MULTI LABEL TWEET CLASSIFICATION AND APPLICATION

Enrollment No	**:**	1613313117
Name	**:**	Vishesh Sharma
Supervisor	**:**	Dr. Mayank Deep Khare

March – 2019

Submitted in partial fulfillment of the Degree of

Bachelor of Technology

Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY

NOIDA INSTITUTE OF ENGINEERING & TECHNOLOGY, NOIDA

TABLE OF CONTENTS

Chapter No.	Topics	Page No.
	Student Declaration	iii.
	Certificate	iv.
	Acknowledgement	v.
	Summary	vi.
	List of Figures	vii.
	List of Tables	viii.
	List of Symbols and Acronyms	ix.
Chapter-1	**Introduction**	1-5
	1.1 General Introduction	
	1.2 List some relevant current/open problems.	
	1.3 Problem Statement	
	1.4 Overview of proposed solution approach and Novelty/benefits	
Chapter-2	**Background Study**	6-13
	2.1 Literature Survey	
	2.1.1 Summary of papers	
	2.1.2 Integrated summary of the literature studied	
	2.2 Summary of field survey, experimental studies, new tools	

Chapter 3: **Analysis, Design and Modeling** 14-22

3.1 Requirements Specifications

3.1.1 Functional Requirements

3.1.2 Non Functional requirements

3.2 Overall Architecture

3.3 Design Documentation

3.3.1 Use Case diagram

3.3.2 Control Flow Diagram

3.3.3 Sequence Diagram

3.4 Risk Analysis and Mitigation Plan

Chapter-4 **Implementation and Testing** 22-28

4.1 Implementation details and issues

4.2 Testing

4.2.1 Testing Plan

4.2.2 Component decomposition and type of testing required

4.2.3 List of Test Cases

Chapter-5 **Findings & Conclusion** 29-31

5.1 Findings

5.2 Conclusion

5.3 Future Work

Appendix 32

References 33

DECLARATION

I/We hereby declare that this submission is my/our own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Noida Signature: ……………………..

Date: March 21,2019 Name: VISHESH SHARMA

Enrollment No.: 1613313117

CERTIFICATE

This is to certify that the work titled “ **MULTILABEL TWEET CLASSIFICATION AND APPLICATION**” submitted by “**VISHESH SHARMA**” in partial fulfillment for the award of degree of **BACHELOR OF TECHNOLOGY** of NOIDA INSTITUTE OF ENGINEERING & TECHNOLOGY University, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor	……………………..
Name of Supervisor	Dr. Mayank Deep Khare
Designation	Assistant Professor (Senior Grade)
Date	March 21,2019

ACKNOWLEDGEMENT

I would like to extend my sincere & heartfelt obligation towards all the people who have helped me in this endeavor. Without their active guidance, help, cooperation & encouragement, I would not have made headway in the project.

I am ineffably indebted to my mentor **DR. MAYANK DEEP KHARE** for conscientious guidance and encouragement to accomplish this assignment. I am extremely thankful to him pay my gratitude towards him for her/his valuable guidance and support on completion of this project in its presently.

I extend my gratitude to my college **NOIDA INSTITUTE OF ENGINEERING & TECHNOLOGY,NOIDA** for giving me this opportunity.

I also acknowledge with a deep sense of reverence, my gratitude towards my parents and member of my family, who has always supported me morally as well as economically.

At last but not least gratitude goes to all of my friends who directly or indirectly helped me to complete this project report.

Signature of the Student ……………………..

Name of Student VISHESH

SHARMA

Enrollment Number 1613313117

Date March 21,2019

SUMMARY

Social Networking today has become a very popular communication tool among Internet users. Among all the social networking sites, Twitter serves as a great place for social web mining because of its openness for public consumption, API documentation, rich developer tool, and broad appeal to users. Among all knowledge that can be extracted from tweets, this project focuses on two aspects: (1) Sentiment of a tweet (2) Topic of a tweet

This Project tends to focus on the problem of classification, i.e., given a set of pre-determined classes, how to identify to which classes an instance belongs to on the basis of both Sentiment as well as Topic of Tweet. Extending the classification Model, we have employed this classification to provide a tool for comparison of products for the consumers as well as provide a tool to the company or manufactures to help them understand their brand reputation on popular social networking site.

Signature of Student: ……………………..	Signature of Supervisor: ……………………..
Name: VISHESH SHARMA	Name: DR. MAYANK DEEP KHARE
Date: March 21,2019	Date: March 21,2019

LIST OF FIGURES

- Figure 1. Research Paper 3: System Architecture
- Figure 2. Research Paper 5: Tweet Topic and Sentiment Labels
- Figure 3. Integrated Summary: Fans Reacting to Underdog Comeback
- Figure 4. Integrated Summary: Tweets Classification under 18 Categories
- Figure 5. Integrated Summary: MTML model
- Figure 6. Integrated Summary: Sample Tweets
- Figure 7. Use Case Diagram
- Figure 8. Control Flow Diagram
- Figure 9. Sequence Diagram

LIST OF TABLES

- Table 1: Risk Analysis and Mitigation Plan
- Table 2: Testing Plan
- Table 3: Component Decomposition and Type of Testing Required
- Table 4: Tweet Sentiment Classification Test Cases
- Table 5: Tweet Topic Classification Test Cases
- Table 6: Polarity Calculation Test Cases

LIST OF SYMBOLS AND ACRONYMS

ix.

- **SVM :** Support Vector Machine
- **CRF:** Conditional Random Field
- **NN:** Neural Networks
- **NB:** Naive Bayes
- **ME:** Maximum Entropy
- **API:** Application Program Interface
- **MTML:** Multi-Task Multi-Label
- **IDE :** Integrated Development Environment
- **GUI:** Graphic User Interface

CHAPTER 1

INTRODUCTION

1.1 General Introduction

Social networking today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life every day. Social Networking web-sites become a valuable sources of people's opinions and sentiments. Users tend to post about products and services they use, or express their political and religious views.

Twitter serves as a great place for social web mining because of its openness for public consumption, API documentation and ease of use, rich developer tool such as Twitter4J, and broad appeal to users from every walk of life. Twitter data is helpful and interesting because tweets happen at the "blink of eye" and are available to use on real time data. Tweets link people in a variety of ways, ranging from short conversational dialogues to expressing strong political religious views.

Among all knowledge that can be extracted from tweets, this project focuses on two aspects:
(1) **Sentiment of a tweet**: such as "positive" and "negative", i.e., the mood of the user;
(2) **Topic of a tweet:** such as "Inquiry", "Mention", and "Service And Support", i.e. , the subject of the respective tweet

This Project tends to focus on the problem of classification, i.e., given a set of pre-determined classes, how to identify which classes an instance belongs to on the basis of both Sentiment as well as Topic of Tweet with a motive of providing the audience a taste of real experiences of people with certain products or services.

Word-of-mouth is the process of conveying information from person to person. In commercial situations, word of-mouth (WOM) involves consumers sharing details, their opinions, or reactions about businesses, products, or services with other people. This WOM branding is a captivating, influential, multifaceted, and typically hard to influence form of product marketing.
Positive WOM branding is considered as a powerful marketing medium for companies to influence consumers. WOM branding is based on social networking and trust: people rely on families, friends, and others in their social network from commercial insights.

Furthermore, extending the classification model, this Project aims to provide Consumers and Company/Manufacturers a tool to serve them developing a brand image.

(1) **Comparison Tool for Consumers:** The above classification into multi-label classification is employed to provide consumers with a comparison tool to help them decide which a better product to go for is.

(2) **Analytical Tool for Company:** The above classification is also employed to help a company understand it's brand image in the social media which serves as electronic word of mouth

1.2 Current/Open Problems

In this project, we examine the expressions of brand attitudes in micro-blog postings. However, there are several unanswered questions concerning this new topic. How prevalent are branding micro-blogs? How do people structure these micro-blogs? What are their effects on online reputation management? These are the questions that motivate this project.

- To explore other classification tasks and extend the MTML model for more realted tasks
- To work upon more wide range of products and services.
- To improve the overall performance of multi-label approach by taking different features and a larger training set into account.
- It may be interesting to compare products on different domains and scales.

1.3 Problem Statement

This project aims to extract the treasury of data obtained from micro-blogging site, Twitter, classify it and exploit its benefits. This Project tends to focus on the problem of classification, i.e., given a set of pre-determined classes, how to identify which classes an instance belongs to on the basis of both Sentiment as well as Topic of Tweet. Also, this classification can further be employed as a tool for comparison for various kinds of products on the basis of real time experiences of people.

Contribution:

A multi-label classification of tweets of different products or services under different categories, either generic or specific, to provide a basis of comparison between any two of the products/ or services. This approach can serve as tool for both consumers and producers of that particular product or service.

For Consumer: to select among different products or services on the basis of various compared categories.

For Producers: to find the loopholes in their product and service image and make a room for improvement.

1.4 Overview of Proposed Solution Approach

- Extract the Tweets according to the keywords entered by the user.
  For example, a user wanting to compare two mobile phones, say, Iphone6 and OnePlusTwo.

- Aims to figure out what sentiments are attached to the extracted tweet.
  For example, “In love with me and camera! #Iphone6” –Positive
  “Software uhh!! Bad Choice #OnePlusTwo” –Negative

- Aims to classify the data set further based on thedata dependentfeatures (different classes, which will be identified after proper collection of data set for a particular kind of product)
  For example, Classes could be “Features”, “Inquiry”, and “Complaint” etc.

- To prepare a comparison tool for different brands of a product providing conclusions on real time data sets.

- To come up with a tool for the producers to help them measure their brand image on social media and exploit its benefits

Novelty:

- The granularity of classification of the data sets

- A comparison tool for products providing an insight of product to the consumers on the basis of real time experiences of people.

- A analytical tool for identifying brand image of a company

CHAPTER 2

BACKGROUND STUDY

2.1 Literature Survey

2.1.1 Summary of Relevant Papers

Research Paper – 1

Title:	Twitter as a Corpus for Sentiment Analysis and Opinion Mining
Author:	Alexander Pak, Patrick Paroubek
Publication:	Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS,Bˆatiment 508,F-91405 Orsay Cedex, France
Publication year:	2010
Web Link:	http://lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

Abstract: This paper focuses on using Twitter for the task of sentiment analysis. With the help of collection of a dataset formed of messages from Twitter , this paper aims to build a sentiment classifier, which is able to determine positive, negative and neutral sentiments attached to particular tweet. They collected a corpus of 300000 text posts from Twitter evenly split automatically between three sets of texts:

1. Texts containing positive emotions, such as happiness, amusement or joy.

2. Texts containing negative emotions, such as sadness, anger or disappointment.

3. Objective texts that only state a fact or do not express any emotions.

They have built a sentiment classifier using the multinomial NaıveBayes classifier that uses N-gram and POS-tags as features. They tried SVM and CRF, however the Naıve Bayes classifier yielded the best results.

Naıve Bayes classifier is based on Bayes’ theorem:

P (s|M) =P (s) · P (M |s)

P (M)where s is a sentiment, M is a Twitter message.

Research Paper – 2

Title: Supervised Multi-Class Classification of Tweets

Author: Zahan Malkani , Evelyn Gillie

Publication: Stanford University, Stanford, California

Publication year: 2012

Web Link: http://cs229.stanford.edu/proj2012/GillieMalkani-SupervisedMulticlassClassificationOfTweets.pdf

Abstract: This paper presents a study of a variety of supervised learning methods used for the problem of Twitter tweet classification. This includes Support Vector Machines (SVM), Neural Networks (NN), Naive Bayes (NB), and Random Forests.
This paper aims to evaluate these methods using their performance on two sources of labeled data
1) Labeling tweets with an attitude from the speaker's perspective, and;
2) Labeling tweets based upon topics

The Case Study they worked upon: "Fans Reacting to an Underdog Comeback". This paper collected approximately 80,000 tweets containing the keywords \Saints" or \Sea-hawks" between 1:00 pm and 4:00 pm on the day of a playoff game between the two teams. The interesting part is that both teams observe patterns related to the score of their teams.

A grid search SVM over a Radial Basis Function Kernel outperformed NB, NN and Random forests.

Research Paper – 3

Title: Twitter Trending Topic Classification

Author: Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary

Publication: 11th IEEE International Conference on Data Mining Workshops

Publication year: 2011

Web Link: http://cucis.ece.northwestern.edu/publications/pdf /LeePal11.pdf

Abstract: It is often difficult to understand what are the trending topics in Twitter. For easier and effective information retrieval, classification of these topics into some categories is therefore important. This paper classifies Twitter Trending Topics into 18 general categories (arts & design, books, business, charity & deals, fashion, food & drink, health, holidays & dates, humor, music, politics, religion, science, sports, technology, TV & movies, other news, and other.) with 2 approaches for topic classification; (i)Bag-of-Words approachand (ii) network-based classification. All the tweets containing a trending topic constitute a document.

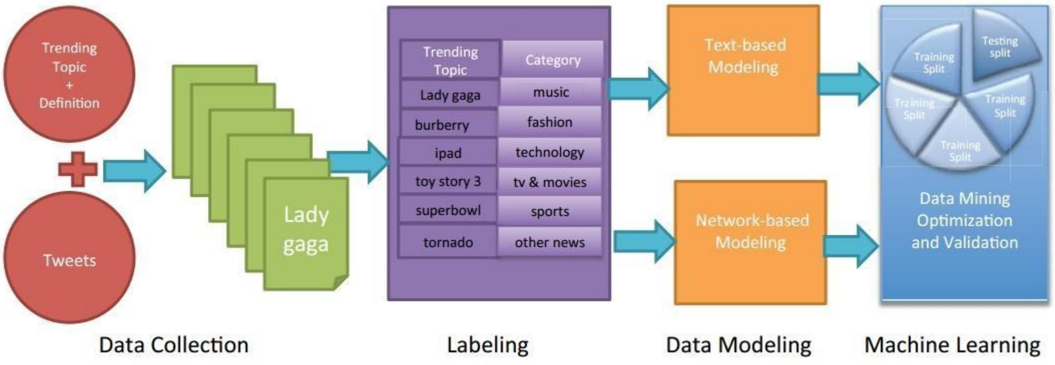


Figure 2. System Architecture. The proposed classification system consists of four stages: (1) Data Collection stage - trending topic, topic definition and tweets are downloaded to compose a document; (2) Labeling stage - over 3000 topics are manually labeled into 18 general classes; (3) Data Modeling stage - (i) Text-based Modeling stage - documents are run through a *string-to-word vector* kernel and converted to tokens with tf-idf weights (ii) Network-based Modeling stage - for each trending topic, 5 most similar topics are computed; (4) Machine Learning stage - various classification schemes are applied using 10-fold cross validation to find the best classifier.

Figure1. Research Paper3 : System Architecture

Research Paper – 4

Title: Evaluating Multi-label Classification of Incident-related Tweets

Author: Axel Schulz, Eneldo Loza Mencía, Thanh Tung Dang, Benedikt Schmidt

Publication: #Microposts2014, 4th Workshop on Making Sense of Microposts

Publication year: 2014

Web Link: http://ceur-ws.org/Vol-1141/paper_01.pdf

Abstract: There is a loss of important information that can be valuable for interpreting certain results if the classification techniques used only assigns a single label to a Tweet. This paper contributes towards analysis of multi-label classification of incident-related tweets.

For example: "THIS CAR HIT THE FIRE HYDRANT AND CAUGHT FIRE. SOMEONE HOLIDAY ALTERED", all labels (”fire”, "crash", and "injuries") should be assigned concurrently instead of a single label. Thus the assignment of multiple labels may not necessarily be an independent process.

The taken approach is composed of three steps.

- Unstructured text -> structured text
- The structured information -> features
- These features are used to train and evaluate a classifier.

The results used four labels consisting of very common and distinct incident types and the injury label: Fire, Shooting, Crash, and Injury.

Research Paper – 5

Title: Sentiment and Topic Analysis on Social Media:
A Multi-Task Multi-Label Classification Approach

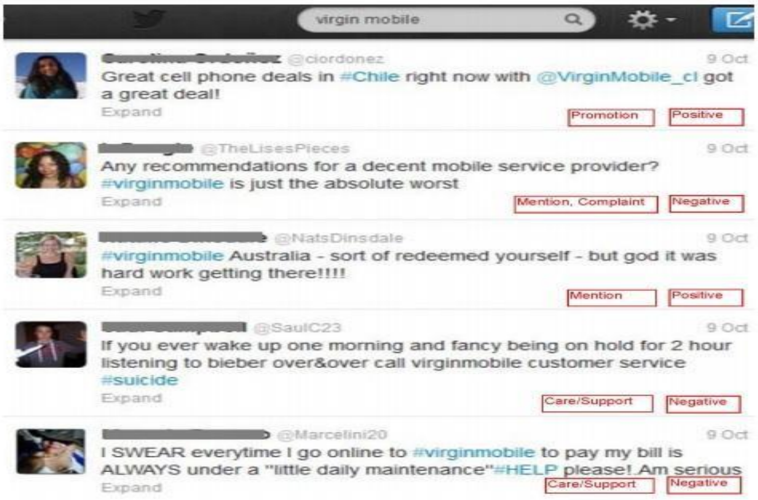
Author: Shu Huang, Wei Peng, Jingxuan Li, Dongwon Lee

Publication: WebSci'13 May 1 – May 5, 2013

Publication year: 2013

Web Link: http://pike.psu.edu/publications/websci13.pdf

Abstract: Customer care and Marketing use Sentiment Analysis and Topic Classification frequently. However, authors believe convectional solutions of classification suffer a drawback as they consider sentiment analysis and topic classification as two isolated topics. Also, they tend to classify a tweet under single sentiment label and single class label. Since social media is noisy, ambiguous and sparse, single label classification may not be able to capture the post classes accurately. Thus, addressing these two problems, this paper proposes a multi-task multi-label (MTML) classification model that performs classification of both sentiments and topics concurrently. It incorporates results of each task from prior steps to promote and reinforce the other iteratively. This paper basically worked upon tweets related to virgin mobiles. The analysis of tweet sentiments and topics can help businesses to get a sense of user opinion towards their products and services. Results show that MTML produces a much higher accuracy of both sentiment and topic classifications.



[redacted] @ciordonez 9 Oct
Great cell phone deals in #Chile right now with @VirginMobile_cl got a great deal!
Expand

[redacted] @TheLisesPieces 9 Oct
Any recommendations for a decent mobile service provider? #virginmobile is just the absolute worst
Expand

[redacted] @NatsDinsdale 9 Oct
#virginmobile Australia - sort of redeemed yourself - but god it was hard work getting there!!!!
Expand

[redacted] @SaulC23 9 Oct
If you ever wake up one morning and fancy being on hold for 2 hour listening to bieber over&over call virginmobile customer service #suicide
Expand

[redacted] @Marcelini20 9 Oct
I SWEAR everytime I go online to #virginmobile to pay my bill is ALWAYS under a "little daily maintenance"#HELP please! Am serious
Expand

Figure 1. Tweets related to "virgin mobile", with topic and sentiment labels.

Figure2. Research Paper 5: Tweet Topic and Sentiment Labels

2.1.2 Integrated Summary

- The figures below show the result of the case study of “Fans Reacting to an Underdog Comeback” which captured that effect by analyzing the change in attitudes (Mocking, Hope and Happiness) distributions over twenty minute intervals.

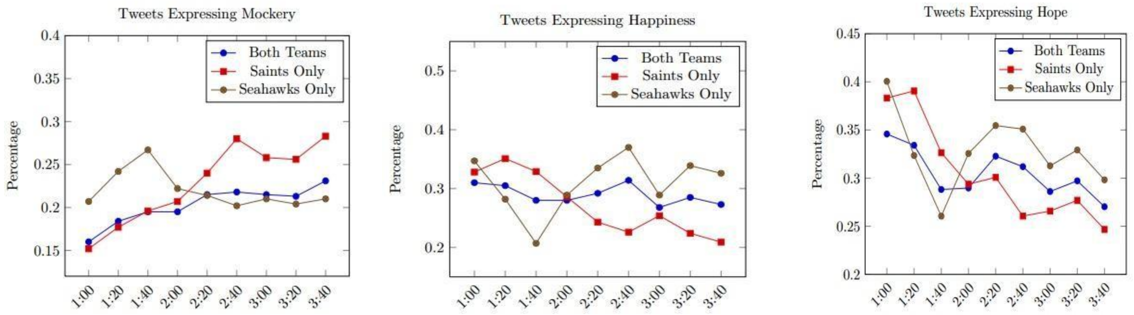


Figure 3. Integrated Summary: Fans Reacting to Underdog Comeback

- The figure below shows the classification of tweets into 18 general category followed by figures depicting the accuracy of the classifications.

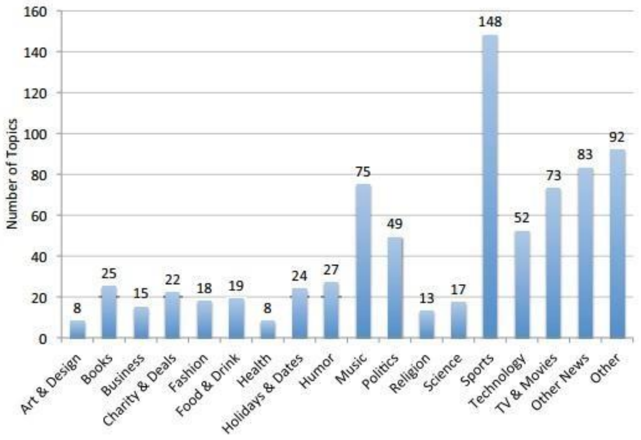


Figure 4. Integrated Summary: Tweets Classification under 18 Categories

- The Figure 5 shows that the MTML model performs better than baseline methods, i.e. Naïve Bayes, Support Vector Machine, Maximum Entropy, EM with Prior on Maximum Entropy, on both sentiment and topic classification.

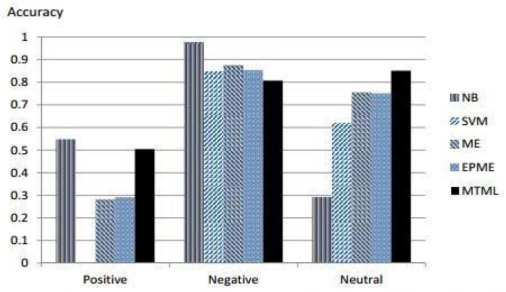


Figure 9. The accuracy of sentiment classification of the MTML model per three sentiment classes

Figure 5. Integrated Summary: MTML model

- Tables below shows 4 sample tweets with their sentiment and topic classification labels. Classification results by NB, SVM, ME and our MTML model.

➢ Tweet #1 has topic Complaint and sentiment Negative.

➢ Tweet #2 has topic Compliment and sentiment Positive.

➢ Tweet #3 has topic Promotion and sentiment Positive.

➢ Tweet #4 has topic Mention and sentiment Neutral.

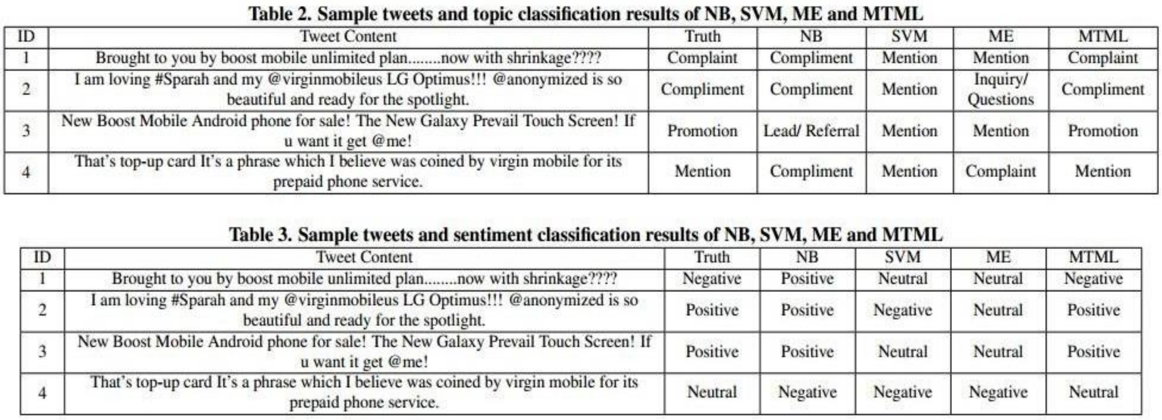


Table 2. Sample tweets and topic classification results of NB, SVM, ME and MTML

ID	Tweet Content	Truth	NB	SVM	ME	MTML
1	Brought to you by boost mobile unlimited plan........now with shrinkage????	Complaint	Compliment	Mention	Mention	Complaint
2	I am loving #Sparah and my @virginmobileus LG Optimus!!! @anonymized is so beautiful and ready for the spotlight.	Compliment	Compliment	Mention	Inquiry/ Questions	Compliment
3	New Boost Mobile Android phone for sale! The New Galaxy Prevail Touch Screen! If u want it get @me!	Promotion	Lead/ Referral	Mention	Mention	Promotion
4	That's top-up card It's a phrase which I believe was coined by virgin mobile for its prepaid phone service.	Mention	Compliment	Mention	Complaint	Mention

Table 3. Sample tweets and sentiment classification results of NB, SVM, ME and MTML

ID	Tweet Content	Truth	NB	SVM	ME	MTML
1	Brought to you by boost mobile unlimited plan........now with shrinkage????	Negative	Positive	Neutral	Neutral	Negative
2	I am loving #Sparah and my @virginmobileus LG Optimus!!! @anonymized is so beautiful and ready for the spotlight.	Positive	Positive	Negative	Neutral	Positive
3	New Boost Mobile Android phone for sale! The New Galaxy Prevail Touch Screen! If u want it get @me!	Positive	Positive	Neutral	Neutral	Positive
4	That's top-up card It's a phrase which I believe was coined by virgin mobile for its prepaid phone service.	Neutral	Negative	Negative	Negative	Neutral

Figure 6. Integrated Summary: Sample Tweets

2.2 Summary of field survey, experimental studies, new tools

The above field survey helped to develop a basic understanding of the utility of micro-blogging, in particular Twitter. Twitter gives this fascinating ability to understand people in context like we've never been able to do before.

Twitter is chosen because it is predominantly suited for data mining because of three key features.

- Twitter's API is well designed and easy to access.
- Twitter data is in a convenient format for analysis.
- Twitter's terms of use for the data are relatively liberal as compared to other APIs. It is in general acceptable that tweets are public and reachable to anyone.

This filed survey also helped with the process of setting up a development environment with Java and Twitter's JAVA based API (Twitter4J), and explore the following:

- Twitter's developer platform and how to make API requests.
- Extracting entities such as user mentions, hash tags from tweets.
- Techniques for performing classification of data on the basis of predefined classes.

Twitter4J is an unofficial Java library for the Twitter API. With Twitter4J, you can easily integrate your Java application with the Twitter service. Twitter4J is an unofficial library.

CHAPTER 3

ANALYSIS, DESIGN AND MODELING

3.1 Requirements Specifications

3.1.1 Functional Requirements

- *RequirementIDR1.01*

  Title: User Authorization

  Description: It is required to be able to connect our program with twitter by proper authorization.

  *Priority:*1

- *RequirementIDR1.02*

  Title: Twitter Data Extraction

  Description: It is required to extract the required data from the Twitter through its API Twitter4J.

  Priority:1

- *RequirementIDR1.03*

  Title: Graphical User Interface Development

  Description: It is required to provide users with a user friendly environment.

  *Priority:*2

- *RequirementIDR1.04*

  Title: Classification under Multiple Labels

  Description: It is required classify the extracted tweets under various classes based on the data set obtained.

  *Priority:*1

- *RequirementIDR1.05*

  Title: Comparison of two products under Consumer Module

  Description: It is required compare the extracted tweets post classification using weightage model to produce a desired result.

  *Priority:*1

- *RequirementIDR1.06*

  Title: Brand Image Description under Company Module

  Description: It is required provide a brand image of a company on the basis the extracted tweets post classification.

  *Priority:*1

3.1.2 Non-Functional Requirements

- *Software Requirements:*

Operating System:	Windows or any flavor of UNIX that supports Java.
Front End:	NetBeans IDE or any UNIX based IDE
Twitter API:	Twitter4J
JVM:	Java 5 or later

- *Performance Requirements:*
    - To be able to provide the user with a system with high rate of accuracy for classification of tweets.
    - To be able to provide the user with a system that can achieve the desired task in decent time frame.

- *Security Requirements:*
    - Only Public Tweets will be extracted through API
    - No access to any user's Personal Details other than Geographical Location (if required)
    - Data extracted from Twitter should not be modified or shared.

- *Safety Requirements:*
    - Authorization shall ensure that Application does not cause any harm to either the user or the system.

- *Reliability Requirements:*
    - The application must be able to extract and classify the most suitable Tweets related to the search words.
    - The classification and comparison results to be reliable enough so that user can make a decision based on them

- *Maintenance Requirements:*
    - The Application requires zero cost of maintenance. However, with time and then, a modification in algorithms should always be welcomed.

3.2 Overall Architecture

Phase 1: Twitter4j Authorization and Tweets Extraction

Twitter API provides interfaces to query Twitter for data based on certain filters.

Twitter provides three kinds of APIs:

- Search API
- REST API
- Streaming API

This Project undertakes the search API of twitter under twitter4j.

Phase 2: Training Phase

Training the classifier with the help of hand classified data set. In this project some, 500 tweets are manually classified to prepare training data set

Phase 3: Classification Phase

Classification with the help of in built JAVA library LMClassifier, classifying the tweets in respective topics and sentiments.

Phase 4: Generating the Desired Results for both the modules,i.e.,consumer and company.

3.3 Design Documentation

3.3.1 Use Case Diagram

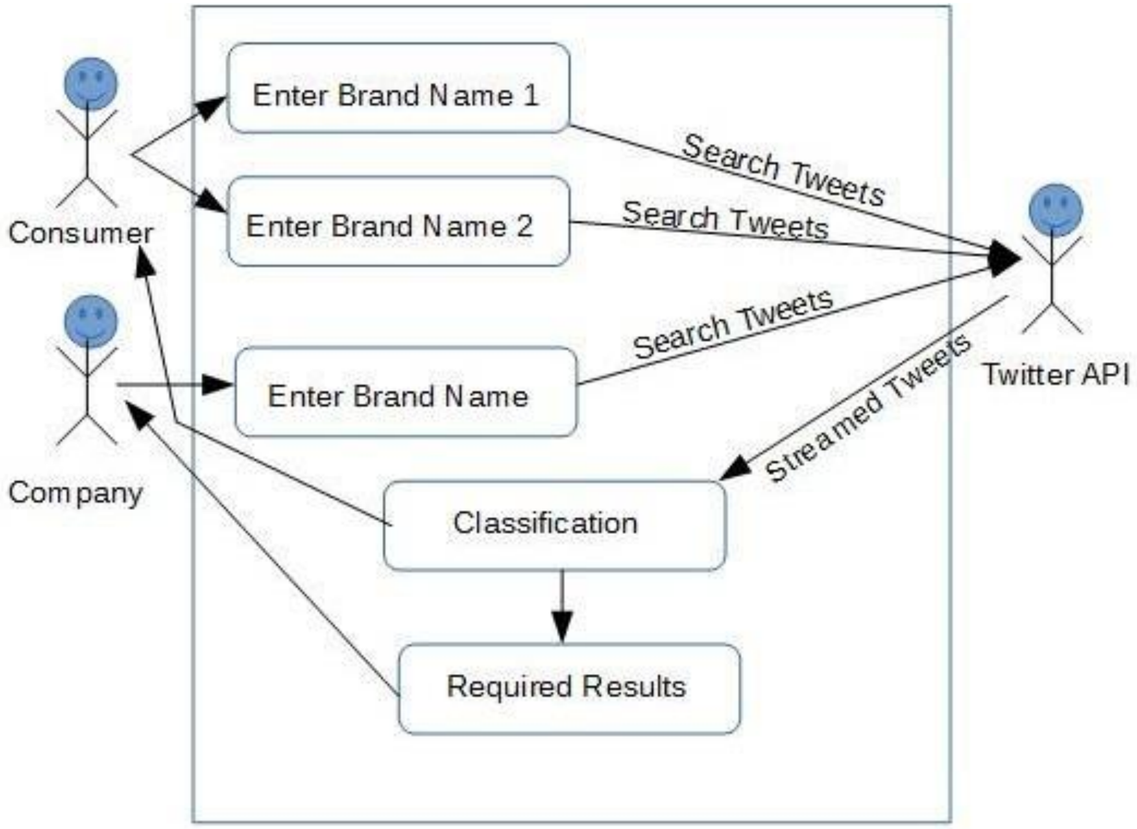


Figure 7. Use Case Diagram

3.3.2 Control Flow Diagram

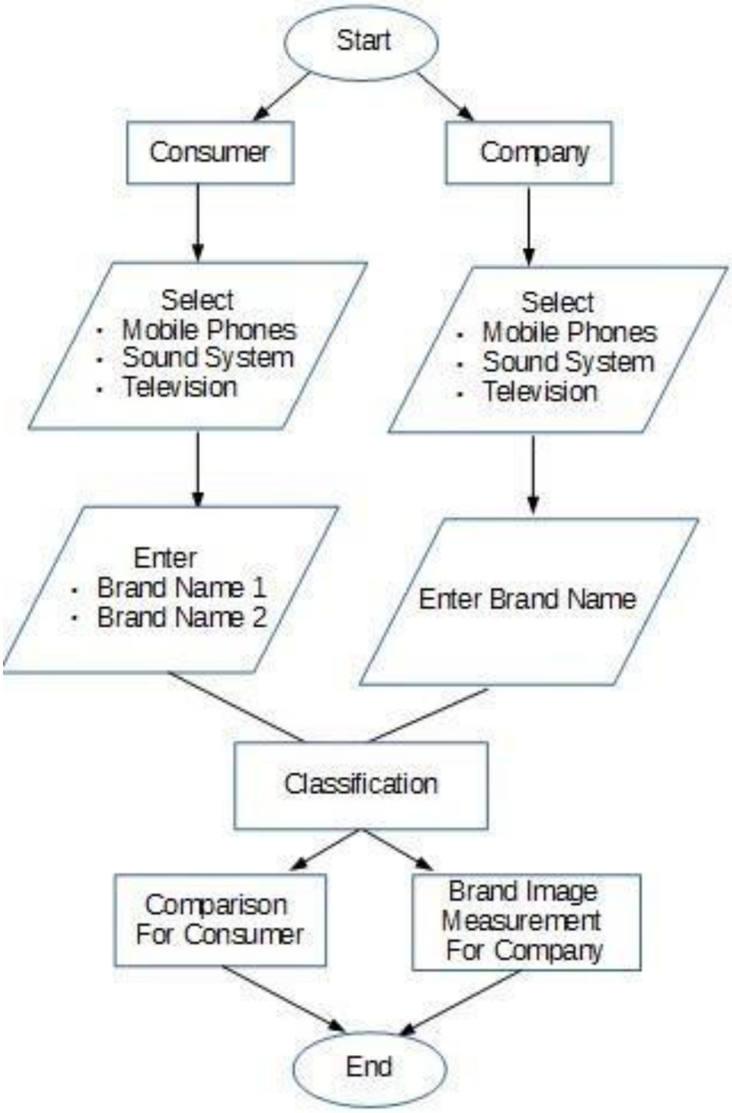


Figure 8. Control Flow Diagram

3.3.3 Sequence Diagram

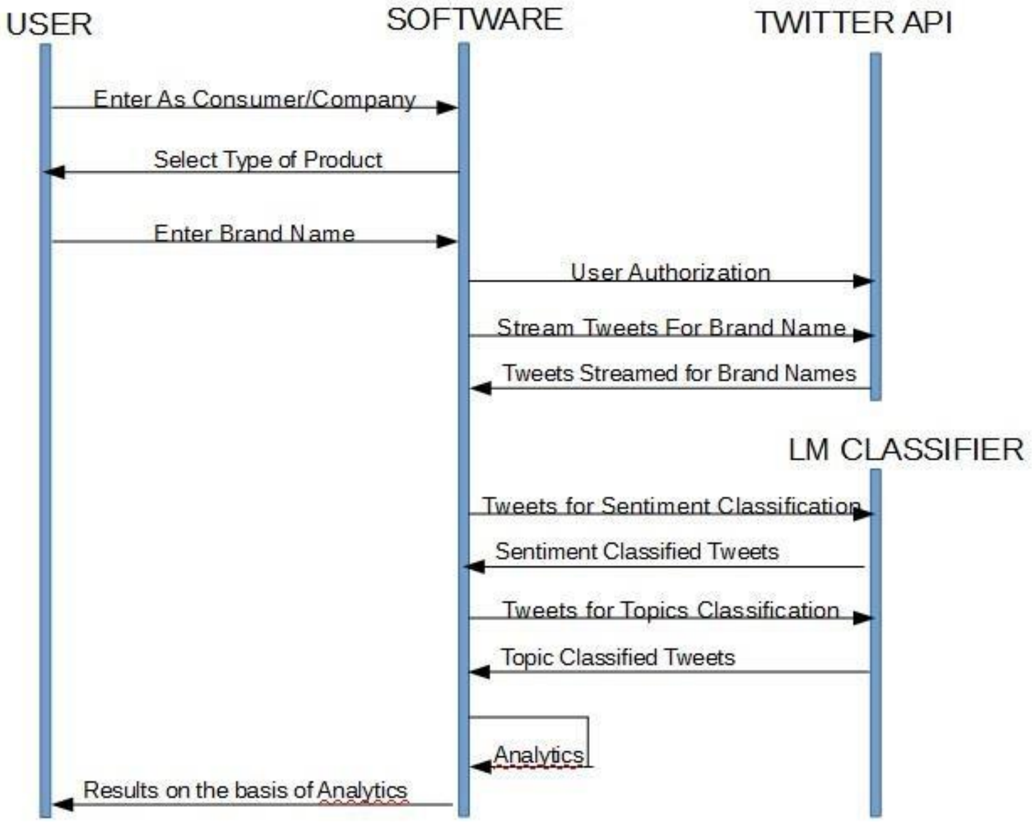

USER
SOFTWARE
TWITTER API
Enter As Consumer/Company
Select Type of Product
Enter Brand Name
User Authorization
Stream Tweets For Brand Name
Tweets Streamed for Brand Names
LM CLASSIFIER
Tweets for Sentiment Classification
Sentiment Classified Tweets
Tweets for Topics Classification
Topic Classified Tweets
Analytics
Results on the basis of Analytics

Figure 9. Sequence Diagram

3.4 Risk Analysis and Mitigation Plan

Software project risk management is crucial for the software development projects. It is used for project planning and control purposes during the project execution. Risk management will help reduce project failure. To mange software project risks, the first step is to identify a list of risks for checklist.

Risk Id	*Description*	*Risk Area*	*Probability (p)*	*Impact (I)*	*RE (p*I)*	*Risk Selected for Mitigation*	*Mitigation Plan*
1.	Unrealistic Time Estimates	P&C	2	4	8	Yes	Work before time estimates
2.	Lack of Data Mining project experience	T	2	3	6	Yes	Practice on new tools
3.	Ambiguous Tweet Classification	Comp	1	3	3	No	
4.	System Failure	H	1	5	5	Yes	Regular Interval Backups

Table 1: Risk Analysis and Mitigation Plan

The software risks comprises of 4 types of Risk Area:

Project Complexity (Comp),
Planning and Control (P&C),
Team (T)
Hardware (H)

CHAPTER-4

IMPLEMENTATION AND TESTING

4.1 Implementation details and Issues

In this project, Sentiment and Topic Classifications are implemented using an in build JAVA class “LMClassifier”. An LMClassifier performs joint probability-based classification of character sequences into non-overlapping categories based on language models for each category and a multivariate distribution over categories. The conditional probability estimate of the category given the input is derived from the joint probability of category and input: P(category input) = P(category, input) / P(input) where the joint probability P(category, input) is determined by the joint probability estimate and the input probability P(input) is estimated by marginalization:

$$P(\text{input}) = \Sigma_{\text{category}} P(\text{category, input})$$

This classification is done with the help of some 1000 hand-classified tweets. The topics are defined as folder name in classification folder.

Also, the comparison between the products is carried out using Polarity of Negative, Positive and Neutral tweets using the formula:

```
if(#Positive_Tweets==0 &&  #Negative_Tweets==0)
        Polarity(topic) = #Neutral_Tweets
else if(#Positive_Tweets > #Negative_Tweets && #Negative_Tweets!=0)
        Polarity(topic) = #Positive_Tweets/(#Neutral_Tweets  + #Negative_Tweets)
else if(#Positive_Tweets > #Negative_Tweets && #Negative_Tweets==0)
        Polarity(topic) = #Positive_Tweets
else if(#Positive_Tweets < #Negative_Tweets && #Positive_Tweets!=0)
        Polarity(topic) = -1 * (#Negative_Tweets/(#Neutral_Tweets  + #Positive_Tweets))
else if(#Positive_Tweets < #Negative_Tweets && #Positive_Tweets==0)
        Polarity(topic) =-1* #Negative_Tweets
```

The net polarity is calculated as the mean of above polarities for different topics under which brand is classified.

$$\text{Net\_Polarity\_of\_Topic}=\frac{\sum \text{Polarity(Topic)}}{\#\text{Total\_Topics}}$$

Also there is a bar graph option showing the variation in sentiments of people over pat few weeks which is developed with the help of unstructured data stored in files.

4.2 Testing

Software testing is an investigation conducted to provide stakeholders with information about the quality of the product or service under test.[1] Software testing can also provide an objective, independent view of the software to allow the business to appreciate and understand the risks of software implementation. Test techniques include the process of executing a program or application with the intent of finding software bugs (errors or other defects).

Software testing involves the execution of a software component or system component to evaluate one or more properties of interest. In general, these properties indicate the extent to which the component or system under test:

- meets the requirements that guided its design and development,
- responds correctly to all kinds of inputs,
- performs its functions within an acceptable time,
- is sufficiently usable,
- can be installed and run in its intended environments, and
- achieves the general result its stakeholders desire.

4.2.1 Testing Plan

TYPE OF TEST	**WILL THE TEST BE PERFORMED?**	**COMMENTS / EXPLANATIONS**	**SOFTWARE COMPONENT**
Requirements Testing	Yes	It is required to test the inclusion of required JAR files	Libraries
Unit Testing	Yes	Unit Testing on Two Different Modules: Consumer & Company Also, Unit Testing on Classification of Different Types of	Mobile Phones, Sound System, Television unit modules under Consumer Unit Module and Company Unit Module

TYPE OF TEST	WILL THE TEST BE PERFORMED?	COMMENTS / EXPLANATIONS	SOFTWARE COMPONENT
		Products Mobile Phones, Sound System and Television	
Integration Testing	Yes	Proper Redirection from One Module to Another on click of Buttons	Complete Software
Performance Testing	Yes	Accuracy of classification	LMClassifier Class and Training Data set
Stress Testing	No	Not Applicable	
Security Testing	Yes	Security Test for User Authorization	Twitter4J User Authorization
Volume Testing	Yes	Number of tweets that can be extracted.	Tweet Extraction

Table 2: Testing Plan

TEST ENVIRONMENT

SOFTWARE ITEMS

Operating System:	Windows 8
Front End:	NetBeans IDE
Twitter API:	Twitter4J
JVM:	Java 5

HARDWARE ITEMS

Laptop

4.2.2 Component Decomposition and Type of Testing required

SNO.	**COMPONENTS (MODULE) THAT REQUIRE TESTING**	**TYPE OF TESTNG REQUIRED**	**TECHNIQUE FOR WRITING TEST CASES**
1.	Tweet Sentiment Classification	Unit	Black Box: Equivalence Classes Testing
2.	Tweet Topic Classification	Unit	Black Box: Equivalence Classes Testing
3.	Polarity Calculation for Comparison	Unit	White Box: Decision Testing
4.	Large Number of Tweets Extraction	Volume	Black Box: Robustness
5.	GUI Testing	Integration	I/O based
6.	Twitter4J Authorization	Security	I/O based
7.	Libraries	Requirement	Testing if JAR files are included or not

Table 3: Component Decomposition and Type of Testing Required

4.2.3 Test Cases

4.2.3.1 Tweet Sentiment Classification Test Cases

TEST CASE ID.	INPUT	EXPECTED OUTPUT	PASS / FAIL
1.	Horrible experience with motog3..display probl..wait for 2 months...new cracked display again..worst service centre	NEGATIVE	PASS
2.	I added a video to a @YouTube playlist	NEUTRAL	PASS
3.	Get top quality listening experience with SKULLCANDY	POSITIVE	PASS

Table 4: Tweet Topic Classification Test Cases

4.2.3.2 Tweet Topic Classification Test Cases

TEST CASE ID.	INPUT	EXPECTED OUTPUT	PASS / FAIL
1.	#Fly captured in my #MotoG3	CAMERA	PASS
2.	MotorolaIndia deposited motoG3 at srvice cntre for sim tray replacement more than a month ago. No progress #careless_attitude #BAD_SERVICE	SERVICE AND SUPPORT	PASS
3.	How to connect Sony BRAVIA TV to CPU [HDMI Cable connection]	INQUIRY	PASS
4.	Got my earphones in an exchange deal with friend #JBL	MENTION	FAIL o/p: Deals
5.	Get top quality listening experience with SKULLCANDY	FEATURES	PASS

Table 5: Tweet Topic Classification Test Cases

4.2.3.3 Polarity Calculation for Comparison

TEST CASE ID.	INPUT	EXPECTED OUTPUT	PASS / FAIL
1.	#Positive_Tweets= 58 #Negative_Tweets= 45 #Neutral_Tweets=30	Net_Polarity= .77	PASS
2.	#Positive_Tweets= 58 #Negative_Tweets= 0 #Neutral_Tweets=30	Net_Polarity=.116	PASS
3.	#Positive_Tweets= 8 #Negative_Tweets= 85 #Neutral_Tweets=125	Net_Polarity=.639	PASS

Table 6: Polarity Calculation Test Cases

CHAPTER-5

FINDINGS & CONCLUSIONS

5.1 Findings

- Our findings show that there are more positive than negative sentiment polarity tweets.
- Positive sentiment polarity tweets spread more than negative tweets in the social sub-network
  of Twitter.
- Most people tweet about the product if they enjoy the product they use and avoid going to products which they don't like by reading reviews and getting recommendations from friends.
- Tweets extraction and classification proved as a useful tool for making a comparison tool on real life experiences of people.
- It proved to be an attractive tool for companies to help them understand their Brand Image.
- The justification of using Twitter as our data field relies on the existence of a social sub-network for ordinary users. Twitter has social network aspects in addition to mass medium characteristics.

5.2 Conclusion

Microblogging nowadays became one of the major types of the communication. It is identified as online word-of-mouth branding.
The large amount of information contained in microblogging web-sites makes them an attractive source of data for opinion mining and sentiment analysis. There are several implications of this project.

These tweets proved as a uselful tool for making a comparison tool on real life experiences of people. Also, it proved to be an attractive tool for companies to help them understand their brand image. Microblogging can be used to provide information and draw potential customers to other online media, such as Websites and blogs. As such, monitoring and leveraging micro-blogging sites concerning one's own brand and the brand of competitors provides valuable competitive intelligent information. Companies can receive positive brand exposure via followers and others who microblog about the company and products. With micro-blog monitoring tools, companies can track micro-blog postings and immediately intervene with unsatisfied customers. Companies can get near realtime feedback, by setting up corporate accounts, from customers using micro-blog polls, and surveys. Organizations can get valuable contents and product improvement ideas by tracking micro-blog postings and following those people who follow their corporate accounts. Companies can leverage contacts made via micro-blogging service to further their branding efforts by responding to comments about the company brand.

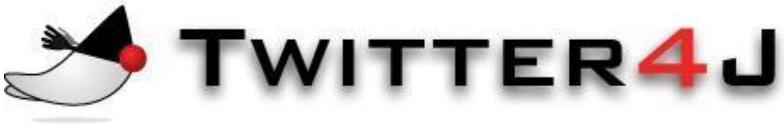
5.3 Future Work

Many research questions remain open for future work:

- Explore other classification tasks and extend our model to incorporate more related tasks.
- While the performance of our model is promising and superior to competing methods, investigate other ideas to further improve the accuracy. For instance, a hybrid approach is promising.
- Aim to add costs to the classifications. For instance, not detecting topic labels should be heavily punished compared to misclassifying.
- Improve the overall performance of our approach by taking different features and a larger training set into account.
- Monitoring and Leveraging micro-blogging sites concerning one’s own brand and the brand of competitors provides valuable competitive intelligent information.
- Companies can track micro-blog postings and immediately intervene with unsatisfied customers.
- Companies can get near real-time feedback, by setting up corporate accounts, from customers using micro-blog polls, and surveys.

APPENDIX

A.DETAILS OF NEW TECHNOLOGY



❖ *Introduction*

Twitter4J is an unofficial Java library for the Twitter API.
With Twitter4J, you can easily integrate your Java application with the Twitter service. Twitter4J is an unofficial library.
Twitter4J is featuring:

✓100% Pure Java - works on any Java Platform version 5 or later

✓Android platform and Google App Engine ready

✓Zero dependency : No additional jars required

✓Built-in OAuth support

✓Out-of-the-box gzip support

✓100% Twitter API 1.1 compatible

❖ *System Requirements*

OS: Windows or any flavor of Unix that supports Java.
JVM: Java 5 or later

❖ *Code Examples*

Sample codes are located at src/twitter4j/examples/ and one can run each classs using bin/*className*.cmd|sh.
To run the example codes, it is needed to have OAuth credentials configured in twitter4j.properties.

B.REFERENCES

- *Evaluating Multi-label Classification of Incident-related Tweets*

Axel Schulz, Eneldo Loza Mencía, Thanh Tung Dang, Benedikt Schmidt
#Microposts2014,4th Workshop on Making Sense of Microposts

- *Micro-blogging as Online Word of Mouth Branding*

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA. ACM 978-1-60558-247-4/09/04

- *Sentiment and Topic Analysis on Social Media:A Multi-Task Multi-Label Classification Approach*

Shu Huang, Wei Peng, Jingxuan Li, Dongwon Lee
WebSci'13 May 1 – May 5, 2013

- *SeNTU:Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning*

PrernaChikersal,SoujanyaPoria,andErikCambria School of Computer Engineering Nanyang Technological University Singapore – 639798

- *Supervised Multi-Class Classification of Tweets*

Zahan Malkani , Evelyn Gillie, Stanford University,2012

- *Twitter Trending Topic Classification*

Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary
11th IEEE International Conference on Data Mining Workshops

- *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*

Alexander Pak, Patrick Paroubek
Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508,F-91405 Orsay Cedex, France,2010