Queen Mary, University of London
Department of Electronic Engineering and Computer Science

Interim Report

Fayimora Femi-Balogun

Submitted in part fulfilment of the requirements for the degree of
BSc Computer Science with Industrial Experience, April 2014

Abstract

Every organisation out there today is constantly looking for ways to improve customer satisfaction. Technology firms like Apple, Samsung and Google want to know if their software/hardware products meets the 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 the 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. The big question is how do they do this?

Social platforms like Facebook and Twitter generate an enormous amount of data on a daily basis. People sometimes use these platforms as an avenue to express their thoughts about products they use. They have discussions with each other about these products and make comparisons.

In this study, we will be making use of Apple Incorporated as a case study. We start by mining Apple related data from Twitter and then we proceed to filtering this data into what is relevant and what isn't. Once we have our relevant data, we will use a mixture of Machine Learning and Natural Language Processing techniques to find common topics in the data. Furthermore, we will analyze the sentiments of the data and investigate how it correlates with the topics. Lastly, we will evaluate the techniques applied to determine which ones work best and why.

‘No amount of experimentation can ever prove me right; a single experiment can prove me wrong.’
Albert Einstein

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Objectives

The main aim of this project is to investigate the use of Machine Learning and Natural Language Processing techniques on social data. Every organisation today is continuously search for new ways to get feedback from their clients/users. Current ways of achieving this include **Surveys** (questionaires or interviews) and **Focus Groups**.

Surveys have the advantage being very easy to create. There are also a variety of tools to help with this. Some of them include SurveyMonkey[1], Google Docs[2]. Unfortunately, Surveys also have a few unpleasant drawbacks like time consumption and labour intensity. It can also be difficult to encourage partcipants to respond. Nonetheless, the main drawback to using Surveys is that the some questions are left unanswered while the answers given may not reflect the truthful sentiments of the participant. [Rubin, 1987] concurs with this and he goes on to discuss how this problem can be solved (to a certain extent) with imputation[3]. [Hayes, 2008] 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.

[1] https://www.surveymonkey.com/
[2] https://drive.google.com
[3] Imputation is the process of inferring plausible values for missing entries

While the above solutions might work, they also 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 linient in other not to sound too negative. This could also somtimes 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 reviews and thoughts about our products and services from customers, voluntarily?*"

1.2 Why Twitter?

Twitter is a social micro-blogging platforms 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 posts 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. This makes it a good source of data.

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.

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. This makes Twitter a perfect source for our data.

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

1.3 Statement of Originality

Statement here.

Chapter 2

Background Theory

2.1 Introduction

2.2 Text Classification

2.3 Topic Modelling

2.4 Sentiment Analysis

2.5 Model Evaluation

Chapter 3

Data Aggregation

First step towards this project is to fetch our data from Twitter. The data is classified into two groups, relevant and irrelevant. We will be spending most of our time with the relevant data.

3.1 Data Classification

To carry out our experiments, we will need to filter out irrelevant tweets. Irrelevant tweets are tweets which we do not really care about. Some examples include:

- *Every day I'm levelling! And now I'm level 19 in #CSRClassics for iPhone! Get it for FREE!*
- *Yes, our apple juice and cider are both GMO-free.*
- *I just had my first carmel apple*

Both tweets could be regarded as relevant but for our use case, they are not. This is because we are only interested in tweets that contain personal opinions about Apple Incorporated. Examples of relevant tweets include: their thoughts

- *Once you get hooked to #Mac, you will definitely go back to #Windows! Lol!*

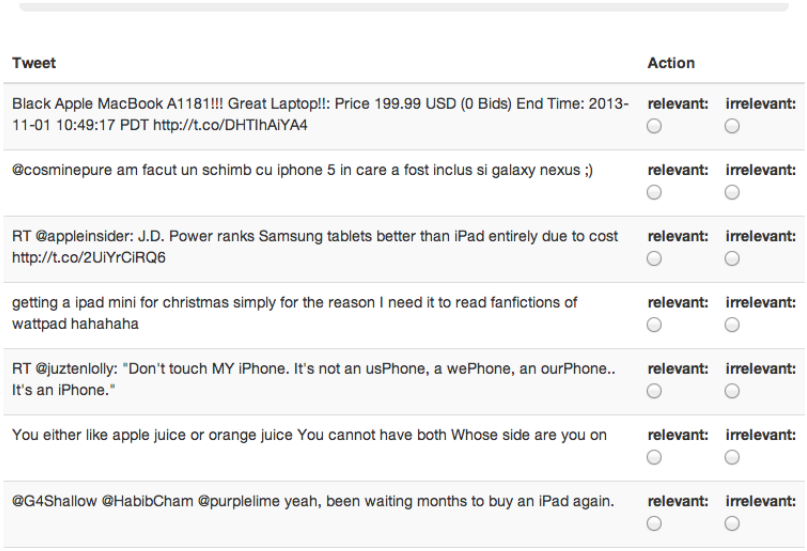

Figure 3.1: The data labelling application

- *If Tim Cook at Apple knows anything about him, it'd be to stay away from Icahn.*

Of course we can manually classify this data but when we have millions of tweets, this becomes impracticable. This is where we employ some classification algorithms to assist us. This is a three step process and we will discuss them in the next sub sections.

3.1.1 Preparing train data

Train data, also known as a training set is a set of data used to train a knowledge database, in this case, a classifier. There are two main ways of getting a training set and they are *a*) creating a new set of data; *b*) labelling a fraction of the actual data

For our purposes, we use the former because we are dealing with natural language and not numbers. People write in different ways on Twitter and trying to create a new training set to encompass all possibilities would be very time consuming and intractable.

I created an application to assist with labelling our train data. It can be found at `http://bit.ly/data_labeller` and a screen shot has been provided in Figure 3.1

3.1.2 Choosing and training a classifier

A Naive Bayes Classifier is a probabilistic classifier which is mainly based on the Bayes Theorem. The classifier works on the assumption that the presence or absence of two features are stochastically independent.

We will train a Naive Bayes Classifier and use it to classify the tweets into relevant and irrelevant groups. **This work is currently in progress**

3.1.3 Classifying tweets

Chapter 4

Topic Modelling

4.1 K-means Clustering

4.2 Latent Semantic Analysis

4.3 Latent Dirichlet Allocation

Chapter 5

Conclusion

5.1 Summary of Report Achievements

Summary.

5.2 Applications

Applications.

5.3 Future Work

Future Work.

Bibliography

[Hayes, 2008] Hayes, Bob E. 2008. *Measuring Customer Satisfaction and Loyalty: Survey Design, use and Statistical analysis Methods*. Third edn. American Society for Quality Press.

[Rubin, 1987] Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.