Chapter 1

Introduction

1.1 Motivation

Organisations today continuously search for new ways to get feedback from their clients in a bid to improve customer satisfaction. Technology firms like Apple, Samsung and Google want to know if their software/hardware products meet their consumers' needs. Merchandise retailers like Walmart and Tesco are constantly trying to make sure they are serving the right products in the right quantity and at for right price. Startups continuously evaluate their products to measure the probability of the company being successful sometime in the future. Postal services like Royal Mail are very interested in how their services are doing and what their customers despise most so they can improve.

Current ways of achieving this include **Surveys** (questionnaires or interviews) and **Focus Groups**. Surveys are very easy to create and distribute. There are also a variety of tools to help with this such as SurveyMonkey[1] and Google Docs[2]. Unfortunately, Surveys also have a few unpleasant drawbacks like time consumption and labour intensity. It can also be difficult to encourage participants to respond. Nevertheless, the main drawback to using Surveys is that some questions are left unanswered while the answers given in answered questions may not reflect the truthful sentiments of the participant. **?** concurs with this and he goes on to discuss how this problem can be solved (to a certain extent) with imputation[3]. **?** also agrees with this point of view and suggests the use of well designed leading questions to put the participant in the right frame of mind. For instance, a leading question like "*How likely will you recommend our service to friends?*" gets the participant thinking about recommendations. While the above solutions might work, they have the same drawbacks as the original problem. Imputation can be very time consuming, labour intensive and error prone while the use of leading questions fails to solve the problem of unanswered questions.

Unfortunately, interviews and focus groups also suffer from false answers due to the fact that they are not anonymous. This means that the participants, in the face of an interviewer, try to be lenient in other not to sound too negative. This could sometimes be due to the fact that participation in the interview/focus group has been incentivised with money or desirable items.

Ideally, the next question we should be asking is "*How can we get the truthful views of our clients about our products and services*"? We need to find a way to get this information without putting any pressure on our clients.

[1] https://www.surveymonkey.com/

[2] https://drive.google.com

[3] Imputation is the process of inferring plausible values for missing entries

1.2 Aims and Objectives

The aim of this project is to investigate other means of getting our customer views and also, how we can make use of Machine Learning and Natural Language Processing techniques to make sense of the data.

Fortunately, the recent surge in the use of social media makes the former relatively easy. People, more often than not, tend to post their truthful feelings about services they use on social media. For instance, Person A buys an iPhone today and realises that the Wi-Fi connectivity is faulty. He/She will most likely post something like "*New iPhone wifi not working #NotCool*" on one or more of the available social networking platforms. From this statement, we can infer that Person A is talking about *the iPhone*, *Wi-Fi* and *Connectivity*. The process of discovering abstract topics in text is called **Topic Modelling** and Chapter 1.5 discusses how we can automate this process.

We try to answer two main research questions. They include:

- Can we use supervised techniques to accurately classify tweets into what is relevant and what is not?
- Can we detect themes/topics in our dataset? If yes, are these topics related to Apple Inc in any way?

1.3 Why Twitter?

Twitter is a social micro-blogging platform where users can share messages in 140 characters. It also allows its users to follow each other. This means, if person A follows person B, A will see public messages from B. These messages are usually referred to as tweets.

Tweets are capped to 140 characters and can contain text, links or a combination of both. They are usually related to either an event, interests or just personal opinion. Facebook posts are mostly always well thought out and each post might include multiple topics. Tweets on the other hand are usually written at the speed of thought.

According to Mashable, DOMO, a Business Intelligence company paired up with Column Five Media to create an infographic[4] about the web back in 2012. It showed that Twitter at the time received around 100,000 tweets per minute. As at 1st February 2014 Twitter claims to receive 500 million tweets a day[5]. That is roughly 350,000 tweets per minute which is over 3 times the amount 2 years before. Twitter also claims to have 241 million monthly users.

Finally, Twitter's data is open compared to other social platforms like Facebook. This means developers are free to tap into this wealth of data in almost real time and free of charge. This makes Twitter a very good source for our data.

1.4 Methodology

This study requires social data and the dataset used was gathered from Twitter between October and November 2013. Each tweet in the dataset is in someway related to Apple Inc and/or their products.

[4]See `http://mashable.com/2012/06/22/data-created-every-minute/`

[5]See https://about.twitter.com/company

We then train a classifier to help filter out as many irrelevant tweets as possible. We briefly analyse different ways to filter the dataset but eventually settle with using Naïve Bayes Classifier. We also look into different ways of analysing the classifier's performance and ways it can be improved.

Finally, we attempt to identify topics/themes in the dataset. We briefly look at Latent Semantic Indexing and why it might not be suitable for our needs. We then look into Latent Dirichlet Allocation, a common approach to topic modelling and use it to detect topics in our dataset. The evaluation of topics generated will be analysed empirically and qualitatively. This means we take a topic and make some assumptions about the semantics of the tweets belonging to that topic. We then analyse the tweets to confirm the validity of our assumption.

1.5 Statement of Originality

This report with any accompanying implementation, is submitted as part requirement for the degree of Computer Science with Industrial Experience at Queen Mary, University of London. I certify that it has not been submitted for any degree or other purposes.

I certify that the intellectual content of this report, to the best of my knowledge, is the product of my own labour except where indicated in the text.