tweets after training. The predictive accuracy regarding the Saudi_Aramco dataset produced higher accuracy for manual sentiments than for automated sentiments; whereas in the case of the BP_America dataset, the accuracy for automated sentiments was higher than for manual sentiments. Third, a hybrid sentiment analysis was applied to both the BP_America and Saudi_Aramco datasets. A novel criterion for calculating the misclassification error was described: using automated sentiments as an adjustment factor to calculate the misclassification errors of actual manual sentiments for both datasets. The results show the misclassification in each cluster which helped to show the resulting well-behaved, moderate and misbehaving clusters.

Chapter 7

Conclusion and Future Work

The final chapter of this thesis provides a summary of the research with the possible future direction. Initially, the chapter explores the thesis contributions. It ends by focusing on future work that is related to this research.

7.1 Thesis Summary and Contributions

Returning to the research hypothesis which examined in this thesis "automated Twitter analysis tools can be a reliable and effective means of data interpretations for companies to make proactive or reactive decisions regarding communications with their stakeholders, provided they take account of their local cultural environment". The aim was to investigate the use of Twitter by "controversial" companies and other users. In particular, it looks at the nature of positive and negative sentiment towards oil companies and asks how this relates to the cultural effects of the network structure. This has required an evaluation of the current automated tools for sentiment analysis, and the development of improved methods based on user classification. In order to achieve this, three main experiments were implemented which are sentiment analysis, text categorization and clustering. These experiments were conducted on datasets extracted from Twitter belonging to two oil companies BP_America and Saudi_Aramco. Datasets divided into two types, the companies' tweets and mentions; along with their details are presented in chapter 3.

In chapter 4 two sentiment analysis methods were applied to the dataset to investigate the effectiveness of automated tools compared to the manual method. The first is a crowdsourced method which involves human annotating tweets with the sentiment polarity conveyed. The second one used is an automated using the AlchemyAPI. To investigate if the automated tool accuracy comparable to the crowdsourcing one, the sentiment analysis was carried out in three phases. In the first phase the tweets and mentions were classified as positive, negative or neutral using both the manual and automated methods. This phase results revealed that automated sentiment is often correctly identified the positive sentiment in companies' tweets but not with

the negative sentiment. On the other hand, there are more negative sentiment conveyed in the mentions when compared to the companies' tweets. This may be explained by the reasoning that, while companies are more likely to tweet positive things, regular users will express their opinions on the activities of the oil companies without constraint. However, the actual variation between manual and automated is quite high. The second phase, classified the tweets and mentions in binary classes namely: positive vs. non-positive or negative vs. non-negative by merging the neutral group once with positive and once with negative respectively. This experiment was applied for further investigation about the nature of misclassification. Motivated by the observation that manual classifiers only differ in interpreting neutral tweets. The results showed that the sentiment methods performed better with binary classification when tweets and mentions were classifies as negative and non-negative (positive + neutral). This indicates that misclassification occurred more with positive and neutral classes. In the third phase the Twitter accounts of users and organisations mentioning the oil companies were categorized into eight main groups, then investigate the accuracy of both sentiment methods within the groups. The aim of this analysis was to investigate the role of user category in the sentiment accuracy. BP_America results showed that the automated method performed better with some user categories such as environmentalist, business analysts, oil company employee and politicians while less performance with other groups. In addition, Saudi_Aramco results revealed that the automated method not performed well with general public and non-government organisation groups while the sentiment accuracy was better with other user categories.

Chapter 4 concludes that automated methods tends not to be accurate enough for all domains and all types of data. Despite this, automated methods allow sentiment classification of large dataset within a short time frame. The automated sentiment analysis showed a better results and reliable when analysing tweets originated from the companies because conventional wisdom would dictate that the companies' tweets are written by PR departments and consequently subjected to intense scrutiny before posting. In addition, the more direct and official tweets such as those that are originated from some users groups who represent professional bodies (e.g. governmental organizations and politicians). As a results of the two oil companies are located in two different culturally distinct, tweets generated from more conservative culture with less of sarcasm or slang terms which easier to analyse automatically. The exploration of manual and automated sentiment analysis in individual tweet revealed that there are problems with a little data such as tweets from some user groups like general public. Companies need to understand the difference between "more data" and "more valuable data" [151]. When they focus on specific types of data which not consider a big data to understand which are the important attributes and which are just noise then tend to solve the small issues. This can provide companies with more valuable insight to Twitter users.

After analysing individual tweets as discussed above the structure of user groups was investig-

ated in chapter 5. Multi-text categorization models were used for automatically categorize the users who mentioning the two oil companies in their tweet into a number of pre-defined classes. In addition, to investigate the role of groups based on the accuracy of the sentiments. Four different machine learning classifiers, namely, support vector machine (SVM), k-Nearest Neighbors (KNN), Naïve Bayes (NB) and Decision Tree (DT) were applied to the datasets. The task was performed by extracting 11 key features from tweets and subjecting them to the classifiers. The value of these features which reflect a deeper understanding of the users and their network structure. Chapter 5 proposed an approach which is similar to predicting the authorship of textual data. From the experiment results, the performance of classifiers was found different based on the characteristic of data. The results clearly showed SVM outperformed other classifiers while classifying Saudi_Aramco datasets, whereas Naïve Bayesian Classifier gave the highest classification accuracy for BP_America dataset. However, some features play an important role in the prediction accuracy such as N-gram and user description. Generally, classifiers showed poor results with BP_America dataset but performed very well with the Saudi_Aramco dataset. This can be attributed to the cultural difference as well, the most likely causes of that is BP_America dataset was containing very noisy tweets with various labels, abbreviations and irregular form which can affect the accuracy of prediction.

Chapter 6 demonstrates the clustering of tweets derived from both oil companies based on different features. The number of tweets was predicted rather than an individual tweet and then investigate the accuracy of the sentiment and find the relationships between the categories of user in each cluster. To achieve that, k-means algorithm was applied to dataset in three different experiments. The first one is the unsupervised model which considering the whole dataset as a training to find the similarity in tweets based on categories of user and the relationships between them in each cluster. The results revealed that some user categories often come together in one cluster for example in Saudi_Aramco dataset general public, media and non-government organisation appeared together in many clusters while environmentalist, government organisation, oil company employee and politicians clustered together in different clusters. That indicates the user groups who usually clustered together have similar interest or behaviour. In the second, predictive model was proposed which clustered the tweets into three sentiment polarities: positive, negative and neutral. The results showed that the predictive accuracy of testing tweets after training. The predictive accuracy regarding the Saudi_Aramco dataset produced higher accuracy for manual sentiments than for automated sentiments; whereas in the case of the BP_America dataset, the accuracy for automated sentiments was higher than for manual sentiments. Third, a hybrid sentiment analysis model. A novel criterion for calculating the misclassification error was described: using automated sentiments as an adjustment factor to calculate the misclassification errors of actual manual sentiments for both datasets. The results showed that the misclassification in each cluster which helped to show the resulting well-behaved, mod-

erate and misbehaving clusters. To conclude from all experiments that have been done in this theses, the cultural differences can be a significant factor as expected to play an important role in shaping the dialogue between oil companies in the Middle East and Western countries and the public. These differences can affect the structure of users, sentiment accuracy and the quality of prediction. At the same time, there are common issues between both companies' datasets. A consequence of the social media sites nature that enabled people to keep their personal information and location it was very difficult to fully explore the cultural differences factor. On the other hand, due to the data limitation in this research the obtained results were not enough to differentiate between both cultures in details.

7.2 Future Work

This thesis has made a substantial step in addressing the accuracy in sentiment and prediction analysis for oil business sector. Throughout this work a number of research problems have been identified and needs to be addressed.

To extend this thesis work further research can be done in the sentiment analysis area. Different automated sentiment analysis method can be used such as SentiStrength which accept various languages. Arabic language is very important to take in account when social media data belong to companies from Middle East are analysed. In addition, comparison between the results outcome from Twitter and Facebook or another popular social media can give a deep insight about people behaviour and cultures which enable companies to improve their online communication strategy with public.

As a future research, all tweets written by a particular user can be used. This extra users' information can be used to increase accuracy of categorization and clustering models. For example, there are certain users termed as 'leaders' (the user who tweet a lot) and there are other users termed as 'follower' who follow the leaders (the users who retweet leader's tweets). These concepts can be applied to increase accuracy of our model and can help in overcoming certain limitation of the system.

Another interesting area for further research is building a framework for detecting signals from Twitter data for oil companies. To get the real value out of social media companies need to analyse large amounts of data in order to find interesting information. This process can be simplified by automatically looking for special data patterns, which are likely to warrant further investigation. Simple algorithms are proposed for this analysis namely *Static Signal Detector* and *Gaussian Signal Detector* and Gaussian Signal Detector which can effectively be used to detect any spikes, signals, and abnormality in given data. This research focuses on the role of PR

executive, Community manager and Campaign manager. That can be helpful for oil companies and their campaign to understand country specific issues, engage to address them, and/or target future PR or marketing at them.

Appendix A

Saudi_Aramco dataset: The count of manual and automated sentiment of each tweet category-wise in each cluster

Cluster	Positive (M)	Neutral (M)	Negative (M)	Positive (A)	Neutral (A)	Negative (A)
0	**92**	**188**	**25**	**305**	**0**	**0**
Business Analyst	0	2	0	2	0	0
Environmentalist	0	2	2	4	0	0
General Public	76	159	20	255	0	0
Government Organization	1	0	0	1	0	0
Media (Journalist, Writers, etc)	0	8	0	8	0	0
Non-Government Organization	3	5	0	8	0	0
Oil company employee	11	10	3	24	0	0
Politician	1	2	0	3	0	0
1	**39**	**567**	**4**	**0**	**0**	**1**
Business Analyst	0	8	0	0	0	8
Environmentalist	0	2	0	0	0	2
General Public	34	483	3	0	0	520
Government Organization	1	4	0	0	0	5
Media (Journalist, Writers, etc)	1	9	1	0	0	11
Non-Government Organization	1	29	0	0	0	30
Oil company employee	2	26	0	0	0	28
Politician	0	6	0	0	0	6
2	**17**	**222**	**5**	**0**	**243**	**1**
Business Analyst	0	7	0	0	7	0
General Public	14	180	5	0	198	1
Government Organization	1	0	0	0	1	0
Media (Journalist, Writers, etc)	0	9	0	0	9	0
Non-Government Organization	1	6	0	0	7	0
Oil company employee	1	17	0	0	18	0
Politician	0	3	0	0	3	0
3	**1**	**1**	**0**	**0**	**0**	**2**
General Public	1	0	0	0	0	1
Non-Government Organization	0	1	0	0	0	1
4	**2**	**30**	**1**	**33**	**0**	**0**
Environmentalist	1	0	0	0	0	1
General Public	1	23	0	0	0	1
Government Organization	0	1	0	1	0	0
Media (Journalist, Writers, etc)	0	1	0	1	0	0
Non-Government Organization	0	5	0	5	0	0
5	**1**	**13**	**0**	**0**	**14**	**0**
General Public	1	13	0	0	14	0
6	**0**	**4**	**0**	**0**	**0**	**4**
General Public	0	4	0	0	0	4

Cluster	Positive (M)	Neutral (M)	Negative (M)	Positive (A)	Neutral (A)	Negative (A)
7	**1**	**27**	**1**	**1**	**26**	**2**
Business Analyst	1	0	1	0	2	0
General Public	0	24	0	1	21	2
Oil company employee	0	2	0	0	2	0
Politicians	0	1	0	0	1	0
8	**0**	**1**	**0**	**0**	**1**	**0**
General Public	0	1	0	0	1	0
9	**312**	**1017**	**106**	**620**	**695**	**120**
Business Analyst	4	26	3	10	20	3
Environmentalist	8	7	2	7	9	1
General Public	225	759	76	407	564	89
Government Organization	2	6	0	3	4	1
Media (Journalist, Writers, etc)	14	58	6	35	36	7
Non-Government Organization	32	80	10	93	23	6
Oil company employee	22	65	7	46	37	11
Politician	5	16	2	19	2	2
10	**4**	**19**	**1**	**2**	**16**	**6**
Business Analyst	0	1	0	0	0	1
General Public	4	13	0	1	12	4
Non-Government Organization	0	4	0	0	4	0
Oil company employee	0	1	1	1	0	1
11	**2**	**8**	**3**	**9**	**0**	**4**
General Public	2	7	3	8	0	4
Government Organization	0	1	0	1	0	0
12	**4**	**21**	**3**	**11**	**15**	**2**
General Public	3	12	2	5	10	2
Media (Journalist, Writers, etc)	0	2	0	0	2	0
Non-Government Organization	1	2	0	3	0	0
Oil company employee	0	4	0	3	1	0
Politicians	0	1	1	0	2	0
13	**0**	**4**	**0**	**1**	**0**	**3**
General Public	0	3	0	0	0	3
Non-Government Organization	0	1	0	1	0	0
14	**16**	**58**	**8**	**75**	**1**	**6**
Business Analyst	0	3	0	3	0	0
General Public	15	45	8	62	1	5
Government Organization	0	1	0	1	0	0
Media (Journalist, Writers etc)	0	2	0	2	0	0
Non-government Organization	1	2	0	3	0	0
Oil company employee	0	5	0	4	0	1
15	**0**	**2**	**0**	**2**	**0**	**0**
Business Analyst	0	1	0	1	0	0
General Public	0	1	0	1	0	0

Cluster	Positive (M)	Neutral (M)	Negative (M)	Positive (A)	Neutral (A)	Negative (A)
16	**11**	**54**	**3**	**4**	**61**	**3**
Business Analyst	0	2	0	0	1	1
General Public	8	43	2	2	49	2
Government Organization	0	1	0	0	1	0
Media (Journalist, Writers, etc)	1	1	1	1	2	0
Non-Government Organization	1	4	0	0	5	0
Oil company employee	1	3	0	1	3	0
17	**0**	**5**	**0**	**0**	**5**	**0**
General Public	0	4	0	0	4	0
Media(Journalist, Writers, etc)	0	1	0	0	1	0
18	**6**	**50**	**1**	**0**	**57**	**0**
Business Analyst	0	2	0	0	2	0
Environmentalist	0	1	0	0	1	0
General Public	5	42	1	0	48	0
Government Organization	1	0	0	0	1	0
Non-Government Organization	0	1	0	0	1	0
Oil company employee	0	4	0	0	4	0
19	**18**	**22**	**0**	**33**	**5**	**2**
Business Analyst	1	0	0	1	0	0
General Public	14	19	0	27	4	2
Media (Journalist, Writers, etc)	1	0	0	1	0	0
Non-Government Organization	0	1	0	0	1	0
Oil company employee	2	2	0	4	0	0
Grand Total	**526**	**2313**	**161**	**1096**	**1139**	**765**

Appendix B

BP_America dataset: The count of manual and automated sentiment of each tweet category-wise in each cluster

Cluster	Positive (M)	Neutral (M)	Negative (M)	Positive (A)	Neutral (A)	Negative (A)
0	184	314	221	184	220	173
Oil company's employee	9	3	0	9	3	0
Business analysts	11	24	4	11	16	6
Enviromentlists	4	10	31	4	13	23
General public	5	16	90	5	18	73
Goverment organisation	40	30	18	40	31	13
Media (Journalists, Writers, etc)	50	147	26	50	93	19
Non-Government organisation	62	74	33	62	41	22
Politicians	1	6	4	1	1	3
FALSE	2	4	15	2	4	14
1	5	18	149	5	0	0
Business analysts	1	0	1	1	0	0
Enviromentlists	0	4	15	0	0	0
General public	0	7	88	0	0	0
Goverment organisation	4	1	4	4	0	0
Media (Journalists, Writers, etc)	0	5	6	0	0	0
Non-Government organisation	0	0	26	0	0	0
Politicians	0	0	1	0	0	0
FALSE	0	1	8	0	0	0
2	1	1	8	1	0	0
Enviromentlists	0	0	6	0	0	0
General public	0	0	2	0	0	0
Goverment organisation	1	0	0	1	0	0
Non-Government organisation	0	1	0	0	0	0
3	3	4	42	3	0	0
Enviromentlists	0	0	20	0	0	0
General public	0	0	5	0	0	0
Goverment organisation	3	0	5	3	0	0
Media (Journalists, Writers, etc)	0	2	7	0	0	0
Non-Government organisation	0	2	4	0	0	0
FALSE	0	0	1	0	0	0
4	0	0	1	0	0	0
FALSE	0	0	1	0	0	0
5	0	0	1	0	0	1
Non-Government organisation	0	0	1	0	0	1
6	0	2	42	0	0	0
Business analysts	0	1	0	0	0	0
Enviromentlists	0	0	26	0	0	0
General public	0	0	4	0	0	0
Goverment organisation	0	0	2	0	0	0
Media (Journalists, Writers, etc)	0	0	7	0	0	0
Non-Government organisation	0	1	3	0	0	0

Cluster	Positive (M)	Neutral (M)	Negative (M)	Positive (A)	Neutral (A)	Negative (A)
7	0	1	4	0	0	4
Enviromentlists	0	0	1	0	0	1
Media (Journalists, Writers, etc)	0	0	1	0	0	1
Non-Government organisation	0	1	1	0	0	2
FALSE	0	0	1	0	0	0
8	0	0	27	0	26	1
Business analysts	0	0	1	0	1	0
Enviromentlists	0	0	12	0	12	0
General public	0	0	5	0	5	0
Goverment organisation	0	0	1	0	0	1
Media (Journalists, Writers, etc)	0	0	1	0	1	0
Non-Government organisation	0	0	5	0	5	0
FALSE	0	0	2	0	2	0
9	24	44	124	24	18	94
Business analysts	0	4	4	0	1	2
Enviromentlists	1	1	14	1	0	14
General public	7	10	65	7	7	40
Goverment organisation	6	3	4	6	3	6
Media (Journalists, Writers, etc)	6	20	11	6	4	17
Non-Government organisation	4	6	20	4	3	10
FALSE	0	0	6	0	0	5
10	3	4	101	3	0	0
Enviromentlists	0	1	55	0	0	0
General public	1	1	8	1	0	0
Goverment organisation	1	0	11	1	0	0
Media (Journalists, Writers, etc)	1	2	14	1	0	0
Non-Government organisation	0	0	9	0	0	0
Politicians	0	0	2	0	0	0
FALSE	0	0	2	0	0	0
11	4	3	86	4	0	0
Enviromentlists	0	1	49	0	0	0
General public	0	0	4	0	0	0
Goverment organisation	3	0	9	3	0	0
Media (Journalists, Writers, etc)	1	0	13	1	0	0
Non-Government organisation	0	2	8	0	0	0
FALSE	0	0	3	0	0	0
12	1	0	8	1	0	0
Enviromentlists	0	0	4	0	0	0
Media (Journalists, Writers, etc)	1	0	0	1	0	0
Non-Government organisation	0	0	2	0	0	0
FALSE	0	0	2	0	0	0

Cluster	Positive (M)	Neutral (M)	Negative (M)	Positive (A)	Neutral (A)	Negative (A)
13	20	40	13	20	11	61
Oil company's employee	1	0	0	1	0	1
Business analysts	2	14	0	2	0	16
Enviromentlists	0	3	8	0	6	4
General public	0	0	1	0	1	0
Goverment organisation	4	3	1	4	1	7
Media (Journalists, Writers, etc)	7	13	2	7	1	21
Non-Government organisation	6	6	0	6	0	12
Politicians	0	1	0	0	1	0
FALSE	0	0	1	0	1	0
14	38	126	55	38	55	61
Oil company employee	4	0	1	4	3	1
Business analysts	1	18	2	1	7	3
Enviromentlists	1	1	6	1	2	4
General public	1	1	12	1	2	9
Goverment organisation	8	16	1	8	8	5
Media (Journalists, Writers, etc)	12	51	22	12	19	31
Non-Government organisation	11	35	10	11	12	7
Politician	0	3	0	0	1	0
FALSE	0	1	1	0	1	1
15	286	343	378	286	195	327
Oil company employee	8	1	2	8	4	1
Business analysts	5	20	9	5	13	7
Enviromentlists	12	14	35	12	3	30
General public	48	63	188	48	51	144
Goverment organisation	49	23	14	49	14	26
Media (Journalists, Writers, etc)	42	116	64	42	50	51
Non-Government organisation	107	73	51	107	44	51
Politicians	2	5	3	2	3	3
FALSE	13	28	12	13	13	14
16	69	55	38	69	85	10
Oil company employee	4	0	1	4	2	1
Business analysts	3	7	2	3	7	0
Enviromentlists	0	0	6	0	5	1
General public	0	2	13	0	12	2
Goverment organisation	15	5	0	15	6	0
Media (Journalists, Writers, etc)	33	23	8	33	32	6
Non-Government organisation	14	16	4	14	17	0
Politicians	0	0	1	0	1	0
FALSE	0	2	3	0	3	0

Cluster	Positive (M)	Neutral (M)	Negative (M)	Positive (A)	Neutral (A)	Negative (A)
17	2	4	80	2	0	0
Enviromentlists	1	0	37	1	0	0
General public	0	0	10	0	0	0
Goverment organisation	1	1	13	1	0	0
Media (Journalists, Writers, etc)	0	2	14	0	0	0
Non-Government organisation	0	1	5	0	0	0
FALSE	0	0	1	0	0	0
18	1	1	2	1	0	0
Enviromentlists	0	0	1	0	0	0
General public	0	0	1	0	0	0
Goverment organisation	1	0	0	1	0	0
Media (Journalists, Writers, etc)	0	1	0	0	0	0
19	3	7	9	3	2	12
General public	2	6	7	2	1	10
Goverment organisation	1	0	1	1	1	1
Media (Journalists, Writers, etc)	0	0	1	0	0	0
FALSE	0	1	0	0	0	1
Grand Total	644	967	1389	644	612	744

Bibliography

[1] H. Aldahawi and S. Allen, "An approach to tweets categorization by using machine learning classifiers in oil business," in *Computational Linguistics and Intelligent Text Processing*, pp. 535–546, Springer, 2015.

[2] H. Aldahawi and S. M. Allen, "Twitter mining in the oil business: A sentiment analysis approach," in *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pp. 581–586, IEEE, 2013.

[3] H. Aldahawi and S. M. Allen, "Analysing cultural effects of social network usage in business," in *8th UK Social Network Analysis Conference. UKSNA*, 2012.

[4] E. Qualman, *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons, 2010.

[5] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[6] N. B. Ellison *et al.*, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[7] M. Trusov, R. E. Bucklin, and K. Pauwels, "Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site," *Journal of marketing*, vol. 73, no. 5, pp. 90–102, 2009.

[8] N. Ellison, C. Steinfield, and C. Lampe, "Spatially bounded online social networks and social capital," *International Communication Association*, vol. 36, no. 1-37, 2006.

[9] J. Brenner and M. Duggan, "The demographics of social media users," *Consultado en*, 2013.

[10] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Life in the network: the coming age of computational social science," *Science (New York, NY)*, vol. 323, no. 5915, p. 721, 2009.

[11] D. J. Watts, “The" new" science of networks,” *Annual review of sociology*, pp. 243–270, 2004.

[12] Y. Amichai-Hamburger and G. Vinitzky, “Social network use and personality,” *Computers in human behavior*, vol. 26, no. 6, pp. 1289–1295, 2010.

[13] Q. Ye, B. Fang, W. J. He, and J. Hsieh, “Can social capital be transferred cross the boundary of the real and virtual worlds? an empirical investigation of twitter,” *Journal of Electronic Commerce Research*, vol. 13, no. 2, pp. 145–156, 2012.

[14] T. Claburn, “Twitter rule change riles developers,” *Information Week*, no. 1342, 2012.

[15] L. Safko, *The social media bible: tactics, tools, and strategies for business success*. John Wiley & Sons, 2010.

[16] M. Coon, “Social media marketing: Successful case studies of businesses using facebook and youtube with an in-depth look into the business use of twitter,” *Unpublished term project, Stanford University, available at http://www. comm. stanford. edu/coterm/-projects/2010/maddy(accessed on 8 April 2011)*, 2010.

[17] S. Muralidharan, K. Dillistone, and J.-H. Shin, “The gulf coast oil spill: Extending the theory of image restoration discourse to the realm of social media and beyond petroleum,” *Public Relations Review*, vol. 37, no. 3, pp. 226–232, 2011.

[18] G. I. Glossary, “Big data,” 2013.

[19] S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson, and T. Seymour, “The history of social media and its impact on business,” *Journal of Applied Management and entrepreneurship*, vol. 16, no. 3, pp. 79–91, 2011.

[20] B. Hogan, “Analyzing social networks,” *The Sage handbook of online research methods*, p. 141, 2008.

[21] P. Holman and T. Devane, *The Change Handbook: Group Methods for Shaping the Future*. Berrett-Koehler Publishers, 1999.

[22] B. B. Bunker and B. Alban, “Introduction to the special issue on large group interventions,” *Journal of Applied Behavioral Science*, vol. 41, no. 1, pp. 9–14, 2005.

[23] M. R. Manning and G. Faisal Binzagr, “Methods, values, and assumptions underlying large group interventions intended to change whole systems,” *The international Journal of organizational analysis*, vol. 4, no. 3, pp. 268–284, 1996.

[24] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series.," *ICWSM*, vol. 11, no. 122-129, pp. 1–2, 2010.

[25] J. Murphy, M. W. Link, J. H. Childs, C. L. Tesfaye, E. Dean, M. Stern, J. Pasek, J. Cohen, M. Callegaro, P. Harwood, *et al.*, "Social media in public opinion research: Report of the aapor task force on emerging technologies in public opinion research," *American Association for Public Opinion Research*, 2014.

[26] K. N. Hampton, L. S. Goulet, C. Marlow, and L. Rainie, "Why most facebook users get more than they give," *Pew Internet & American Life Project*, vol. 3, 2012.

[27] J. Murphy, J. Edgar, and M. Keating, "Crowdsourcing in the cognitive interviewing process," in *Annual Meeting of the American Association for Public Opinion Research, Anaheim, CA*, 2014.

[28] E. Dean, B. Head, and J. Swicegood, "Virtual cognitive interviewing using skype and second life," *Social Media, Sociality, and Survey Research*, pp. 107–132, 2013.

[29] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, and I. Ounis, "Can twitter replace newswire for breaking news?," 2013.

[30] D. Ediger, K. Jiang, J. Riedy, D. Bader, C. Corley, R. Farber, W. N. Reynolds, *et al.*, "Massive social network analysis: Mining twitter for social good," in *Parallel Processing (ICPP), 2010 39th International Conference on*, pp. 583–593, IEEE, 2010.

[31] K. Lovejoy and G. D. Saxton, "Information, community, and action: how nonprofit organizations use social media*," *Journal of Computer-Mediated Communication*, vol. 17, no. 3, pp. 337–353, 2012.

[32] R. D. Waters and J. Y. Jamal, "Tweet, tweet, tweet: A content analysis of nonprofit organizations' twitter updates," *Public Relations Review*, vol. 37, no. 3, pp. 321–324, 2011.

[33] I. K. Ngugi, R. E. Johnsen, and P. Erdélyi, "Relational capabilities for value co-creation and innovation in smes," *Journal of small business and enterprise development*, vol. 17, no. 2, pp. 260–278, 2010.

[34] N. G. Barnes and J. Andonian, "The 2011 fortune 500 and social media adoption: Have america's largest companies reached a social media plateau?," *University of Massachusetts, Dartmouth (available at www. umassd. edu/cmr/studiesandresearch/bloggingtwitterandfacebookusage)*, 2011.

[35] S. Rybalko and T. Seltzer, "Dialogic communication in 140 characters or less: How fortune 500 companies engage stakeholders using twitter," *Public Relations Review*, vol. 36, no. 4, pp. 336–341, 2010.

[36] S. Wigley and B. K. Lewis, "Rules of engagement: Practice what you tweet," *Public Relations Review*, vol. 38, no. 1, pp. 165–167, 2012.

[37] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American society for information science and technology*, vol. 60, no. 11, pp. 2169–2188, 2009.

[38] J. Yoo, S. Choi, M. Choi, and J. Rho, "Why people use twitter: social conformity and social value perspectives," *Online Information Review*, vol. 38, no. 2, pp. 265–283, 2014.

[39] L. R. Men and W.-H. S. Tsai, "How companies cultivate relationships with publics on social network sites: Evidence from china and the united states," *Public Relations Review*, vol. 38, no. 5, pp. 723–730, 2012.

[40] M. J. Paul, "Interactive disaster communication on the internet: A content analysis of sixty-four disaster relief home pages," *Journalism & Mass Communication Quarterly*, vol. 78, no. 4, pp. 739–753, 2001.

[41] D. K. Wright and M. D. Hinson, "Examining how public relations practitioners actually are using social media," *Public Relations Journal*, vol. 3, no. 3, pp. 1–33, 2009.

[42] B. R. Watson, "Is twitter an alternative medium? comparing gulf coast twitter and newspaper coverage of the 2010 bp oil spill," *Communication Research*, p. 0093650214565896, 2015.

[43] M. Bush, "Tailspin gets a new meaning. american airlines' pr blitz struggles to contain anger from groundings," 2008.

[44] S. J. Andriole, "Business impact of web 2.0 technologies," *Communications of the ACM*, vol. 53, no. 12, pp. 67–79, 2010.

[45] S. Utz, "The (potential) benefits of campaigning via social network sites," *Journal of Computer-Mediated Communication*, vol. 14, no. 2, pp. 221–243, 2009.

[46] M. W. DiStaso, T. McCorkindale, and D. K. Wright, "How public relations executives perceive and measure the impact of social media in their organizations," *Public Relations Review*, vol. 37, no. 3, pp. 325–328, 2011.

[47] Y. Kim, D. Sohn, and S. M. Choi, "Cultural difference in motivations for using social network sites: A comparative study of american and korean college students," *Computers in Human Behavior*, vol. 27, no. 1, pp. 365–372, 2011.

[48] A. Vasalou, A. N. Joinson, and D. Courvoisier, "Cultural differences, experience with social networks and the nature of "true commitment" in facebook," *International Journal of Human-Computer Studies*, vol. 68, no. 10, pp. 719–728, 2010.

[49] Twitter, "About twitter company," 2015.

[50] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network?: The structure of the twitter follow graph," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pp. 493–498, International World Wide Web Conferences Steering Committee, 2014.

[51] M. E. Newman, "Assortative mixing in networks," *Physical review letters*, vol. 89, no. 20, p. 208701, 2002.

[52] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[53] M. E. Newman and J. Park, "Why social networks are different from other types of networks," *Physical Review E*, vol. 68, no. 3, p. 036122, 2003.

[54] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proceedings of the 2nd ACM workshop on Online social networks*, pp. 37–42, ACM, 2009.

[55] L. Rossi and M. Magnani, "Conversation practices and network structure in twitter.," in *ICWSM*, 2012.

[56] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417–424, Association for Computational Linguistics, 2002.

[57] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1275–1284, ACM, 2009.

[58] 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 preprint arXiv:1309.6202*, 2013.

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

[60] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcripts," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 327–335, Association for Computational Linguistics, 2006.

[61] F. Zhu and X. Zhang, "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," *Journal of marketing*, vol. 74, no. 2, pp. 133–148, 2010.

[62] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 478–514, 2012.

[63] A. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," in *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, pp. 1–7, ACM, 2012.

[64] P. Burnap, M. L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, and A. Voss, "Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–14, 2014.

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

[66] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems (TOIS)*, vol. 21, no. 4, pp. 315–346, 2003.

[67] O. Chapelle, B. Schlkopf, and A. Zien, "Semi-supervised learning," 2010.

[68] H. Saif, Y. He, and H. Alani, "Semantic smoothing for twitter sentiment analysis," 2011.

[69] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in *The Semantic Web–ISWC 2012*, pp. 508–524, Springer, 2012.

[70] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, "Precise tweet classification and sentiment analysis," in *Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on*, pp. 461–466, IEEE, 2013.

[71] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[72] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.," in *LREC*, vol. 10, pp. 2200–2204, 2010.

[73] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proceedings of the first ACM conference on Online social networks*, pp. 27–38, ACM, 2013.

[74] E. Cambria, C. Havasi, and A. Hussain, "Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis.," in *FLAIRS conference*, pp. 202–207, 2012.

[75] B. Han, P. Cook, and T. Baldwin, "Lexical normalization for social media text," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 1, p. 5, 2013.

[76] R. González-Ibánez, S. Muresan, and N. Wacholder, "Identifying sarcasm in twitter: a closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 581–586, Association for Computational Linguistics, 2011.

[77] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[78] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," 2007.

[79] Z. Ghahramani, "Unsupervised learning," in *Advanced Lectures on Machine Learning*, pp. 72–112, Springer, 2004.

[80] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, vol. 24, no. 1, pp. 175–186, 2014.

[81] Z. K. Malik, A. Hussain, and J. Wu, "Novel biologically inspired approaches to extracting online information from temporal data," *Cognitive Computation*, vol. 6, no. 3, pp. 595–607, 2014.

[82] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 2014.

[83] Y.-W. Lee, C. Gentile, and R. Kantor, "Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores," *Applied Linguistics*, p. amp040, 2009.

[84] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[85] L. M. Wills, "Automated program recognition: A feasibility demonstration," *Artificial Intelligence*, vol. 45, no. 1, pp. 113–171, 1990.

[86] N. Jindal and B. Liu, "Review spam detection," in *Proceedings of the 16th international conference on World Wide Web*, pp. 1189–1190, ACM, 2007.

[87] T. Zhang, A. Popescul, and B. Dom, "Linear prediction models with graph regularization for web-page categorization," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 821–826, ACM, 2006.

[88] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting stock market indicators through twitter "i hope it is not as bad as i fear"," *Procedia-Social and Behavioral Sciences*, vol. 26, pp. 55–62, 2011.

[89] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[90] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44, ACM, 2010.

[91] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.

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

[93] M. H. Dunham, *Data mining: Introductory and advanced topics*. Pearson Education India, 2006.

[94] P. Cunningham and S. J. Delany, "k-nearest neighbour classifiers," *Multiple Classifier Systems*, pp. 1–17, 2007.

[95] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[96] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[97] R. H. Turi, *Clustering-based colour image segmentation*. Monash University PhD thesis, 2001.

[98] C. Carpineto and G. Romano, "A lattice conceptual clustering system and its application to browsing retrieval," *Machine learning*, vol. 24, no. 2, pp. 95–122, 1996.

[99] R. Tryon and D. Bailey, *Cluster Analysis*. New York, NY: McGraw Hill, 1970.

[100] B. O'Connor, M. Krieger, and D. Ahn, "Tweetmotif: Exploratory search and topic summarization for twitter.," in *ICWSM*, 2010.

[101] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287, Association for Computational Linguistics, 2010.

[102] J. Pöschko, "Exploring twitter hashtags," *arXiv preprint arXiv:1111.6553*, 2011.

[103] D. Antenucci, G. Handy, A. Modi, and M. Tinkerhess, "Classification of tweets via clustering of hashtags," *EECS*, vol. 545, pp. 1–11, 2011.

[104] A. Karandikar, *Clustering short status messages: A topic model based approach*. PhD thesis, University of Maryland, 2010.

[105] M. Cheong and V. Lee, "A study on detecting patterns in twitter intra-topic user and message clustering," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 3125–3128, IEEE, 2010.

[106] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the first workshop on social media analytics*, pp. 115–122, ACM, 2010.

[107] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, pp. 851–860, ACM, 2010.

[108] O. Phelan, K. McCarthy, and B. Smyth, "Using twitter to recommend real-time topical news," in *Proceedings of the third ACM conference on Recommender systems*, pp. 385–388, ACM, 2009.

[109] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506, ACM, 2009.

[110] D. Arthur and S. Vassilvitskii, "How slow is the k-means method?," in *Proceedings of the twenty-second annual symposium on Computational geometry*, pp. 144–153, ACM, 2006.

[111] S. Har-Peled and B. Sadri, "How fast is the k-means method?," *Algorithmica*, vol. 41, no. 3, pp. 185–202, 2005.

[112] C. D. Manning, P. Raghavan, H. Schütze, *et al.*, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.

[113] R. C. De Amorim, "Learning feature weights for k-means clustering using the minkowski metric," 2011.

[114] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[115] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, pp. 675–684, ACM, 2011.

[116] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proceedings of the 20th international conference companion on World wide web*, pp. 57–58, ACM, 2011.

[117] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in *LREC*, vol. 10, pp. 1320–1326, 2010.

[118] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 359–367, Association for Computational Linguistics, 2011.

[119] E. Intelligence, "Petroleum intelligence weekly," 2013.

[120] BP, "Bp america annual report," 2015.

[121] Aramco, "Saudi aramco annual report," 2015.

[122] Twitter, "Twitter help centre: Hashtags," 2014.

[123] H.-C. Chang, "A new perspective on twitter hashtag use: diffusion of innovation theory," *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–4, 2010.

[124] Twitter, "Twitter help center: Links," 2014.

[125] Twitter, "Twitter help center: Retweet," 2014.

[126] Twitter, "Twitter help center: Types of tweets," 2014.

[127] C. KRAUSS and J. SCHWARTZ, "Bp will plead guilty and pay over $4 billion," 2012.

[128] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271, Association for Computational Linguistics, 2004.

[129] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?," *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.

[130] C. Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 286–295, Association for Computational Linguistics, 2009.

[131] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in twitter data with crowdsourcing," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 80–88, Association for Computational Linguistics, 2010.

[132] J. Sprouse, "A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory," *Behavior research methods*, vol. 43, no. 1, pp. 155–167, 2011.

[133] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision making*, vol. 5, no. 5, pp. 411–419, 2010.

[134] J. H. McDonald, *Handbook of biological statistics*, vol. 2. Sparky House Publishing Baltimore, MD, 2009.

[135] J. V. Freeman and S. A. Julious, "The analysis of categorical data," *Scope*, vol. 16, no. 1, pp. 18–21, 2007.

[136] M. K. Buckland and F. C. Gey, "The relationship between recall and precision," *JASIS*, vol. 45, no. 1, pp. 12–19, 1994.

[137] D. Billsus and M. J. Pazzani, "User modeling for adaptive news access," *User modeling and user-adapted interaction*, vol. 10, no. 2-3, pp. 147–180, 2000.

[138] M. A. Ghazanfar, *Robust, scalable, and practical algorithms for recommender systems*. PhD thesis, University of Southampton, 2012.

[139] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in *Proceedings of the fifth ACM conference on Digital libraries*, pp. 195–204, ACM, 2000.

[140] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations," in *AAAI/IAAI*, pp. 187–192, 2002.

[141] K. Aas and L. Eikvil, "Text categorisation: A survey," *Raport NR*, vol. 941, 1999.

[142] M. J. Giarlo, "A comparative analysis of keyword extraction techniques," 2005.

[143] T. Tran, "Combining collaborative filtering and knowledge-based approaches for better recommendation systems," *Journal of Business and Technology*, vol. 2, no. 2, pp. 17–24, 2007.

[144] M. A. Ghazanfar, A. Prügel-Bennett, and S. Szedmak, "Kernel-mapping recommender system algorithms," *Information Sciences*, vol. 208, pp. 81–104, 2012.

[145] U. Y. Nahm and R. J. Mooney, "Text mining with information extraction," in *AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, vol. 1, 2002.

[146] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The adaptive web*, pp. 325–341, Springer, 2007.

[147] M. A. Ghazanfar and A. Prugel, "The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved svd-based recommendations," *Informatica*, vol. 37, no. 1, 2013.

[148] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[149] S. Alag, *Collective intelligence in action*. Manning New York, 2009.

[150] D. Godfrey, C. Johns, C. Meyer, S. Race, and C. Sadek, "A case study in text mining: Interpreting twitter data from world cup tweets," *arXiv preprint arXiv:1408.5427*, 2014.

[151] A. Murdock, "Little data vs. big data: Which one should you use?," 2015.