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.

Contents

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. As we see later in this study, most if not all of these techniques have already been tested on textual data by other researchers but not much has been done on social data.

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[1] 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.

1.3 Statement of Originality

Statement here.

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

Chapter 2

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.

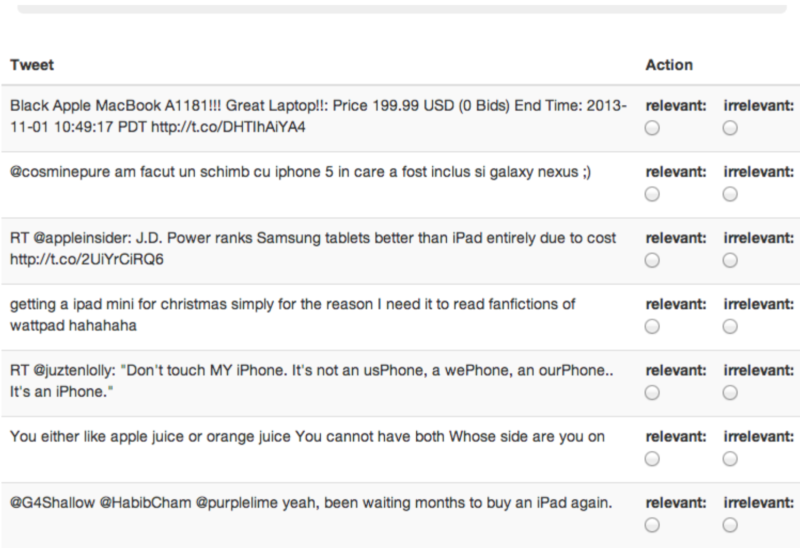
2.1 Data Classification

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!*



Tweet	Action	
Black Apple MacBook A1181!!! Great Laptop!!: Price 199.99 USD (0 Bids) End Time: 2013-11-01 10:49:17 PDT http://t.co/DHTIhAiYA4	relevant:	irrelevant:
@cosminepure am facut un schimb cu iphone 5 in care a fost inclus si galaxy nexus ;)	relevant:	irrelevant:
RT @appleinsider: J.D. Power ranks Samsung tablets better than iPad entirely due to cost http://t.co/2UiYrCiRQ6	relevant:	irrelevant:
getting a ipad mini for christmas simply for the reason I need it to read fanfictions of wattpad hahahaha	relevant:	irrelevant:
RT @juztenlolly: "Don't touch MY iPhone. It's not an usPhone, a wePhone, an ourPhone.. It's an iPhone."	relevant:	irrelevant:
You either like apple juice or orange juice You cannot have both Whose side are you on	relevant:	irrelevant:
@G4Shallow @HabibCham @purplelime yeah, been waiting months to buy an iPad again.	relevant:	irrelevant:

Figure 2.1: 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.

2.1.1 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.

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 2.1

2.1.2 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**

2.1.3 Classifying tweets

Chapter 3

Plan for Semester 2

The table below shows a lists of tasks to be completed and an estimated completion date.

Task	**Completion date**
Evaluate the use of Naive Bayes Classifier and k-means to classify data	Week 2
Cluster tweets by topics using Topic Modelling techniques	Week 5
Run Sentiment Analysis on topics	Week 8
Evaluate techniques used	Week 10

Chapter 4

Risk Assessment

Risk	**Likelihood**	**Impact**	**Measures taken to prevent occurence**
Data loss	Low	High	Initial data is currently backed up to external drives. Labelled data will be backed up when labelling is complete
Algorithm run time	Medium	Medium	For some cases, I can test the programs with a little dataset until I feel it is ready to take on the large dataset
Missing deadlines for the above tasks	Medium	High	I have created a Trello board with all tasks and plan to follow it strictly.