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

Title of my project

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 techinques on social data.

1.2 Why Twitter?

Twitter is a social micro-blogging platforms where users can share their thoughts 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 usualy 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

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

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.

Chapter 2

Data Aggregation

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

2.1 Data Collection

2.2 Data Filtration

Chapter 3

Background Research

3.1 Introduction

Text of the Background.