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

Draft Report

Fayimora Femi-Balogun

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

Abstract

NOT COMPLETE

Every organisation out there today is constantly looking for ways to improve customer satisfaction. *Talk a little bit about 1 or two companies and how they do it (3 sentences...ish)*

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.

Acknowledgements

Dedication

Dedication here.

‘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

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 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. 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. Some of them include 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. (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)

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

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 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 lenient in other not to sound too negative. This could also 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.

1.2 Aims and Objectives

The main aim of this project is to investigate other means of getting our data 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*. We could also infer that the sentiment of the user, with respect to those topics, is somewhat *negative*. The process of discovering abstract topics in text is called **Topic Modelling** while the process of discovering sentiments in text is known as **Sentiment Analysis**. Chapters 4 and 5 discuss how we can automate these processes, respectively.

1.3 Methodology

How do I plan to approach this research?

1.4 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.

1.5 Statement of Originality

Statement here.

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

Chapter 2

Background Theory

2.1 Introduction

Automatic Text Classification or Text Categorization is a rapidly growing field in Machine Learning and Natural Language Processing. This mainly due to the amount of electronic data we currently generate. The main task is to assign one or more classes to a given text document. Applications of text classification include *Email Spam Detection* and *Language Detection*. The former involves trying to distinguish spam emails from legitimate ones while the latter involves the identification of the language a document was written in.

However, this study makes use of classification techniques for data filtration (removing irrelevant documents from a list of documents, similar to spam filtering), topic modelling (extracting topics from a list of documents) and sentiment analysis (predicting the sentiment of the author of a document). This chapter explains a few background concepts and reviews some relevant research previously done in this area.

2.2 Naive Bayes Classifier

The naive Bayes classifier is the simplest classifier that can be used and this is due to the fact that is is based on simple Bayes Theorem. It is a probabilistic classifier which assumes that all features of the documents are independent of each other.

Bayes theorem states that the probability of A given B is the probability of B given A times the probability of A divided by the probability of B. Mathematically, this is written as:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.1}$$

Applying this logic to text classification, the probability that a document $d_i \in D$ belongs to a class c is denoted as:

$$p(c|d_i) = \frac{p(d_i|c)p(c)}{p(d_i)} \tag{2.2}$$

Although other techniques like Maximum Entropy, Random Forests or Support Vector Machines tend to perform better, a naive Bayes classifier will require less memory and CPU cycles. It is also computationally less complex and simpler to implement. With regard to performance, (Huang *et al.* , 2003) showed using multiple datasets from (Blake & Merz, 1998) that the naive Bayes classifier in many cases performs as good as other complex classifiers and (Zhang, 2004) goes further to explain why it performs well. Other studies have also found Bayesian classifiers to be effective without being affected by its simple independence assumption (Langley *et al.* , 1992; Manning *et al.* , 2008).

2.3 Topic Modelling

Topic Modelling is a process by which abstract topics/themes are extracted from a collection of documents. This process is usually carried out with the aid of topic models, a suite of algorithms used for topic modelling. It has been applied in a variety of fields like Software Analysis where (Linstead *et al.* , 2009) used topic modelling to find topics embedded in code and (Gethers & Poshyvanyk, 2010) used topic modelling to capture coupling among classes. (Kireyev *et al.* , 2009) applied topic models on disaster related data from Twitter in an effort to determine what topics were discussed within the time span of a natural disaster. (Hospedales *et al.* , 2009) introduced a new topic model that can be used to analyze videos with complex and crowded scenes in other to discover regularities in the videos. A system built on such model will be able to answer a question like "What interesting events happened in the last 5 hours". Other fields include Audio Analysis (Smaragdis *et al.* , 2009), Influence modelling (Gerrish & Blei, 2009), Finance (Doyle & Elkan, 2009), Writer Identification (Bhardwaj *et al.* , 2009) and many more.

There are a number of topic models but the two main ones are ***Latent Semantic Indexing*** (LSI) and ***Latent Dirichlet Allocation*** (LDA) and we discuss them further in the following sections.

2.3.1 Latent Semantic Indexing

Latent Semantic Indexing, sometimes referred to as *Latent Semantic Analysis*, is an indexing technique leverages matrix-algebra computations[1] to identify any patterns in relationships between a collection of text documents. It works based on the assumption that words used in the same context tend to have homogeneous meanings (Deerwester *et al.* , 1990; Dumais, 2004; Landauer, 2006).

2.3.2 Latent Dirichlet allocation

lda is blah blh blah

2.4 Model Evaluation

[1]Specifically, it uses Singular Value Decomposition which is a factorization of a complex matrix. See `http://en.wikipedia.org/wiki/Singular_value_decomposition`

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!*
- *Yes, our apple juice and cider are both GMO-free.*
- *I just had my first carmel apple*

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

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 3.1: The data labelling application

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

Sentiment Analysis

Building my own sentiment classifier would be awesome but if time does not permit, I might just make use of an available API. Maybe CHatterbox or anything similar.

Chapter 6

Conclusion

6.1 Summary of Report Achievements

Summary.

6.2 Applications

Applications.

6.3 Future Work

Future Work.

Appendices

Appendix A

Sample Appendix

The content of the appendix

Bibliography

Bhardwaj, Anurag, Malgireddy, Manavender, Setlur, Srirangaraj, Govindaraju, Venu, & Ramachandrula, S. 2009. Writer identification in offline handwriting using topic models. *In: Proceedings of the NIPS 2009 Workshop on Applications of Topic Models: Text and Beyond*.

Blake, Catherine L, & Merz, Christopher J. 1998. UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California. *Department of Information and Computer Science*, **460**.

Deerwester, Scott C., Dumais, Susan T, Landauer, Thomas K., Furnas, George W., & Harshman, Richard A. 1990. Indexing by latent semantic analysis. *JASIS*, **41**(6), 391–407.

Doyle, Gabriel, & Elkan, Charles. 2009. Financial topic models. *In: NIPS 2009 Workshop on Applications of Topic Models: Text and Beyond*.

Dumais, Susan T. 2004. Latent semantic analysis. *Annual review of information science and technology*, **38**(1), 188–230.

Gerrish, Sean, & Blei, David. 2009. Modeling Influence in Text Corpora.

Gethers, Malcom, & Poshyvanyk, Denys. 2010. Using relational topic models to capture coupling among classes in object-oriented software systems. *Pages 1–10 of: Software Maintenance (ICSM), 2010 IEEE International Conference on*. IEEE.

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

Hospedales, Timothy, Gong, Shaogang, & Xiang, Tao. 2009. A markov clustering topic model for mining behaviour in video. *Pages 1165–1172 of: Computer Vision, 2009 IEEE 12th International Conference on.* IEEE.

Huang, Jin, Lu, Jingjing, & Ling, Charles X. 2003. Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. *Pages 553–556 of: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on.* IEEE.

Kireyev, Kirill, Palen, Leysia, & Anderson, K. 2009. Applications of topics models to analysis of disaster-related twitter data. *In: NIPS Workshop on Applications for Topic Models: Text and Beyond*, vol. 1.

Landauer, Thomas K. 2006. Latent semantic analysis. *Encyclopedia of Cognitive Science.*

Langley, Pat, Iba, Wayne, & Thompson, Kevin. 1992. An analysis of Bayesian classifiers. *Pages 223–228 of: AAAI*, vol. 90.

Linstead, Erik, Hughes, Lindsey, Lopes, Cristina, & Baldi, Pierre. 2009. Software analysis with unsupervised topic models. *Page 52 of: NIPS Workshop on Application of Topic Models: Text and Beyond*, vol. 50.

Manning, Christopher D, Raghavan, Prabhakar, & Schütze, Hinrich. 2008. *Introduction to information retrieval.* Vol. 1. Cambridge University Press Cambridge.

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

Smaragdis, Paris, Shashanka, Madhusudana, & Raj, Bhiksha. 2009. Topic Models for Audio Mixture Analysis. *Applications for Topic Models: Text and Beyond, Whistler.*

Zhang, Harry. 2004. The optimality of naive Bayes. *A A*, **1**(2), 3.