Telco Feedback Management System using Sentiment Analysis

Suhail Jamaldeen

149212 K

MSc in IT

University of Moratuwa

March 2018

Supervised by

Mr. Saminda Premaratne

Senior Lecturer

University of Moratuwa

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Suhail Jamaldeen

Date:

Supervised by

Mr. Saminda Premaratne

Date:

Dedication

This thesis is dedicated my father

Late Mr. HL. Jamaldeen
(Senior Superintendent of Police)

who sacrificed his life to our mother nation Sri Lanka

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Mr. Saminda Premarathna, for the continuous support of my MSc study and research, for his patience, motivation, enthusiasm, and immense knowledge. I have been amazingly fortunate to have a supervisor who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps faltered.

Besides my supervisor, I would also like to thank my family specially my mother, my wife and daughter for the support they provided me through my entire life and in particular.

I must also acknowledge my friends Chinthuja Varathalingam, Saranya Varathalingam and Arunnirthana Arulanantham for their support throughout my thesis submission.

Abstract

Social Media such as Facebook, Twitter, blogs, etc. and social interactions are currently growing in an exploding speed. This leads Social Media in becoming a large repository of valuable opinions of a large scale of different people on numerous products and services. Twitter and Facebook plays a major contribution in social interactions. As such, people share their opinion on Telecommunication service providers using social media.

Usually any telecommunication provider in the world provide more than one services, such as Home Broadband, Mobile Devices, Prepaid, Postpaid, Mobile Broadband and Televisions.

When people share their opinions and feedbacks on the services which given by any telecommunication providers through social media. Telecommunication providers face a huge challenge in

- Categorizing the opinions as positive feedback, negative feedback, questions and statements
- Identifying for what services, the opinion is belong to
- Assigning the categorized opinion to a support engineer depending on the services which are provided by the telecommunication provider
- Replying to opinions, following them and managing them

When it’s come to categorizing the opinion, we need a best approach to analyze the text and the data analysis approach should be able to detect only explicitly expressed opinions.

And then when it’s come to assigning categorized opinion to a support engineer; it cannot be done from a social media itself. None of the social media allows us to use their platform to manage the individual’s opinion.

Same as assigning; replying and managing individual’s opinion cannot be done using any of the social media.

This research study will propose a best approach to analyze the text, which are mentioned a telecommunication provider and this study will provide a mechanism or a software solution to manage the opinion raised by the subscriber of a telecommunication.

In this thesis project, we focus on using Twitter, one of the most popular social media.

Table of Contents

List of Figures

List of Table

Chapter 1

Introduction

1.1 Prolegomena

Social Media such as Twitter, Facebook and blogs has become huge repository of valuable user generated data. These data are generated by different kind of people on number of different products and services all over the world. Hence reading these text content source containing opinions is impossible. These texts could contain numbers of attributes (or features); forming complex opinion relationships hidden inside sentences. The complexity of identifying valuable opinions from these unstructured texts leads the problem to use data mining techniques. Data mining exposes patterns and relationships using intelligence techniques, which might otherwise have remained undetected. (Sumathi & Sivanandam, 2006)

Opinion mining is a sub type of web content mining, which is defined as application of data mining techniques to discover patterns and extracting useful information from online web content (Liu, 2011). Opinion mining is also referred as sentiment analysis. In our case, the content mining source will be social media. Opinion mining comprises Information Retrieval, Information Extraction and Data Mining tasks to address the problem with Text Mining, Natural Language Processing (NLP) and Text Classification applications. (Kalaria & Prajapati, 2016)

1.2 Background and Motivation

During this digital era, amount of available information grows dramatically; making the information retrieval and availability of best sources for a user with least effort is a canonical problem. In recent past, special attention has been given to the valuable User Generated Content (UGC) available on the World Wide Web, particularly in social media. Various studies have been under gone over recent years on capturing customer opinions from this UGC, since it represents a valuable unique type of information, which symbolizes the voice of ordinary customers.

Social media is one of the areas of the WWW, which contains highly active users, and highly dynamic UGC where individuals and communities share, create, discuss, and modify. This content contains of opinions on various subjects, products and services, from different people having different point of views.

Telecommunication service provider is a type of communication service provider who providers telephony services and related services. Nowadays Telecommunication providers provide a vast number of services from GSM, Internet, Television, wireless communication and others.

According to a Central Bank report, every 100 people in Sri Lanka owns 113 mobile phones and 13 fixed lines by the end of the year 2015 (Anon., 2016). Telecommunication providers provide 24 X 7 customer support for their subscribers via voice call, online chatting facility, email and social media.

Customer support through Voice call, Chatting and email can be tracked, maintained and archived. The communication through these channels can be maintained in a secure manner where public does not have accessed to.

However, social media is very different from the other support systems. Social media is open to any individual; others can access one subscriber's feedback and comments.

Moreover since social media feedback and comments can be a decision point of common users, so the telecommunication providers should be able to track them in an effective manner as well.

1.3 Problem Statement

Social media becomes the largest user generated content repository when individuals rise complains and give feedbacks through social media to their Telco providers. However, Telco providers are facing difficulties in mining the data that related to them and their services, categorizing and assigning them to their support engineers and following them up.

In addition, the Telco providers has no evidence and relation mapping to their services and external factors, which are interrupting their services. For an example there is no evidence of relation that if there is a rain the 4G speed drops down.

A social media post which related to a telecommunication provider can be categorized as "Negative Feedback, Positive Feedback, Question and Statement" and depending on the service category (Internet, Broadband, 4G and others) they can be classified. For an example a social media post “I was able to browse the home broadband internet without any disruption” can be categorized as “**Positive feedback** related to **Home Broadband**”

Currently the Telco providers manage the issues raised by the subscribers through social media using excel sheets, which are difficult to manage and maintain.

1.4 Hypothesis

The user-generated content in social media, which are related to telecommunication providers can be mined, categorized according to the feedback type and maintained using a software solution. To achieve this, a data mining and language processing approach should be identified and a software solution should be built in order to manage mined data.

1.5 Objectives

The aim of this project is to find a novel approach to analyze the tweets, which are mentioned to a telecommunication provider and categories them as positive feedback, negative feedback, questions and statements related to the services provided by the particular telecommunication provider.

In addition, this study will provide a mechanism or a software system to manage the opinion raised by the subscriber of a telecommunication.

Basically the main objectives of this study can be outlined as follows.

- Identify a mechanism to crawl the tweets which are related the particular telecommunication provider.
- Store the crawled tweets into the database.
- Identify a better natural language processing mechanism and sentimental analysis methodology and apply on the crawled tweets so that they can be categorized into positive feedback, negative feedback, questions and statements.
- Identify a methodology to evaluate for what service category a tweet is belongs to and assign support engineers depending on the service category.
- The support engineer should be able to comment on the tweets.

- The commented replies should be able to seen by the subscriber
- The tweets, reply and status should be maintained by the portal and should be tracked.

1.5 Proposed solution

Using a crawling technology the system will crawl the tweets tweeted by mentioning a telecommunication provider.

Then the software solution will identify the crawled tweets using an appropriate sentiment analysis technique and categorize them as “Negative Feedback, Positive Feedback, Question and Statement” and also the system will identify to what service is the tweet belong to, and will assign support engineers who are related to specific services (Peo TV, GSM, Broadband, Mobile devices and so on). For this system a natural language processing mechanism and a proper sentiment analysis mechanism will be implemented.

Then the support engineers will act according to the tweets and the tweets will be tracked and evaluate.

1.6 Scope of the thesis project

In this research project, the tweets which are in English language only will be crawled and evaluated.

Usually user generated content in social media contain many grammatical mistakes and miss-spelled words, incomplete words and playful words known as internet slang (Boiy et al., June 2007). Tweets which are to be analyzed are assumed to be syntactically and grammatically correct.

The tweets which are to be analyzed will be based on the subjective detection for a telecommunication provider and their services.

Input dataset used for the training are crawled from twitter which mention a particular telecommunication provider.

All the tweets are automatically labeled as “Negative Feedback, Positive Feedback, Question or Statement” using the appropriate sentiment analysis technique.

Also the appropriate tweets will be categorized automatically depending on the service category, and the support engineer for the category can be assigned.

The comment and reply will be posted automatically to the tweet as a reply in the tweet.

Tweets and feedback with images will not be considered and crawled since it falls under image processing.

1.6 Structure of the Thesis

The rest of the thesis is organized as follows.

Chapter 2, Literature Review and Background Study will give an in depth knowledge and understanding on sentiment analysis and how the sentiment analysis technologies are used for text analysis.

This chapter will further discuss on current techniques in sentiment analysis, their strengths and weaknesses as well.

This chapter will discuss on social media, particularly about twitter, its features and limitations. In addition, this chapter will illustrate how the telecommunication providers use sentiment analysis and manage the feedback given by the subscribers currently. This chapter will also give an idea on problem definition.

Next, the Chapter 3, Propose Approach and Technologies Adoption. This chapter describes proposed approach to the defined problem and technologies are used. Authentication and authorization on twitter API, what are the inputs, how the inputs are crawled, how natural language input are processed, algorithms used for the training are further explained in the chapter.

The Chapter 4; Design and Implementation. This chapter describes on software solution architecture. How the proposed technologies are used and in what layer they are used also will be discussed in this chapter. This chapter also discusses the how the technologies are implemented in the system. Software Development methodology used, how the data flows work, UI and UX capabilities will be discussed in this chapter.

Chapter 6 is on evaluation of the new solution. The expected output and the results will discussed in this chapter.

Chapter 7 concludes the research with a note on further work. This will describe the future enhancements and how the limitations can be avoided and make the solution more suitable for the specified domain.

Chapter 2

Literature Review and Background Study

2.1 Introduction

In chapter 1, I have tried to present you an overall introduction to this research project.

This chapter will present in-depth knowledge on sentiment analysis, current techniques and technologies used in sentiment analysis, their strengths and weaknesses. This chapter also describes how the sentiment analysis are used for text based content. This chapter will also discuss on how the sentiment analysis used with the content from social media.

At the end of the chapter problem definition also described.

2.2 Sentiment and Sentiment Analysis

This section discusses on what is sentiment, what is sentiment analysis, current methods and techniques of sentiment analysis, strengths and weaknesses of their usage and their various applications.

2.2.1 Sentiment

Sentiment can be interpreted as feelings, which determines if an expression is positive, negative, or neutral, and to what degree (Liu, 2011). “What other people think” has always been an important piece of information for most of us during the decision-making process (Prager, 2006).

2.2.2 Sentiment Analysis

Sentiment analysis is a field of study of identifying and analyzing individuals’ opinions (negative, positive), attitudes, emotions, and evaluations from text, speech, tweets and database sources through Natural Language Processing (NLP) (Wilson, et al., 2005). It's also referred as subjectivity analysis, opinion mining, and appraisal extraction.

The words opinion, sentiment, view and belief are used interchangeably but there are differences between them.

- Opinion: A conclusion open to dispute (because different experts have different opinions)
- View: subjective opinion
- Belief: deliberate acceptance and intellectual assent
- Sentiment: opinion representing one's feelings

(Kharde & Sonawane, 2016)

Sentiment Analysis is broadly classified in the two types.

1. Feature or aspect based sentiment analysis.

   Mining opinions from text about specific entities and their aspects.

2. Subjectivity or Objectivity based sentiment analysis.

   Objectivity based sentiment analysis does the exploration of the tweets which are related to the emotions like hate, miss, love etc.

Sentiment Analysis is one of the most trending research areas in natural language processing. Sentiment Analysis and is also widely studied in data mining, Web mining, and text mining. (Liu, 2012)

Opinion Mining and Sentiment Analysis systems are being developed and used in almost every business and social domain.

2.2.3 Early Study on Sentiment Analysis

Document-Level Sentiment Analysis

The study on sentiment analysis was started with document level sentiment analysis. Document-level sentiment analysis is known as the simplest form of sentiment analysis. Labelling the entire document as containing overall positive or negative polarity, or rating scores of reviews is called document level sentiment analysis. These systems mainly based on supervised approaches relying on manually labeled samples.

Document level sentiment analysis starts with assumption that document contains opinion about one entity. For example a book or a document can be reviewed by an individual can show his/her positive or negative attitude toward that book. However, sentiments were not only expressed at document-level, nor they are limited to a single target.

There are two key approaches to document-level sentiment analysis, supervised learning and unsupervised learning.

Document Level Sentiment Analysis Using Supervised Learning

Sentiment analysis can be formulated as a supervised learning with positive, negative and neutral classes. Supervised learning uses a known dataset (called the training dataset) to make predictions. The training dataset will have input data and response values. Using the training dataset the supervised learning algorithm will build a model which can make predictions and will assign the response values for the new dataset.

Typically document-level supervised learning methods are widely used in sentiment analysis of product reviews. Any existing supervised learning methods can be applied to sentiment analysis, e.g., Naïve Bayesian algorithm, Maximum Entropy and Support Vector Machine, .etc. (Pang et al., 2002) used supervised learning to classify movie reviews into two classes, positive and negative. They have proved that using a bag-of-word unigram as features can upgrade the performance of both Naïve Bayesian and Support Vector Machine classifications.

Support Vector Machine method has been shown to be to be higher accurate in the study by Pang et al. (2002).

Supervised learning process contains two main steps.

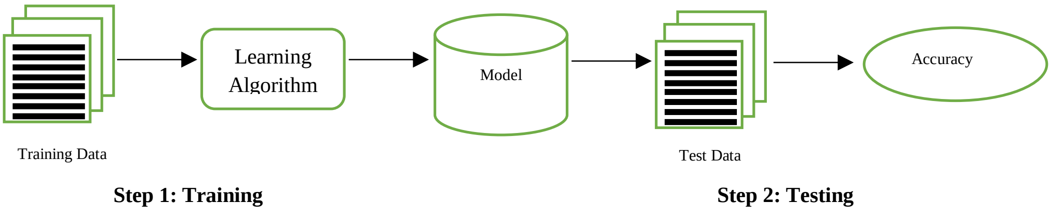


Figure 1 - Document Level Sentiment Analysis using Supervised Learning

1. Learning (Training)

   This methodology will learn a model using the training sample data.

2. Testing

Test the model using unseen test data to assess the model accuracy.

$$Accuracy = \frac{Number\ of\ correct\ classification}{Total\ number\ of\ test\ cases}$$

Document Level Sentiment Analysis Using Unsupervised Learning

Unsupervised approach is based on finding Semantic Orientation (SO) of a particular phrase within the document. If the average value of these SO is higher than a pre-defined threshold value, the document is classified as positive, else negative. Turney in his proceedings “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews” proposed a ‘thumbs up’ or ‘thumbs down’ classification for product reviews using unsupervised learning method.

To estimate the SO of phrases Turney used one of the classical techniques, the difference between Pointwise Mutual Information (PMI) values of the phrase and two sentiment words. (Turney, 2002)

$$PMI\ (word_1, word_2) = \log_2 \frac{P\ (word_1 \& word_2)}{P\ (word_1)\ P\ (word_2)}$$

PMI values are calculated by comparing its similarity to a positive reference word “excellent” with its similarity to a negative reference word “poor”. These words were chosen because, in a common five star review rating system, one star rating is called “poor” while five stars called “excellent”.

$$SO\ (Phrase) = PMI(Pharse, "excellent")\ -\ PMI\ (Pharse, "poor")$$

The classification was done with three steps; (1) extract phrases containing adjectives or adverbs, (2) estimate the SO of each phrase and (3) classify the review considering the average semantic orientation value of the phrases. Results showed that average of 74% accuracy can be achieved with this method. Turney suggests that automatic ‘thumbs up’ or ‘thumbs down’ classification can be applied to search engines. For example, given the query “Akumal travel review”, a search engine could report, “There are 5,000 hits, of which 80% are thumbs up and 20% are thumbs down”

(Turney, 2002). The search results could be sorted by average SO, so that the user could easily filter out what are most criticized reviews and what are most supportive.

Segment Level Sentiment Analysis

With the mature of document level sentiment analysis, segment level sentimental analysis arose, where the sentimental segments will be recognized from non-sentimental segments. e.g., graph-based techniques and identification of syntactic phrases. In graph-based technique the document will be segmented based on their subjectivity, or by performing a classification based on some fixed syntactic phrases that are likely to be used to express opinions. When it's come to identification of syntactic phrases method n some fixed syntactic phrases that are likely to be used to express opinions.

Sentence Level Sentiment Analysis

After the maturity of the segment based sentiment analysis, the sentiment analysis started based on sentence level. (Cambria, 2013). This is more useful than the document and segment level analysis because a single document and a segment of text can contain several opinion for the same entity. To get a deeper resolution of the opinions it carries out, sentence level sentiment analysis become helpful. (Ortony, et al., 1990)

Feature based sentiment analysis

Feature based sentiment analysis is used for situations where people express opinions about multiple features (or attributes) of an entity within same sentence, like battery life or display size of a camera. In many cases users do not directly express their opinion about one product, but provide comparable opinions on two products and this problem is addresses in comparative sentiment analysis. Sentiment Lexicon Expansion is the technique used to generate list of positive and negative opinion-baring words, and is further discussed in this section. Further it explores the literature on available paths, especially in concept-level sentiment analysis which aims to go beyond a mere word level analysis of text and provide a more semantic analysis of text through the use of web ontologies or semantic networks (Becker et al, 2013).

Word Level Sentiment Analysis

Word level sentiment analysis is based on a single adjective as feature. The two methods of automatically annotating sentiment at the word level are:

1. Dictionary-Based Approaches
2. Corpus-Based Approaches.

2.2.4 Main Approaches to Sentiment Analysis

According to Cambria in his journal “An Introduction to Concept-Level Sentiment Analysis” says that existing approaches to sentiment analysis can be grouped into three main categories: 1. Keyword Spotting, 2. Lexical Affinity, and 3. Statistical Methods.

Keyword Spotting

Keyword spotting is known as the most basic but popular approach of all three. In this approach text is classified into categories based on clear-cut words which will be the decision points; the words can be ‘happy’, ‘sad’, ‘afraid’, and ‘bored’.

This method is well suited when a sentence convey a direct meaning like “The trip was fun”. This approach is likely to fail on a sentence which doesn’t contain a direct meaning like “This trip wasn’t much a fun at all”. Another weakness of this methodology is that most of the sentences convey emotions through underlying meaning rather than affect adjectives. For example, the text “My husband just filed for divorce and he wants to take custody of my children away from me” certainly evokes strong emotions, but uses no affect keywords, and therefore, cannot be classified using a keyword spotting approach. (Cambria, 2013)

Lexical Affinity

Lexical affinity methodology is slightly more sophisticated than keyword spotting. In this method rather than detecting affect and adjective keywords, assigns probability ‘affinity’ for a particular emotion. For example, the key word ‘war’ can be assigned 80% probability of being indicating a negative sentiment, as in ‘civil war’ or ‘damaged by war. These probabilities are usually trained from linguistic corpora.

According to Cambria, lexical affinity approach has two main problems. First problem lexical affinity can be easily tricked by sentences like “I avoided an accident” (negation) and “I met my girlfriend by accident” (other word senses). This is because lexical affinity operating solely on the word-level categorization. Second, lexical affinity probabilities are often biased toward text of a particular genre, dictated by the source of the linguistic corpora. This makes it difficult to develop a reusable, domain-independent model. (Cambria, 2013)

Statistical Methods

Statistical methods, such as Bayesian inference and Support Vector Machines (SVM), have been popular for affect classification of texts in many researches (Pang, et al., 2002). In this methodology a large number of training corpus of affectively annotated texts fed into the machine learning algorithm. The machine learning algorithm does not only process affect keywords, but also take the valence of other arbitrary keywords (like lexical affinity), punctuation, and word co-occurrence frequencies into account.

This approach requires a large training corpus to be able to predict the effect of arbitrary words and would be only suited to classifying larger blocks of text, not just sentences.

Cambria (2002) in his journal “An Introduction to Concept-Level Sentiment Analysis” says that traditional statistical sentiment analysis methods are generally semantically weak and statistical text classifiers only work with acceptable accuracy when given a sufficiently large text input.

So, statistical methods is best suited to classify user’s text on the page- or paragraph -level, they do not work well on smaller text units such as sentences or clauses.

2.2.5 Sentiment Classification Techniques/Algorithms

Naïve Bayes Classifier

Naïve Bayes classifier is based on Conditional Probability and Bayes Rule. It’s a baseline classification algorithm. Naïve Bayes classifier assumes that the classes for classification are independent. (Gupte et al., 2014)

Conditional Probability: Assuming the probability of something to happened based on data of something that has been already happened.

Naïve Bayes classification compares the contents with the list of words to classify the documents to their right category or class. Let d be the tweet and c* be a class that is assigned to d, where

$$C^{*} = \arg mac_{c} P_{NB}(c \mid d)$$

$$P_{NB}(c \mid d) = \frac{(P(c))\sum_{i=1}^{m} p(f \mid |c)^{n_{i(d)}}}{P(d)}$$

From the above equation, ‘f’ is a ‘feature’, count of feature (fi) is denoted with $n_i(d)$ and is present in d which represents a tweet. Here, m denotes no. of features. Parameters P(c) and P(f|c) are computed through maximum likelihood estimates, and smoothing is utilized for unseen features. To train and classify using Naïve Bayes Machine Learning technique, we can use the Python NLTK library. (Kharde & Sonawane, 2016)

Maximum Entropy Classifier

In Maximum Entropy Classifier, no assumptions are taken regarding the relationship in between the features extracted from dataset. This classifier always tries to maximize the entropy of the system by estimating the conditional distribution of the class label.

Maximum entropy even handles overlap feature and is same as logistic regression method which finds the distribution over classes. The conditional distribution is defined as MaxEnt makes no independence assumptions for its features, unlike Naive Bayes.

The model is represented by the following:

$$P_{ME}(c \mid d, \lambda) = \frac{\exp[\sum_{i} \lambda_i f_i(c,d)]}{\sum_{c} \exp[\sum_{i} \lambda_i f_i(c,d)]}$$

Where c is the class, d is the tweet and λ_i is the weight vector. The weight vectors decide the importance of a feature in classification.

Support Vector Machine

Support vector machines are supervised learning model which analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space.

The input data are two sets of vectors of size m each. Then every data which represented as a vector is classified into a class. Nextly we find a margin between the two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SUPPORT VECTOR MACHINE also supports classification and regression which are useful for statistical learning theory and it also helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully.

Bayesian Network

Bayesian network is a type of Probabilistic Graphical Model that can be used to build models from data and/or expert opinion. They can be used for a wide range of tasks including prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction and decision making under uncertainty. (Anon., 2018)

Gradient Boosted Regression Trees

The Gradient Boosted Regression Trees (GBRT) model (also called Gradient Boosted Machine or GBM), is one of the most effective machine learning models for predictive analytics.

Neural Network

Neural Network has emerged as an important tool for classification for the past few years and it has established as a promising alternative to various conventional classification methods.

2.2.6 Evaluation of Sentiment Classification

The performance and the accuracy of sentiment classification can be evaluated by the below mentioned four indexes and equations.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{F1} = (2\ \text{X Precision X Recall})/(\text{Precision} + \text{Recall})$$

In the above TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances, as defined in the below table.

Table 1 - Confusion Matrix

	Predicted Positives	Predicted Negatives
Actual Positive	TP	FN
Actual Negative	FP	TN

2.3 Machine Learning

Machine learning is a way of using artificial intelligence (AI) that provides systems the capability to automatically learn and train themselves and improve from experience without being explicitly programmed.

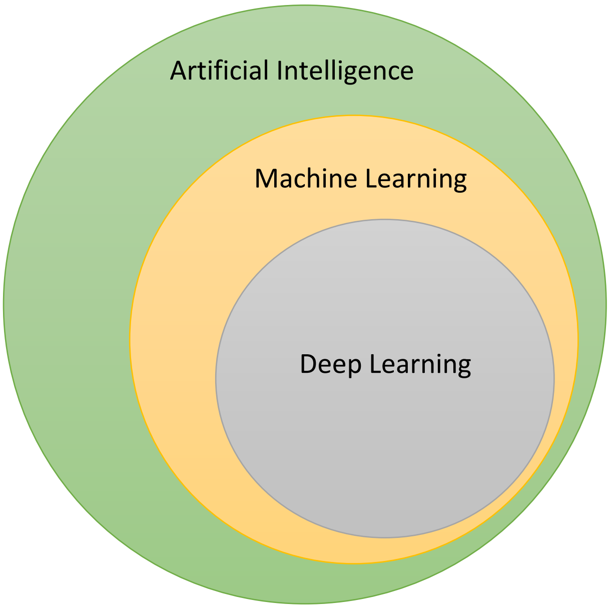


Figure 2 - Artificial Intelligence vs Machine Learning vs Deep Learning

Machine learning systems finds patterns in explicit provided large amount of data classify them into classes and get trained.

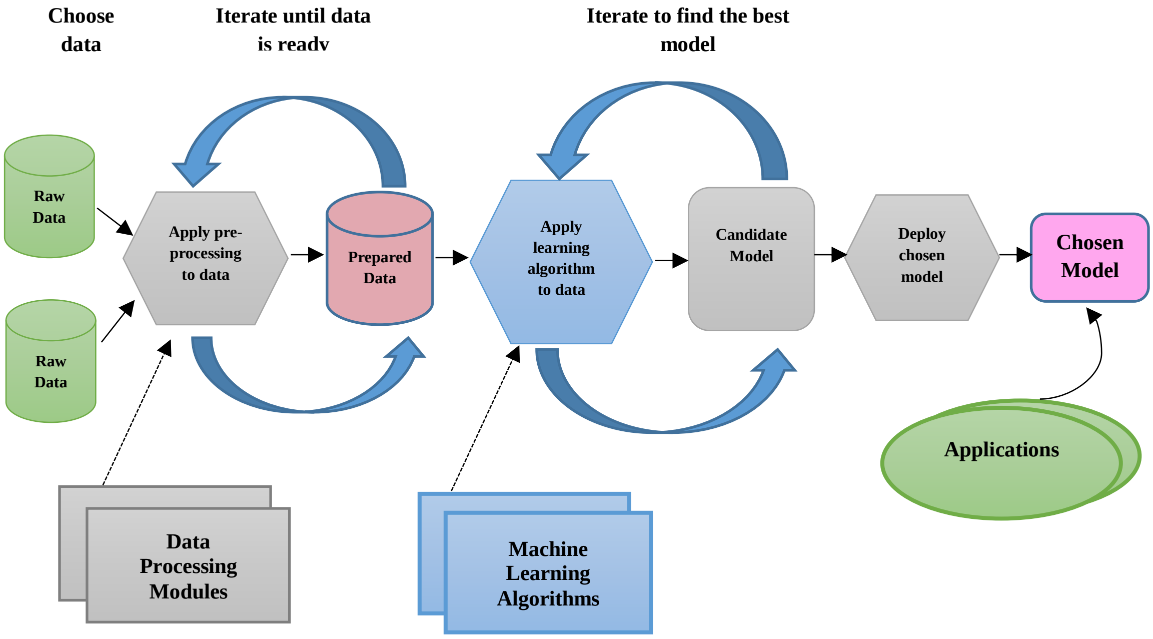

Choose data
Iterate until data is readv
Iterate to find the best model
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model
Applications
Data Processing Modules
Machine Learning Algorithms

Figure 3 - Machine Learning Process

For a better machine learning approach we need three things:

1. Lots of data to train
2. Lots of compute power
3. Effective machine learning algorithms.

As discussed in the section – document level sentiment analysis this approach contains two types of machine learning techniques:

1. Unsupervised learning
2. Supervised learning.

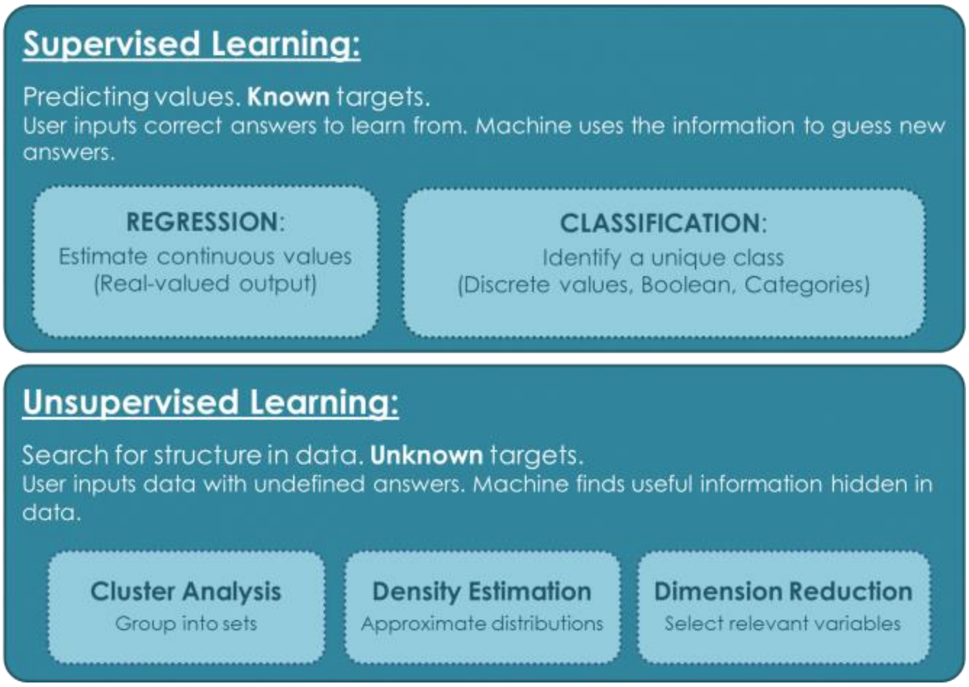

Supervised Learning:
Predicting values. Known targets.
User inputs correct answers to learn from. Machine uses the information to guess new answers.
REGRESSION:
Estimate continuous values
(Real-valued output)
CLASSIFICATION:
Identify a unique class
(Discrete values, Boolean, Categories)
Unsupervised Learning:
Search for structure in data. Unknown targets.
User inputs data with undefined answers. Machine finds useful information hidden in data.
Cluster Analysis
Group into sets
Density Estimation
Approximate distributions
Dimension Reduction
Select relevant variables

Figure 4 - Supervised Learning and Unsupervised Learning

The success and accuracy of the both learning methods are mainly depends on the selection and extraction of the specific set of features used to detect sentiment. (Kharde & Sonawane, 2016)

Machine learning for text sentiment analysis need two sets of data.

1. Training Data Set

   The prepared data used to create a model.
2. Test Data Set

A test dataset is often used to validate the model.

Machine learning starts with the collection of dataset for training. Then using the training data set the classifier should be trained. Then the selection of feature should be taken place which is very important this will tell how the documents are represented.

The most commonly used features in sentiment classification are

- Term presence and their frequency
- Part of speech information
- Negations
- Opinion words and phrases

(Kharde & Sonawane, 2016)

2.3.1 Machine Learning Offerings

SAS Analytics Suite

SAS analytics suite is a software developed by SAS Institute during 1960s for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics.

SAS is a software suite that can mine, alter, manage and retrieve data from a variety of sources and perform statistical analysis on it. (Anon., 2018)

RapidMiner Studio

RapidMiner is a data science software platform which enables to build software application for data science teams that unites data prep, machine learning, and predictive model deployment.

RepidMiner provides environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. (RapidMiner, 2018)

Alteryx Analytics

Alteryx was founded in 1997 and Alteryx Analytics is one of the products from Alteryx. Alteryx Analytics is preferred solution for big data analysis.

IBM SPSS

SPSS is one of the oldest ad widely used data analytics software which used for logical batched and non-batched statistical analysis. IBM acquired SPSS in 2009.

Now IBS SPSS platform offers advanced statistical analysis, a vast library of machine learning algorithms, text analysis, open source extensibility, integration with big data and seamless deployment into applications. (IBM, 2018)

Microsoft Azure Machine Learning

Microsoft Azure Machine Learning is a cloud based vendor by Microsoft. This was started with the development of Cortana which was the intelligence assistant for windows and some other products. (Anon., 2018)

Azure ML provides Machine Learning Studio which is a collaborative drag and drop tool to build, test, and deploy predictive analytics solutions using the data we feed into the studio. Also machine learning studio enable to publish the models as web services where the web services can easily be consumed by custom developed apps or BI tools such as Power BI and Excel. Azure Machine Learning Studio contains a wide scale of built in library of algorithms from regression, clustering, classification, and anomaly detection families. We can make use of them as needed depending on the different type of machine learning problem we are going to address. (Ericson & Rohm, 2017)

Another Microsoft Azure feature is “Machine learning algorithm cheat sheet for Microsoft Azure Machine Learning Studio”. This cheat sheet helps the developers to choose the correct algorithm for a predictive analytics model.

Amazon Machine Learning

Amazon been investing deeply in artificial intelligence for over 20 years and came up with a cloud based solution. Amazon Machine Learning provides visualization tools and wizards that guide you through the process of creating machine learning (ML) models without having to learn complex ML algorithms and technology.

(Amazon, 2018)

Microsoft Language Understanding Intelligent Services (LUIS)

LUIS is a service provided by Microsoft as a part of Microsoft Cognitive Services, which is formally known as Project Oxford. LUIS is designed to identify valuable information in conversations, Language Understanding interprets user goals (intents) and distills valuable information from sentences (entities), for a high quality, nuanced language model. LUIS uses machine learning and allow developers to build application which use natural language as input and extract the meaning from it (Microsoft, 2017).

LUIS enables the developers to create their own language understanding models in an easy way. LUIS offers a set of language understanding tools to achieve it.

2.4 Future challenges of Sentiment Analysis

There are many works has been done in sentiment analysis over the past years and the studies are being done rapidly in new directions with the development of technologies and techniques. Although there are many automatic classifiers, features, and datasets used to detect sentiment; there are some of the challenges still remaining; questions that have not been explored sufficiently.

In this section we will be discussing some of the challenges which are not addressed in sentiment analysis domain.

- Language: Most of the sentiment analysis studies has been done for English, French, Spanish and some other languages. But still there is a huge space for other languages like Sinhala, Tamil, Japanese and many other. So studies can be done in these language and fill the gap.

- Automatic Reply and Chatbot: There are chatbot created using natural language understanding and artificial intelligence. A chatbot is a computer application that attempts to simulate the conversation of a human being via text or voice interactions. A human being can ask a chatbot a question or can make a command, then the chatbot will respond and perform the request. Using natural language understanding, systems and chatbots like these can be developed with highly accurate.

2.5 Twitter

Twitter is a free social networking site, which can signed up and used by anyone around the world. Twitter attracts over 330 million monthly active users and generates over 500 million Tweets daily. (Aslam, 2018)

Users can interact with any other users with direct messages or posts, which called as "tweets", restricted to 280 characters. Posts / tweets can be read by anyone depending on the privacy setting of the twitter user. Users can access and user Twitter through browser, SMS or a mobile device.

There are several conventions used by micro bloggers to convey information within the limit of 2 characters.

- Hashtags (#) followed by a word or code e.g. #icwsm, are used to group related posts together.
- Posts may be directed to a particular person by putting a @username at the beginning of the post. Even though the post is directed to a person others can still view it, provided the account is public.
- Micro-bloggers can 'retweet' someone else's post by copying the post and the person's username.
- Micro-bloggers often add URLs to a post. To keep within the character limit, they use a URL shortening service.

(Ehrlich & Shami, 2010)

2.5.1 Twitter API and Libraries

Twitter provides endpoints through Twitter API. Twitter endpoints and API are bundled into many libraries and programming languages such as C#, Java, VB.Net, Ruby, Python, C++, Object C and many other.

There are two API types provided by Twitter in order to communicate.

1. Rest API

   Twitter Rest API allows to query and search Twitter feeds and the Twitter users' data (which are public and which we have the authorization to access). Rest API work as request and response way. The client (our application) will send request to server (Twitter) and the server reply back to client as Response.

2. Streaming API

   Streaming API, push the tweets based on search terms or for specific users you request for a listener; where the listener will process the data which are captured. Steaming API needed continuous net connection which makes the push delivered in real time.

(Twitter, 2017)

2.5.2 Authentication and Authorization on Twitter API

The Twitter API allows a consumer application to use an OAuth Request Token to request user authorization where the consumer can use Twitter service impersonating as a Twitter user.

There are two forms of authentication.

- User authentication

  This is the most common form of resource authentication in Twitter's OAuth 1.0a implementation.

- Application-only authentication

  Application-only authentication is a form of authentication where an application makes API requests on its own behalf, without a user context. API calls are still rate limited per API method, but the pool each method draws from belongs to the entire application at large, rather than from a per-user limit

(Twitter, 2017)

The consumer application will have to be created as an app from Twitter developer dashboard (https://apps.twitter.com/) where we will receive Consumer Key (API Key) and other tokens and those keys can be used in the application as impersonation of a user.

2.6 Related Works

In this section we will be discussing and evaluating the past research and software systems which are related to sentiment analysis for twitter and other social media.

2.6.1 Amolik et al., (2016)

In a research project by Amolik et al., (2016); they used Twitter feeds to analyse the movie reviews. They prefer machine learning technique over the lexian technique. They used feature vector and classifiers such as Support vector machine and Naïve Bayes as the machine learning classifiers. They classifies tweets as positive, negative and neutral to give sentiment of each tweet.

The results show that we get 75 % accuracy form SUPPORT VECTOR MACHINE and 65% accuracy form Naïve Bayesian classifier.

2.6.2 Luo et al., (2013)

Lue and his colleques in their research paper mentiones that they developed a bag of features into a learning-to-rank scenario which demonstrated excellent power for ranking problem.

They used ranking function such as BM25 (based on Okapi BM25 [36]) and VSM (based on vector space model).

(Luo et al., 2013)

2.6.3 Pak & Paroubek (2010)

Pak & Paroubek (2010) in their report titled “Twitter as a Corpus for Sentiment Analysis and Opinion Mining” proposed a model to classify the tweets into three sentiments as objective, positive and negative. They collected a corpus of text posts by crawling the tweets using Twitter API and automatically annotating those tweets using emoticons.

- Happy emoticons: “:-)”, “:)”, “=)”, “:D” etc.
- Sad emoticons: “:-(”, “:(”, “=(”, “;(” etc.

Using that corpus, they developed a sentiment classifier based on the multinomial Naive Bayes method that uses features like N-gram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

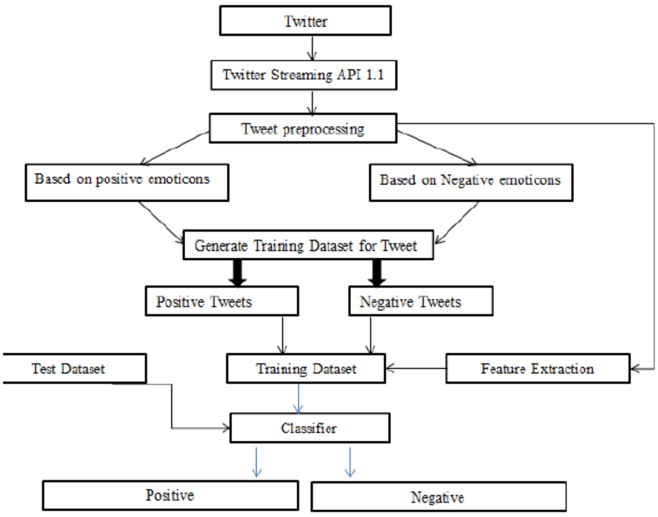


Figure 5 - Sentiment Classification Based On Emoticons

2.6.4 Davidov et al.,(2010)

Davidov et al., (2010) proposed an approach to utilize hashtags in tweets which are posted by the users; as a classification of sentiment type using punctuation, single words, n-grams and patterns as different feature types, which are then combined into a single feature vector for sentiment classification. They used K-Nearest Neighbor strategy to assign sentiment labels by constructing a feature vector for each example in the training and test the trained models.

2.6.5 Agarwal et al., (2011)

Agarwal and his colleagues in their research project developed three models for classifying sentiment into positive, negative and neutral classes.

They experimented with models such as: unigram model, a feature based model and a tree kernel based model. In tree kernel model they represented tweets as a tree. The feature based model uses 100 features and the unigram model uses over 10,000 features. The tree kernel based model outperformed the other two models.

(Agarwal, et al., 2011)

2.6.6 Parikh and Movassate (2009)

Parikh & Movassate (2009) in their reseach they implemented two models to classify tweets. 1. Naive Bayes bigram model and 2. Maximum Entropy model. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model. (Parikh & Movassate, 2009)

2.6.7 Kolchyna et al., (2015)

Kolchaya from UK did a research on “Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination” on 2015. They used Twitter dataset from the SemEval-2013 competition, task 2-B. Their results shows that machine learning method based on Support Vector Machine and Naive Bayes classifiers outperforms the lexicon method.

(Kolchyna, et al., 2015)

2.6.8 Overall Related Work Summery

The below table gives a summarized version of the related works which related our thesis.

Table 2 - Overall Related Work Summery

	Method	**Data Set**	**Acc.**	**Author**
Machine Learning				
	SVM	Movie reviews	86.40%	(Pang & Lee, 2004.)
		Twitter	86.62%	(Kolchyna, et al., 2015)
	CoTraining SVM	Twitter	82.52%	(Liu, 2012)
	Deep Learning and Neural Network	Stanfond Sentiment Treebank	85.70%	Richard
Lexical based	Corpus	Product reviews	74.00%	(Turney, 2002)
	Dictionary	Amazon’s mechanical	---	Taboada

		Turk		
Cross-lingual	Ensemble	Amazon	81.00%	Wan,X
	Co-Training	Amozon, ITI68	81.30%	Wan,X.
	EWGA	IMDb movie review	>90%	Abbasi,A.
	CLMM	MPQA, NTCIR, ISI	83.02%	Mengi
Cross-domain	Active Learning	Book, DVD, Electronics, Kitchen	80% (avg)	Li, S
	Thesaurus			Bollegala
	SFA			Pan S J

2.7 Existing Sentiment Analysis Architecture for Twitter feed

In this section we will be discussing on existing sentimental analysis architecture and steps implemented to analysis the sentiment with related to tweets.

2.7.1 Preprocessing

- Remove non English tweets.
- Remove names of the nouns; such as person, place
- Convert the tweets into the lower case.
- Remove retweeted tweets. As a convention retweeted tweet starts with "RT @username". We will have to remove this since the main tweet will be crawled by the system.
- Remove all the urls (eg. www.suhail.cloud) and unwanted under name targets.
- Correct the spelling and repeated characters will be removed.
- Replace emoticons with the sentiments.
- Remove Stop Words. Stop word is a commonly used words (such as, at, be) which has very minor meaning.
- Remove unwanted punctuations, symbols and numbers

- Expand Acronyms (For expansion an acronym dictionary can be used)

2.7.2 Feature Extraction

The preprocessed dataset contains many distinctive properties. In feature extraction phrase, the extraction of the aspects and distinctive properties of the preprocessed dataset will be done. There after the extracted distinctive properties are used to analyze the positive and negative polarity in a sentence. This is where the determination of the opinion of the individuals are captured. To achieve this, models like unigram, bigram can be used.

2.7.3 Training

Training the classifier can be lead to easier prediction of unknown data.

2.7.4 Classification

When it's comes to classification problem supervised learning technique can be used.

2.8 Sentiment Analysis Approaches for Twitter Data

According to Kharde V., and Sonawane S., there are two best approaches to analyze the sentiment for twitter feeds: 1. Machine Learning Approach, 2. Lexicon-Based Approach.

Machine Learning Approach

Couple of machine learning techniques have been identified to classify the tweets into models. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and Support Vector Machines (SVM) and neural network have achieved great success in sentiment analysis.

According to Amolik et al., machine learning technique is very easier and efficient than symbolic techniques. These techniques are easily applied to twitter sentiment analysis. Further states that among the different machine learning classifiers Naïve Bayesian and Support vector machine performs well and also provide higher accuracy.

Lexicon-Based Approach

A sentiment dictionary will be used in Lexicon based method. The dictionary uses with opinion words and match them with the data to determine polarity. Then sentiment scores will be assigned to the opinions how Positive, Negative and Objective based on the words contained in the dictionary.

Lexicon-based approaches mainly rely on a sentiment lexicon, i.e., a collection of known and precompiled phrases, sentiment terms, and even idioms, developed for traditional genres of communication, such as the Opinion Finder lexicon.

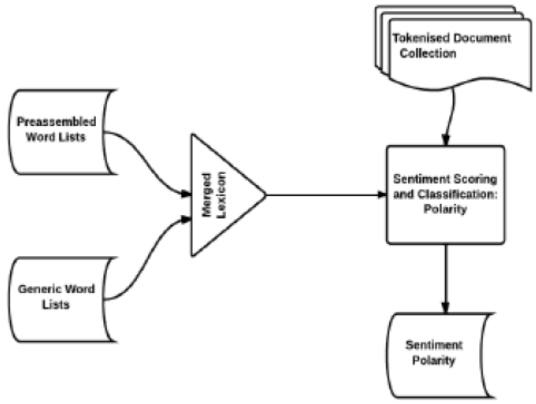

Tokenised Document Collection
Preassembled Word Lists
Merged Lexicon
Sentiment Scoring and Classification: Polarity
Generic Word Lists
Sentiment Polarity

Figure 6 - Lexicon-Based Model

There are two sub classifications for this approach

1. Dictionary based

   It is based on the usage of terms (seeds) that are usually collected and annotated manually. This set grows by searching the synonyms and antonyms of a dictionary. An example of that dictionary is WordNet, which is used to develop a thesaurus called SentiWordNet. The main drawback of this method is that can't deal with domain and context specific orientations. (Kharde & Sonawane, 2016)

2. Corpus based

   In corpus based approach the dictionaries are used with related to specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either statistical or semantic techniques. (Kharde & Sonawane, 2016)

2.9 Problem Definition

Social media becoming the largest user generated content repository when individuals rise complains and give feedbacks through social media like Twitter and Facebook to their telco providers. However, telco providers are facing difficulties in analyzing the data that related to them and their services, categorizing and assigning feedbacks to their support engineers and following them up.

Currently the telco providers manage the issues raised by the subscribers through social media using excel sheets, which are difficult to manage and maintain.

2.10 Summery

In this chapter, I have presented a comprehensive literature review on Sentiment Analysis and identified the research problem, as there are no software solution available for the telecommunication providers where the tweets are analyzed, categorized and maintained.

Chapter 3

Propose Approach and Technologies Adoption

3.1 Introduction

This chapter describes the proposed approach and the solution for the identified problem in this study. This chapter will describe

- Selection of Twitter API
- Authentication and Authorization on Twitter API
- Methodology of crawling the tweets
- How the crawled data are extracted and processed
- How the system allows to track and maintain the feedback
- Twitter Libraries used in order to crawl the tweets
- Azure services used to implement the system
- Data storage mechanism

3.2 Selection of Twitter API

As we discussed in the literature review chapter there are two APIs available in twitter. Among those we have selected Rest API.

This is because Rest API will not maintain a persistent HTTP connection between the server (in our case it's Twitter) and the client (in our case it's out application) where streaming API will maintain a persistent connection. Since there can be multiple Tweets published at the same time our listener will have to analysis them at once; which will need more computational power. At the other hand in Rest API we can request the Twitter time to time in a short interval and crawl the tweets and analyze them as we need.

3.3 Authentication and Authorization on Twitter API

According to the literature review there are two Twitter API authentication models available (User authentication and Application-only authentication) and we will be using User authentication model because of the following reasons.

Application only authentication model will not be able to post Tweets. Our application should be able to respond to the feedback provided by the telecommunication subscriber.

Also in future our system can be customized as to analyze the sentiment and send automatic replies to the direct messages. For this the system should be able to access the direct messages where "User authentication model" can and the Application-only model cannot access the direct messages in twitter.

3.4 Data Storage Mechanism

The data, which related to our application, will be store in Microsoft Azure under Database as a service offering. Azure dataset service is selected because the database is hosted in Microsoft Cloud and the downtime of the database server is minimal to exclusion.

Entity Framework is used in order to communicate with the database and perform CRUD operations in the database.

Entity Framework is an object-relational mapper (O/RM) that enables .NET developers to work with a database using .NET objects. It eliminates the need for most of the data-access code that developers usually need to write. (Microsoft, 2017)

We will be using the latest version of the Entity Framework which is Entity Framework 6.

3.5 Overall Sentiment Analysis Techniques

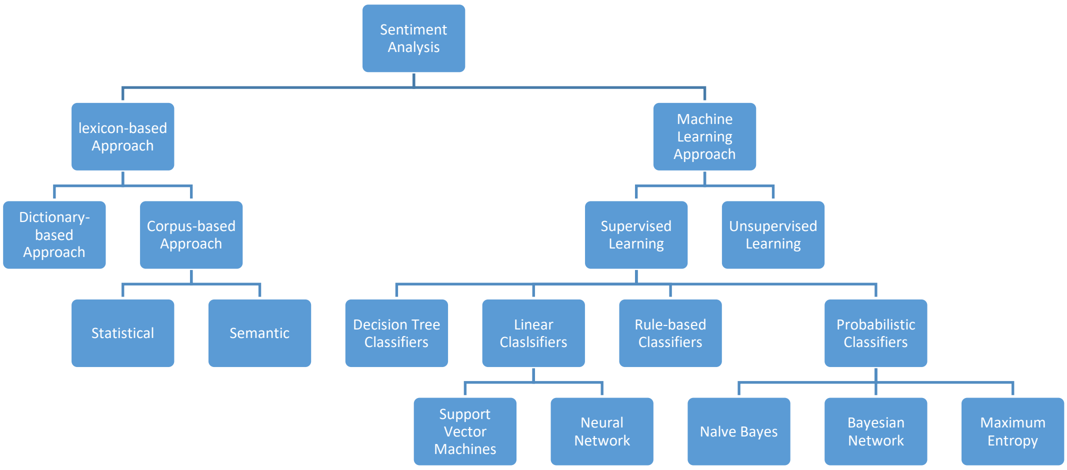


Figure 7 - Overall Sentiment Analysis Techniques

3.6 Selection of Sentiment Analysis Approach

Among the two (1. Machine Learning Approaches, 2. Lexicon Based Approaches.) I have selected machine learning technique.

Machine learning technique uses training data which does not require database of words like used in knowledge-based approach. Therefore, machine learning techniques is better and faster. If the system can analyze the tweets as faster as possible the tweet can be categorized and can be assigned to the support engineer. So that the support engineer can act soon and the response time for a feedback can be reduced.

In the related works section it is clearly mentioned that the Machine learning approaches were performing high accuracy than the Lexicon Based Approaches.

3.7 Selection of Machine Learning Types

Among the two machine learning types (1. Supervised and 2. Unsupervised) we have selected supervised machine learning technique.

The existing training data plays a major contribution on selection of machine learning type. In supervised technique the machine will be able to predict the response by using the exiting training data where for unsupervised we have to provide entire data always to identify the pattern. Since we have training data and we don’t want the pattern to be find using cluster data; we recommend supervised machine learning technique.

And in supervised learning technique we will be using Multi class neural network. We select Multi class neural network because past few years and it has established as a promising alternative to various conventional classification methods.

According to the study carried in Chapter 2, it is clearly mentioned that the Support Vector Machine and Neural Network holds the highest accuracy value than any other algorithms.

Since SVM is suitable for binary classification the multiclass neural network is more suitable for the classification more than two.

3.8 Selection of Machine Learning Offering

We have selected Microsoft Azure Machine learning as the machine learning offering. In Azure machine learning offering we can utilize the inbuilt tools and we can built our own or customized existing tools as we need for cleaning the raw data. This will be easier when it's come to preprocessing data.

Another important aspect of machine learning is programming the algorithm. Usually data scientist use Python or R to code algorithm. In Azure machine learning environment we have inbuilt Python and R code snippets and also Azure machine learning studio provides a wide range of algorithm libraries which we can easily customize and use the depending on the machine learning problem we are going to address.

Another notable advantage of using Azure Machine Learning is that it provides a cheat sheet for machine learning algorithms which helps the developers to choose the right algorithm for a predictive analytics model.

In Azure Machine Learning Studio we can compare more than one selected algorithm. We can pass the same dataset to two different algorithms at the same time and evaluate the performance and accuracy of the algorithms in graphs.

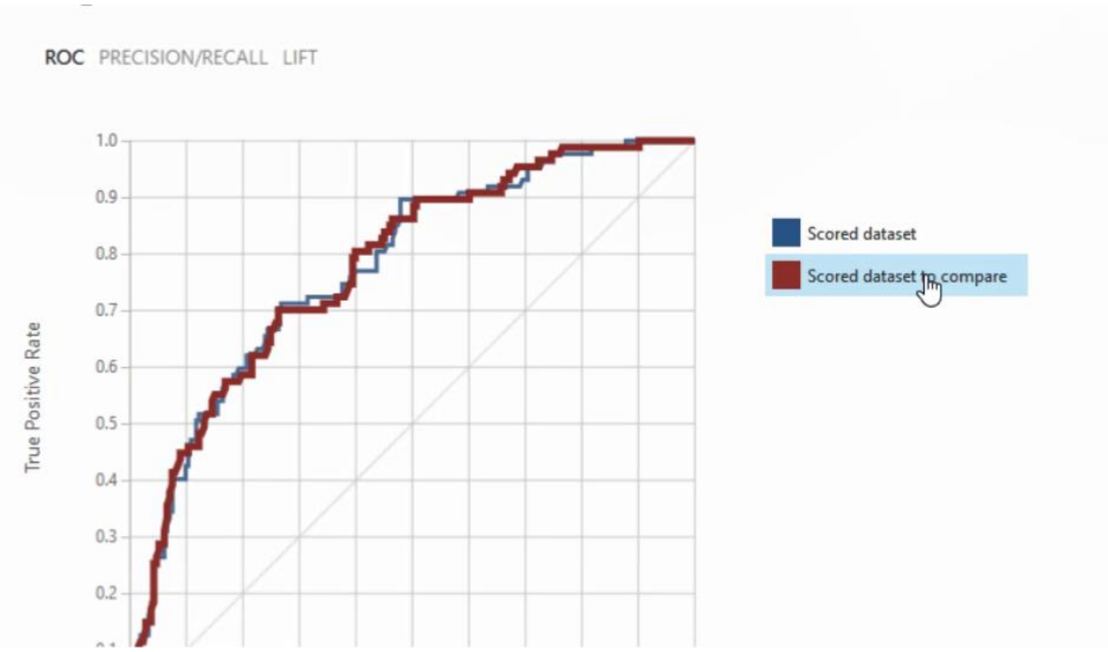


Figure 8 - Compare Performance of Two Algorithms in Machine Learning Studio

The blow diagram will show us, in what phrases the Azure Machine Learning will be used.

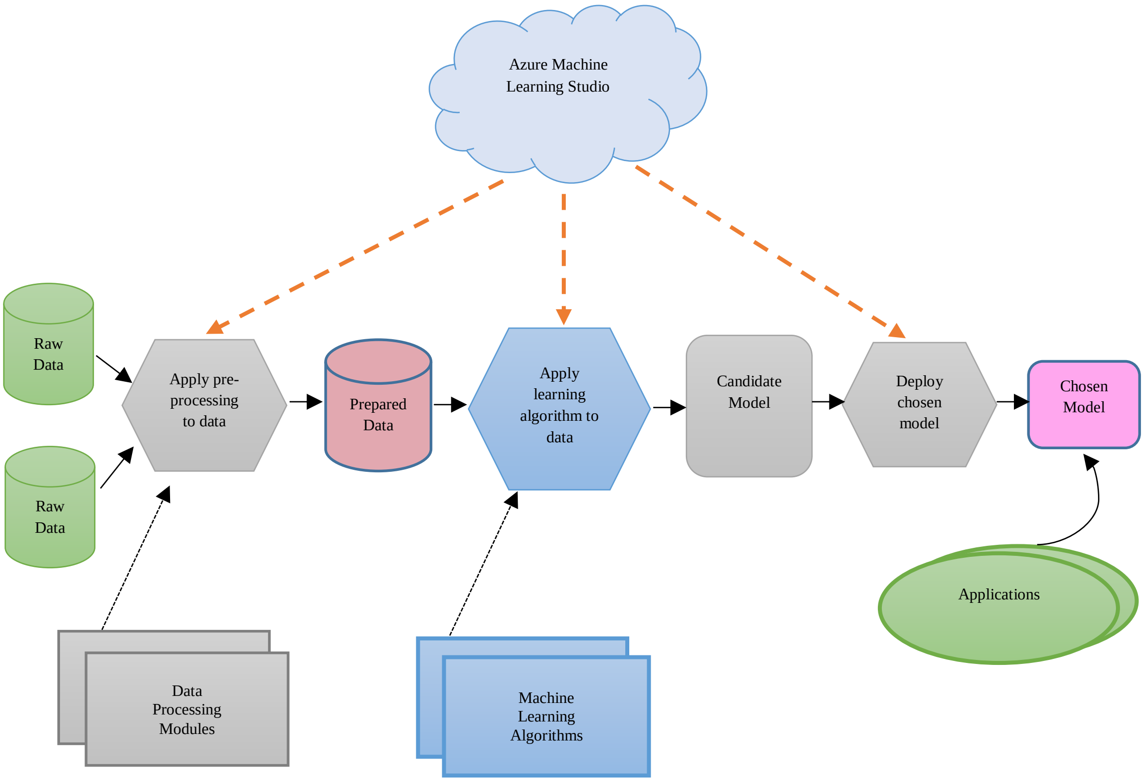

Azure Machine Learning Studio
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model
Applications
Data Processing Modules
Machine Learning Algorithms

Figure 9 - Use of Azure Machine Learning Studio for Machine Learning

3.9 Deployment of Machine Learning Solutions

We have to deploy our machine learning solution somewhere so that the clients can consume our solution and make use of them. In our case it's our own system which will consume the machine learning solution.

We will be deploying the azure machine learning solution as Azure Web App. Azure Web Apps enables to build and host web applications in the programming language of your choice without managing infrastructure. (Anon., 2017)

3. 10 Selection of Software Architecture

Our thesis project consists of a software solution. We will be using the most popular and commonly used three tier architecture 1. Presentation Layer, 2. Business Logic Layer 3. Database Layer).

The benefits of using the three tier architecture can be describes as follows.

- Easy to manage: We can manage each and every tier separately. Since our system connects with external components like Azure function and Web hooks it should be maintained and decoupled.

- Easy to add new features: If we in need of to introduce a new feature, you can add it to the appropriate tier without affecting the other tiers.

In general, three tier architecture can be described as in the below diagram.

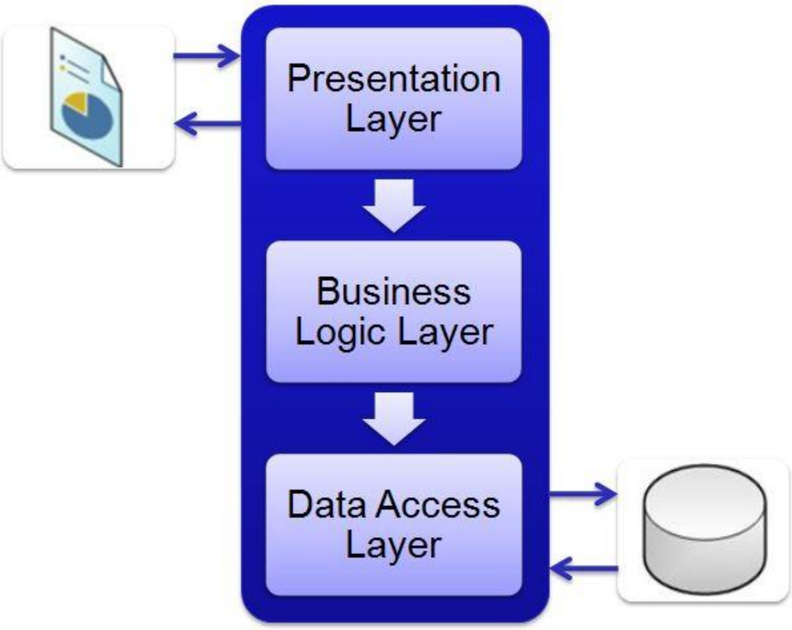


Figure 10 - Three Layer Architecture

3.10.1 Presentation Layer

This is the topmost level of the application and in simple terms it is called as Graphical User Interface (GUI). This layer contains the components that implement and display the user interface and manage user interaction through the controllers like button, textboxes and checkboxes.

3.10.2 Business Logic Layer

The Business logic layer is typically a web service (Rest or SOAP) or framework. This can be considered as the brain of a three-layered system and is responsible for receiving and transmitting requests and responses from and to the presentation client to the database. It also acts as the intermediary to the back-end database.

3.10.3 Data Access Layer

The database tier is typically a database server such as MS SQL, My SQL or Oracle, but it also can be any place where stores the data for future retrieval.

3.11 Hosting and Cloud Offering

Our software solution and sentiment analysis component will be interconnected through. So we will be using any cloud offerings in order to host the system solution. We will be using the Microsoft Azure as our cloud offering.

These are the reasons for using Microsoft Azure than the other cloud solutions like Amazon Web Services (AWS) and Google Cloud.

Internet Information System (IIS) for web applications: Our business logic layer is developed using ASP.Net Web API2 technology. ASP.Net Web API2 should be published and deployed in IIS which as a hosting mechanism for .Net environment. In Azure they have built in Web Apps functionalities where we can host the Web API 2 solutions.

Easy Publish: Using Visual Studio 2017 we can easily publish any solution to Azure environment.

Cost benefits and pricing model: Azure pricing will be based on consumption, still there are free service plans where we can use for development and testing.

Database in Cloud: Azure has its own Database as service offering where without installing MS SQL Server we can directly use a MS SQL database.

3.11 Software Used

In this section we will be discussing what are the software and integrated development environment (IDE) used in order to develop the solution.

Visual Studio Code

Visual Studio Code is an open source code editor developed by Microsoft. We used VS Code do the development for the presentation layer.

Our presentation layer is developed using Angular 5, we had to install certain other extension for the IntelliSense.

Visual Studio 2017

Visual Studio 2017 is IDE developed by Microsoft. We use Visual Studio to develop server side components such as rest services and business logic for the application.

Microsoft MS SQL Server Management Studio

SQL Server Management Studio (SSMS) is a software application by Microsoft which used for configuring, managing and administrating all the components within MS SQL Server and MS SQL Database.

We use “Azure SQL Database as a Service” in order to store data. So SSMS is the ideal tool to communicate with the SQL database which is in Azure.

Chapter 4

Design and Implementation

4.1 Introduction

This chapter describes the software architecture and the four main modules of the system namely

- Tweet Crawling Module
- Text Preprocessing module
- Data Extraction and Sentiment Analysis Module
- Categorizing Module
- Feedback Management Module

Each processing step and modules are further explained in this chapter.

4.2 Research Methodology

The main objective of this research is to investigate and develop a methodology where the system will crawl the tweets made by subscribers of a telecommunication provider, extract the data and categorize them as “feedbacks, complains and compliments”.

Also the system will automatically categorize the crawled tweets into services provided by the particular telecommunication provider. This is based on the score given to the keywords in aligned with the telecom service.

This research is conducted exploring several related techniques used in opinion mining and sentiment analysis. For identifying requirements and system architecture, the information collected from the literature survey are used. From several paths of obtaining the final objective, the following design describes the best method selected by analyzing the literature.

4.3 Design Assumptions

The major part of the research investigates the crawling of tweets, sentiment analysis and categorizing them. Sentiment analysis and opinion mining deals with natural language, thus there are some constraints and limitations.

Input language is limited only to English for this research, and multilingual sentiment analysis is not focused. All the input sentences are assumed syntactically correct, spelled correct and grammatically correct. Abbreviations are considered as out of the scope.

4.4 System Architecture

This section presents the abstract architectural and functional design of the proposed Telco Feedback Management System using Opinion Mining. As described in the introduction section of this chapter the system consists of three main modules.

- Tweet Crawling Module
- Text Preprocessing module
- Data Extraction and Sentiment Analysis Module
- Categorizing Module
- Feedback Management and Archiving Module

Figure 5.1 illustrates overall design architecture of this project including all the main modules.

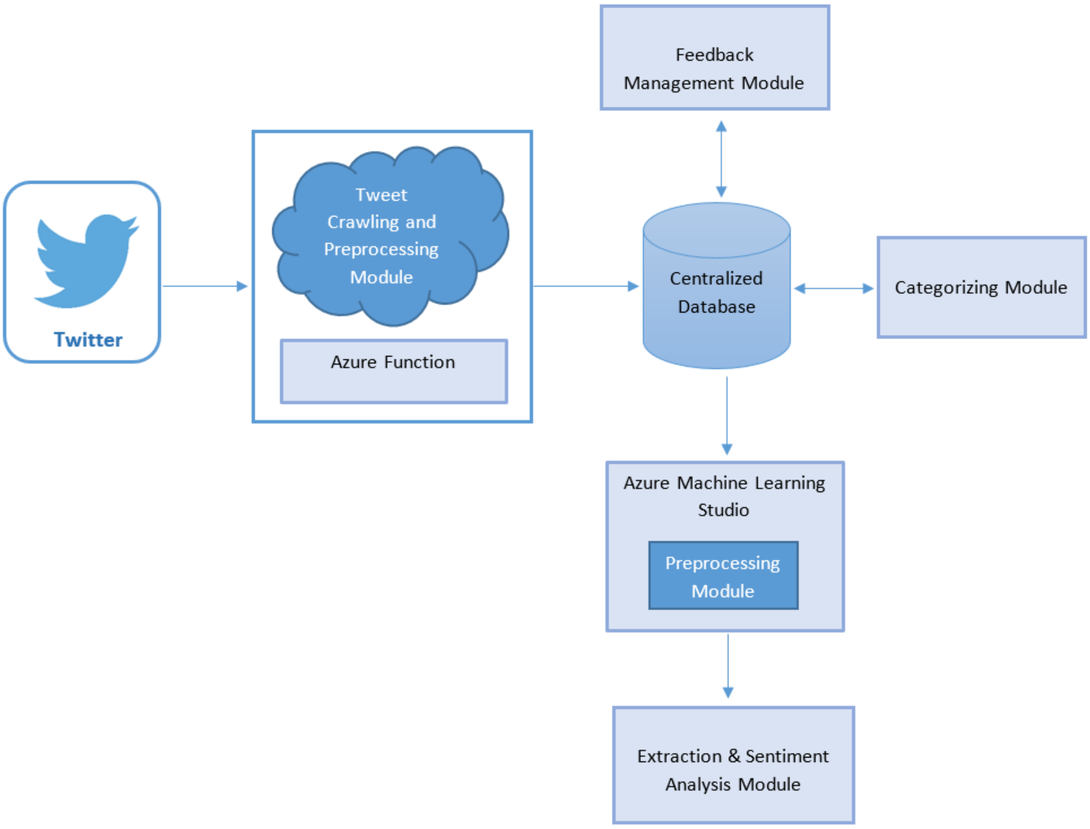

Feedback Management Module
Twitter
Tweet Crawling and Preprocessing Module
Azure Function
Centralized Database
Categorizing Module
Azure Machine Learning Studio
Preprocessing Module
Extraction & Sentiment Analysis Module

Figure 11 - Software Architecture

4.5 Design Components

This section describes functionalities of major modules of the system architecture as described in the section 3.4. Each module is responsible for completing specific abstract task, and together all, these modules complete the sentiment analysis and feedback management system.

4.5.1 Tweet Crawling Module

Our system analyses tweets that are tweeted for a telecommunication provider. In order to analyze the tweets our system should be able to crawl the tweets first.

Crawling component should only crawl the tweets sends on to a specific telecommunication provider. This module should be developed as such it should not crawl the data that are already crawled, also the tweets should be crawled within a frequent timely interval.

Once the tweets are crawled, they will be stored in centralized database and the data will be used by the next module which is preprocessing.

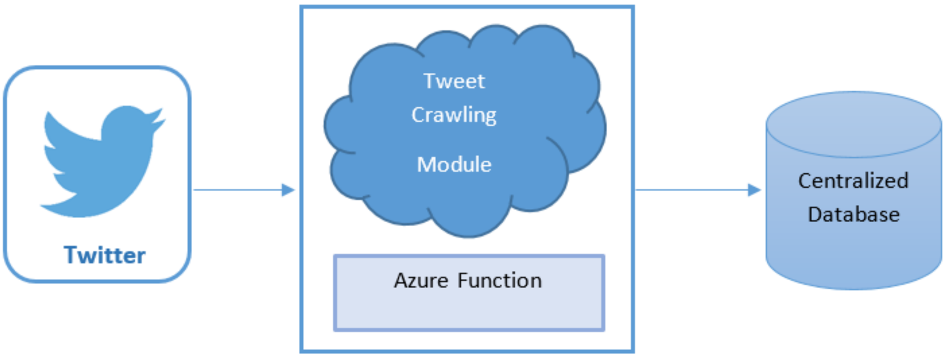


Figure 12 - Tweet crawling module

4.5.2 Preprocessing Module

Tweets may be malformed, incorrectly spelled, grammatically incorrect. In order to prepare the data for the further process, each crawled tweet has to be normalized using some simple filters.

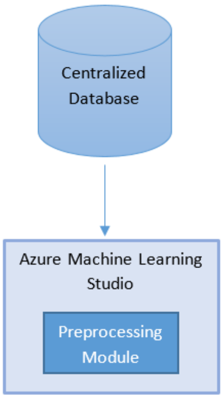


Figure 13 - Preprocessing Module

- Remove non English tweets. Since our system is scoped for only English we have to remove the non-English tweets.
- Replace punctuation, special characters and digits with space. Removing and replacing punctuation, special characters and digits will not affect the features of sentences, ex: “Ph.D.” is kept unchanged while interjections like “hmm. . .” are removed.
- Convert the text into lowercase.
- Remove all the url, emails and unwanted under name targets.

- Correct the spelling and repeated characters will be removed.
- Remove Stop Words. Stop word is a commonly used words (such as, at, be) which has very minor meaning.
- Remove unwanted punctuations, symbols and numbers
- Remove Acronyms

Acronyms and abbreviations are considered as out of the scope while some of the acronyms will be interpreted correctly. Please see Appendix 1 for lists of acronyms that will be interpreted correctly.

4.5.2 Data Extraction and Sentiment Analysis Module

The selection of sentiment analysis technique is discussed in Chapter 3. Data extraction and sentiment analysis module is developed and deployed in Azure Machine Learning Studio.

The final solution is exposed through a web service and we can consume the service and evaluate the tweet send to the system through API.

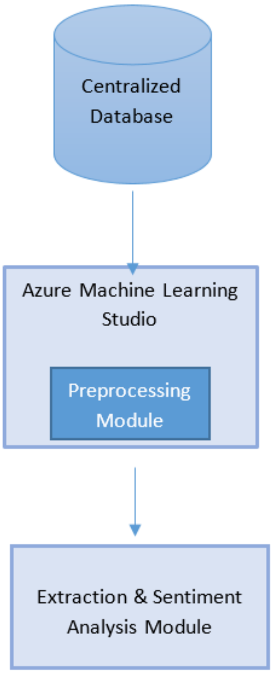

Centralized Database
Azure Machine Learning Studio
Preprocessing Module
Extraction & Sentiment Analysis Module

Figure 14 - Sentiment Analysis Module

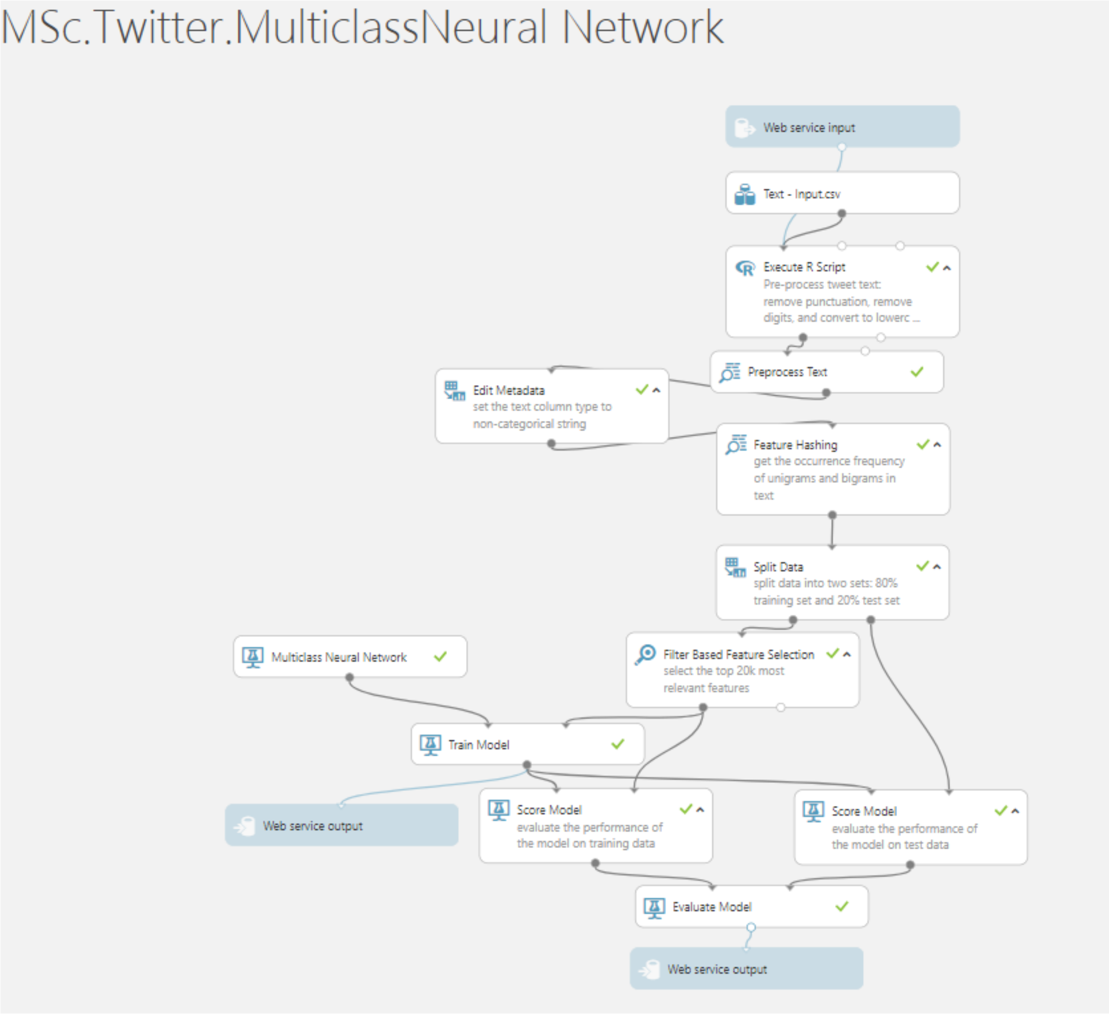

MSc.Twitter.MulticlassNeural Network
Web service input
Text - Input.csv
Execute R Script
Pre-process tweet text: remove punctuation, remove digits, and convert to lowerc ...
Preprocess Text
Edit Metadata
set the text column type to non-categorical string
Feature Hashing
get the occurrence frequency of unigrams and bigrams in text
Split Data
split data into two sets: 80% training set and 20% test set
Multiclass Neural Network
Filter Based Feature Selection
select the top 20k most relevant features
Train Model
Web service output
Score Model
evaluate the performance of the model on training data
Score Model
evaluate the performance of the model on test data
Evaluate Model
Web service output

Figure 15 – Multiclass Neural Network Method

4.5.3 Feedback Management Module

The feedback management module is developed using the separation of concerns as three layers as discussed in the Chapter 3.

This is the module where the end users will directly involve with the application solution. The administrator will assign issues/tweets to the supports engineer and the support engineer will comment on the issue. The comment will be directly posted in the user’s twitter feed then and there.

Also there are automatic response will be send as direct message to the subscribers if the issue is a positive feedback.

The feedback management module of our system will architecture as follows.

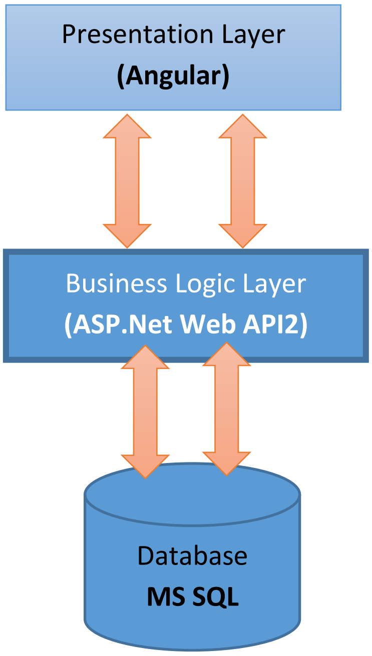


Figure 16 - Feedback Management Module

We will be using the technologies in the layers as follows.

Presentation Layer: Angular 5

Business Logic Layer: Web API 2

Database Layer: MS SQL

Chapter 5

Evaluation

In this chapter we will be discussing on the evaluations and testing carried out after the system development and implementation.

5.1 Evaluation

For the evaluating purposes we crawled and gathered around 200 random tweets which were mentioned to Dialog Axiata Sri Lanka (@dialoglk) and Mobitel Sri Lanka (@MobitelSriLanka). Using the implemented system we tested the crawled tweets.

Sentiment Analysis

The below results shows the sentiment results and its accuracy percentage

Table 3 - Evaluation Results for Sentiment Analysis

Sentiments	Number of Samples	Number of Correct Result	Accuracy (%)
Positive	24	22	91.67
Negative	26	20	76.92
Question	26	22	84.61
Statement	32	30	9.75

Service Category

The below table displays the service category evaluated results.

Table 4 - Evaluation Results for Service Category

Service Category	Number of Samples	Number of Correct Result	Accuracy (%)
Broadband	11	10	90
TV	25	22	88

Prepaid	26	25	96
Postpaid	27	22	81
General / Others	30	30	100

Chapter 6

Discussion

All the previous chapters discussed the identified problem, literature review, proposed approaches and proposed solution. In this chapter we will be discussing on limitations of the current system, challenges faced and future works.

6.1 Challenges Faced During the Sentiment Analysis

Sentiment Analysis is a very challenging task. Following are some of the challenges faced in Sentiment Analysis of Twitter.

- **Sarcasm Detection**

  Sarcastic sentences express negative opinion about a target using positive words in unique way. Basically when it's come to a telecommunication provider more people use sarcasm words to mock a service or any support engineer.

  E.g. “Nicely done team. You are fixing this issue for last 5 hours”

- **Entity Recognition**

  One tweet can hold more than one sentiments towards different entities. Separating out the text about a specific entity and then analyzing sentiment towards the tweet is a difficult process.

  E.g. “Your Broadband is worst while your mobile connection is awesome”

  A simple bag-of-words approach will label it as neutral, however, it carries a specific sentiment for both the entities present in the statement.

  In any case, where both positive and negative sentiments taken place the system will mark it as negative because the negative should be addressed by a support engineer.

- **Handling Comparisons**

When there are comparison tweets which related to services of the particular telecommunication provider identifying analyzing those are challenged. In most of the cases the system marked as neutral.

E.g. “I have used both Dialog Mobile Broadband and Home Broadband, Mobile broadband is better than the Home Broadband”

- **Selection of Machine Learning Techniques and Algorithms**

When it’s come to machine learning technique, earlier steps should be chosen wisely after doing a better research since the later steps are depend on the first stages. And unfortunately we cannot travel through backwards; we have to start from the very initial stage which is preprocessing data.

6.2 Future Work

Current system is capable of sentiment classification only in English language. Making the Tweets analyzed in other languages and managing the feedback for them can be done in future.

Also current system is not capable auto responding to the tweets by the telecommunication subscribers. Some questions like “what are the broadband packages available” can be answered automatically by a chatbot; to achieve this we can use artificial intelligence APIs and natural language processing.

References

Pang, B., Lee, L. & Vaithyanathan, S., 2002. *Thumbs up? Sentiment Classification using Machine Learning Techniques.* Philadelphia, Association for Computational Linguistics.

Agarwal, A. et al., 2011. *Sentiment Analysis of Twitter Data,* New York: Department of Computer Science, Columbia University.

Amazon, 2018. *Amazon Machine Learning.* [Online]
Available at: https://aws.amazon.com/aml/

Amolik, A., Jivane, N., Bhandari, M. & D., 2016. Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques.. *International Journal of Engineering and Technology,* p. 2038.

Anon., 2016. *Internet usage statistics in Sri Lanka.* [Online]
Available at: http://www.digitalmarketer.lk/internet-usage-statistics-in-sri-lanka.html

Anon., 2017. *Web Apps Documentation.* [Online]
Available at: https://docs.microsoft.com/en-us/azure/app-service/

Anon., 2018. *Azure Machine Learning Studio.* [Online]
Available at: https://azure.microsoft.com/en-us/services/machine-learning-studio/

Anon., 2018. *Bayesian networks - an introduction.* [Online]
Available at: https://www.bayesserver.com/docs/introduction/bayesian-networks

Anon., 2018. *SAS Institute Inc..* [Online]
Available at: https://www.sas.com/en_us/home.html

Aslam, S., 2018. [Online]
Available at: https://www.omnicoreagency.com/twitter-statistics/

Becker, L., Erhart, G., Skiba, D. & Matula, V., 2013. AVAYA: Sentiment Analysis on Twitter with Self-Training. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic,* p. 333–340.

Boiy, E., Hens, P., Deschacht, K. & Moens, M. F., June 2007. Automatic Sentiment Analysis in On-line Text. *Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria.*

Cambria, E., 2013. *An Introduction to Concept-Level Sentiment Analysis,* s.l.: Temasek Laboratories, National University of Singapore.

Davidov, D., Tsur, O. & Rappoport, A., 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. *Coling 2010: Poster Volume*, August, p. 241–249.

Doyle, D., n.d. *Ranks NL Webmaster Tools.* [Online]
Available at: https://www.ranks.nl/
[Accessed 13 03 2018].

Ehrlich, K. & Shami, N. S., 2010. Microblogging Inside and Outside the Workplace. *IBM TJ Watson Research Center and Center for Social Software.*

Ericson, G. & Rohm, W. A., 2017. *What is Azure Machine Learning Studio?.* [Online]
Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio

Grandi, R. & Neri, F., 2014. Sentiment Analysis and City Branding. *New Trends in Databases and Information Systems: 17th East European Conference on Advances in Databases and Information Systems,* pp. 339-344.

Gupte, A., Joshi, S., Gadgul, P. & Kadam, A., 2014. Comparative Study of Classification Algorithms used in Sentiment Analysis. *International Journal of Computer Science and Information Technologies,* pp. 6261 - 6264.

IBM, 2018. *IBM SPSS Software.* [Online]
Available at: https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software

Kalaria, A. & Prajapati, Z., 2016. Opinion Mining for Information Retrieval: Survey. *International Journal of Computer Science and Network,* pp. 934 - 940.

Kharde, V. A. & Sonawane, S., 2016. Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications (.*

Kolchyna, O., Souza, T. T. P., Treleaven, P. C. & Ast, T., 2015. *Twitter Sentiment Analysis: Lexicon Method, Machine,* London: Department of Computer Science, UCL, Gower Street,.

Liu, B., 2011. *Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data.* s.l.:Springer.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies,* p. 167.

Luo, Z., Osborne, M. & Wang, T., 2013. An effective approach to tweets opinion. *World Wide Web*, 5-6 November.

Microsoft, 2017. *Entity Framework Documentation.* [Online]
Available at: https://docs.microsoft.com/en-us/ef/#pivot=entityfmwk&panel=entityfmwk1

Microsoft, 2017. *Language Understanding (LUIS).* [Online]
Available at: https://azure.microsoft.com/en-us/services/cognitive-services/language-understanding-intelligent-service/
[Accessed 12 3 2018].

Ortony, A., Clore, G. L. & Collins, A., 1990. *The Cognitive Structure of Emotions.* Cambridge : Cambridge University Press.

Pak, A. & Paroubek, P., 2010. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining,* Bˆatiment: Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS.

Paltoglou, G. & Thelwall, M., 2010. A study of Information Retrieval weighting schemes for sentiment analysis. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* p. 1386–1395.

Pang, B. & Lee, L., 2004.. A sentimental education: Sentiment analysis using. *In Proceedings of the 42nd annual meeting on Association for Computational Linguistics,* p. 271.

Parikh, R. & Movassate, M., 2009. *Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques,* s.l.: s.n.

RapidMiner, 2018. [Online]
Available at: https://rapidminer.com/us/

Sumathi, S. & Sivanandam, S., 2006. *Introduction to Data Mining Principles, Studies in Computational.* s.l.:Springer-Verlag Berlin Heidelberg.

Turney, P. D., 2002. *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.* Philadelphia, s.n., pp. 417-424.

Twitter, 2017. *Twitter Developer Portal.* [Online]
Available at: https://developer.twitter.com/

Twitter, 2017. *Twitter Developer Portal.* [Online]
Available at: https://developer.twitter.com/

Wilson, T., Wiebe, J. & Hoffmann, P., 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language,* pp. 347 - 354.

Appendixes

1.0 Acronyms and their meaning used in preprocessing

- UGC : User Generated Content
- WWW: World Wide Web
- ASAP : As soon as possible
- 3G : Third Generation (mobile communication system)
- 4G : Fourth Generation
- ADSL : Asymmetric Digital Subscriber Line
- CCTV : Closed Circuit Television
- CRUD : Create, read, update and delete
- AML : Azure Machine Learning
- LUIS : Microsoft Language Understanding Intelligent Services
- API : Application Programming Interface

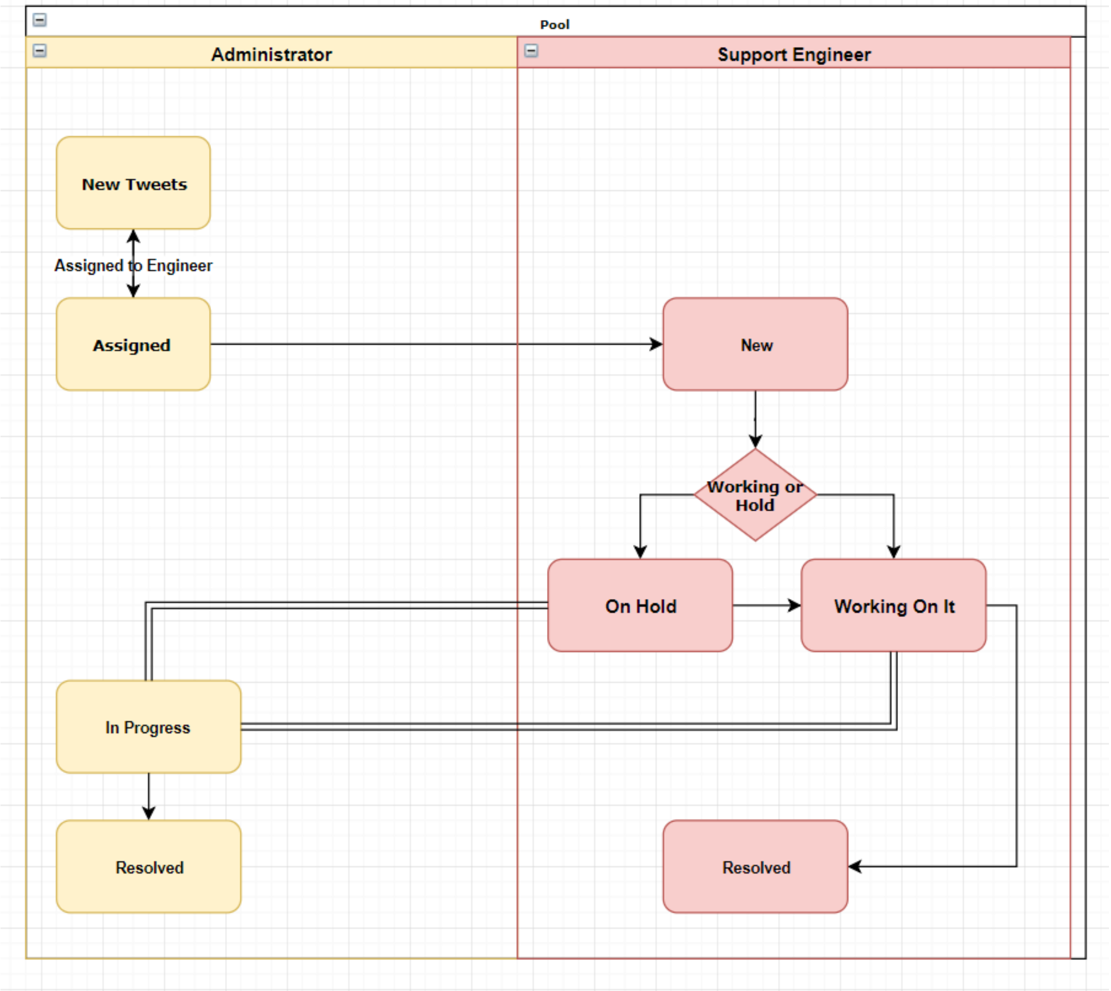
2.0 Stopwords

- a
- about
- above
- after
- again
- against
- all
- am
- an
- and
- any
- are
- aren’t
- as
- at
- be
- because
- been
- before
- being
- below
- between
- both
- but
- by
- did
- didn’t
- do
- does
- doesn’t
- doing
- don’t
- down
- during
- each
- few
- for
- from
- further
- had
- hadn’t
- has
- hasn’t
- have
- haven’t
- having
- he
- he’d
- he’ll
- he’s
- herself
- him
- himself
- his
- how
- how’s
- i
- i’d
- i’ll
- i’m
- i’ve
- if
- in
- into
- is
- isn’t
- it
- it’s
- its
- itself
- let’s
- me
- more
- most
- mustn’t
- not
- of
- off
- on
- once
- only
- or
- other
- ought
- our
- ours
- ourselves
- out
- over
- own
- same
- shan’t
- she
- she’d
- she’ll
- she’s
- should
- shouldn’t
- so
- some
- the
- their
- theirs
- them
- themselves
- then
- there
- there’s
- these
- they
- they’d
- they’ll
- they’re
- they’ve
- this
- those
- through
- to
- too
- under
- until
- up
- very
- was
- wasn’t
- we’ve
- were
- weren’t
- what
- what’s
- when
- when’s
- where
- where’s
- which
- while
- who
- who’s
- whom
- why
- why’s
- with
- won’t
- would
- wouldn’t
- you
- you’d
- you’ll
- you’re
- you’ve

can’t	her	my	such	we	yours
cannot	here	myself	than	we’d	your
could	here’s	no	that	we’ll	yourself
couldn’t	hers	nor	that’s	we’re	yourselves

(Doyle, n.d.)

3.0 Cycle of an Issue


Pool
Administrator
Support Engineer
New Tweets
Assigned to Engineer
Assigned
New
Working or Hold
On Hold
Working On It
In Progress
Resolved
Resolved