Chittagong Universioty of Engineering & Technology
Department of Computer Science & Engineering
Chittagong-4349

(Project/Thesis Proposal)

Application for the approval of B. Sc. Engineering Project/Thesis
(Computer Science & Engineering)

Date: 23-9-2019

1.	**Name of the Student**	: Simon Islam
	Roll No.	: 1504062
	Session	: 2018-2019
2.	**Present Address**	: Room No: 110 Shahid Tarek Huda Hall Chittagong University of Engineering & Technology Chittagong-4349
3.	**Name of the Supervisor**	: Animesh Chandra Roy
	Designation	: Assistant Professor Department of Computer Science & Engineering Chittagong University of Engineering & Technology Chittagong-4349
4.	**Name of the Department**	: Computer Science & Engineering
	Program	: B.Sc. Engineering
5.	**Date of First Enrollment in the Program**	: 25 February, 2016
6.	**Tentative Title**	: Multi-Label Emotion Classification on Trending Tweets using Machine Learning

7. Introduction

Twitter is currently one of the biggest social network in the world. As of september, 2019, it has 330 million active users worldwide. It is famous for it's microblogging feature. Microblogging is the habit of making tiny posts to a microblog. It can range from various topics, pictures, gifs or sharing a link. Before november, 2017, a user can write a tweet consists of at most 140 characters. But since then, twitter has doubled this capacity to 280 characters. When comparing with other social media giants like facebook, the characteristics that sets twitter apart is it's reach. Twitter has a global reach. It is not limited to friends and family. Instead, it encompasses the poster and his follwers. In other social network sites like facebook, if someone wants to see someone's posts both has to accept each other as friends. But in twitter all someone has to do is to follow the other person. This one way follower relationship has given twitter a global reach.

In today's world, trends are changing on daily basis. Twitter detects this current trends by calculating the volume of a keyword over a period of time. The spike in the volume in a short time classify the keyword as trends. And the tweets having those keywords in that time are called trending tweet. Due to twitter having this global reach & vast number of users, it has become a global platform to express peoples emotion online. From average day people to superstars, everyone use twitter to express their emotions on various topics. In return, their followers gives a feedback of his own emotions. To understands what kind of sentiments are they creating around the world, it is important to be able to extract emotions from their tweets. It can help researchers to understand how people are feeling about a certain topic. During the recent age of social media, sentiment analysis has been one the most studied field by the researchers. It is the process of identifying and categorizing sentiments over a piece of text. It extracts valuable features from the texts. Sometimes confused with opinion mining, sentiment analysis deals with sentiments of the text where in opinion mining extract opinion.

In this work, we aim to perform multi-label emotion analysis of tweets posted in twitter. We wish to classify the tweets in one of 8 emotion classes. They are: sadness, disgust, joy, interest, admiration, anger, surprise & fear. We will collect trending tweets from twitter and will try to label them with emotions that describe them best. It is a multi-label classification because more than one emotion may be needed to properly clssify a tweet. Although many work has been done on sentiment analysis, very few of them attempted to classify them in multiple labels. Our main goal in this work is to properly label them accoring to their emotional sentiments.

8. Background and Present State of the Problem

In this current age of social media, people has began to discuss their thoughts, daily lives and works on social media. They comment on various happenings around the world and how it is affecting their lives or the lives of other people. This huge dependency of peole on social media has attracted researchers to study the behaviors. Twitter, due to it's global reach and accessibility has attracted many researchers. Some of the researches include the use of slangs, emoticons and how they developed over a period of time.[1, 2] Most of the sentiment analysis done on twitter has been binary classification. In binary

classification, the tweets are marked from negative to positive. Some classification also contains a neutral class. They include polarity of products[3], movies[4] and democratic election.[5] There are also some advanced work that went deeper into the classification and tried to asses the strength of the sentiment. Their range included from very negative to very positive, in some cases giving scores depending on the sentiment intensity.[6, 7]
In the recent years, the works on multi-label classification has started to gain some traction. Instead of labelling a tweet positive, negative or neutral, a multi-class classification labels it according to the emotions found in the text. Bhowmick, Basu et al.[8] used multilabel sentiment classification on news extracted from Times of India newspaper. Liu et al.[9] used a multi-label classification approach to perform multi-label classification of microblogs. They performed this on gatherd data. Bouazizi et al.[10] used 3 pairs of opposite sentiment to label their tweets. They didn't consider statements of opposite polarity. We wish to perform multi-label emotion classification on real time trending tweets on twitter.

9. Aims with Specific Objectives & Possible Outcomes

The main objectives & the possible outcomes of this work are mentioned in the following:

- To extract real time trending tweets from twitter
- To perform Multi-Label Emotion Analysis of the tweets (sadness, disgust, joy, interest, admiration, anger, surprise & fear)

10. Outline of Methodology

The key objective of our work is to develop a system that can process the tweets and label them according to their emotions. We first need to gather a dataset of tweets. Twitter provides their own api that can collect tweets from twitter. Then we need to clean the dataset. We do this by removing links and pictures from the tweets. Then we separate the senteces, hashtags, emoticons from the tweets. From the senteces, we will detect phrases, expand emoticons, process elongations. We also need to detect negation in a sentence or does it represent two different emotions with conjunctions. For this, we need to consider each and every tweet individually and preprocess it before sending it to train our model. Another important step before the training is feature extraction. When all the steps are finally complete, we use our data to train multi-label sentiment classification model. Finally, the model is tested using the test cases that we create.And it will be evaluated and deployed in a real time enviroment, where it will detect emotions in currently trending tweets. A brief overview of our system is given below:

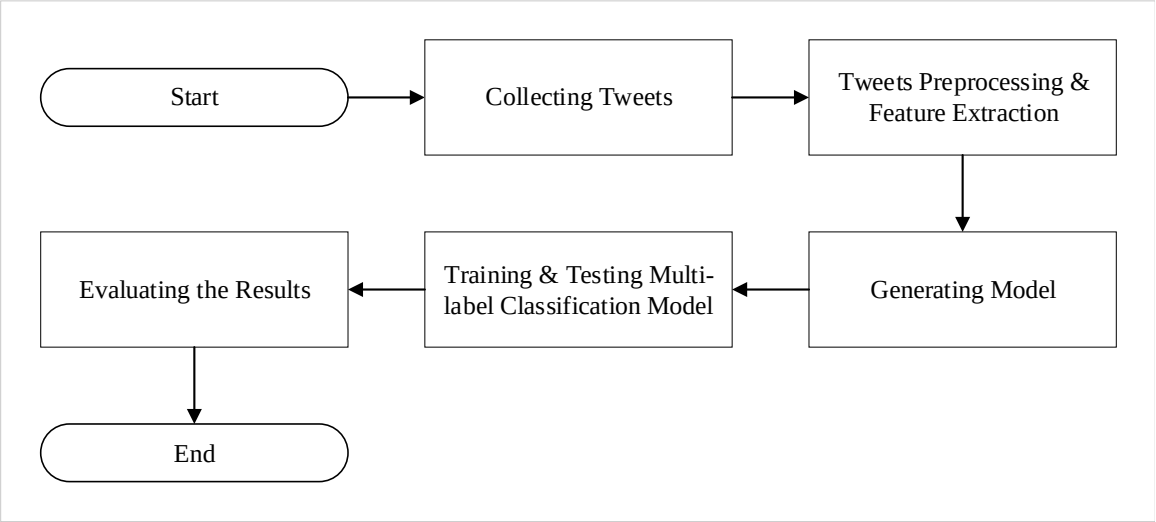

Figure 1: Overview of the work

Collecting Tweets

Twitter has it's own API that can be used to collect tweets from twitter. It can be used to collect data in real time or from a selected user or regarding a specific topic. We will use it to collect tweets in real time.

Preprocessing Tweets

To be able to train our model, it is important to process our data properly. Unlike a sentence, a tweet is not a collection of words. People uses sentences, punctuations, phrases, acronyms, emoticons to properly convey their messages. Our job here is to process the text in a way to make it clean. And we will also extract features from other aspects of tweet. Data preprocessing phase is described as follows:

- Remove pictures, web links and e-mails
- Separate sentences, punctuations, hashtags & emoticons
- Process the hashtag to seperate it by words
- Count the number of question marks and exclamation mark after each word
- Expand the acronyms & parts of speech tagging
- Recognize intensifier word classes and negation words
- Remove stopwords, process elongation & perform word stemming and lemmatisation of word
- Map the emoticons according to their emotions

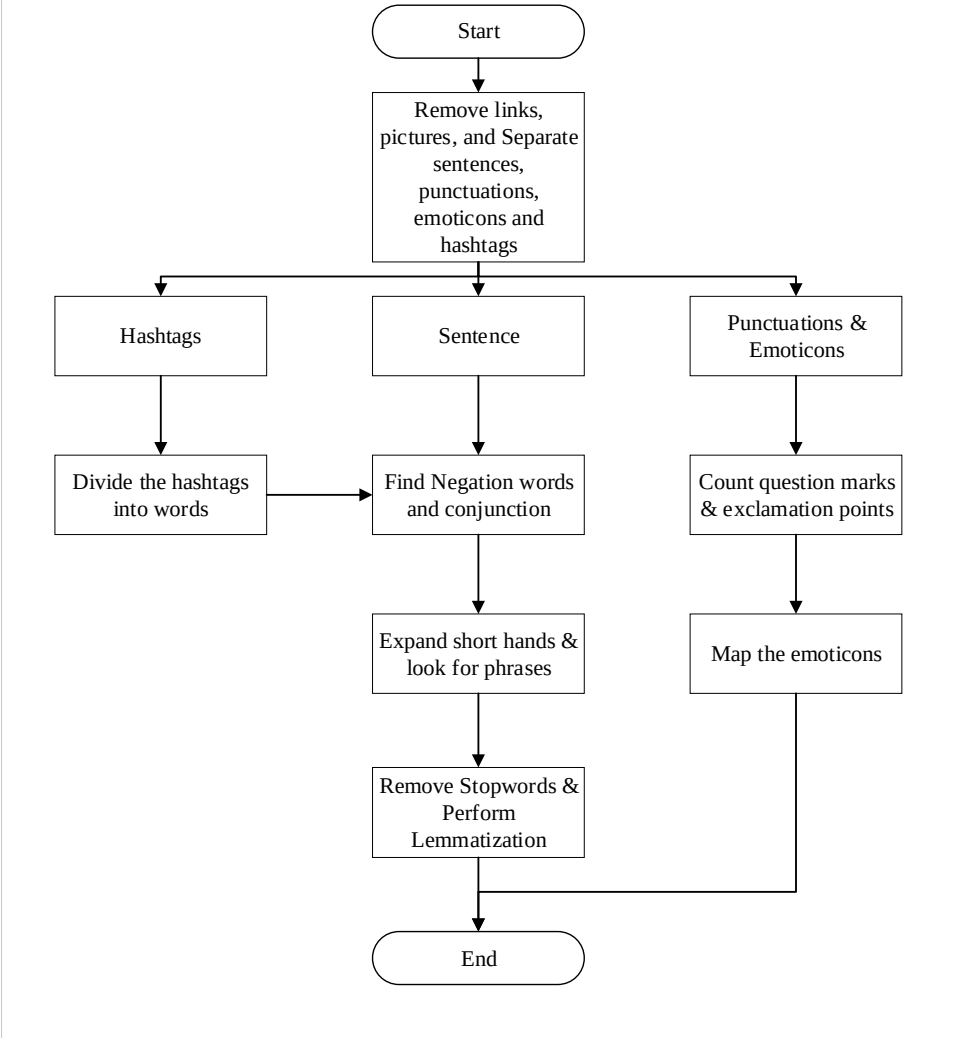

Figure 2: Tweet Preprocessing Flowchart

Training and Testing the Data

There are two ways to approach sentiment analysis. One is lexicon based and the other is machine learning based. In the our work, we wish to perform multi-label emotion classification using machine learning. There are many machine learning algorithms that are suited for this work. Some of them being Naive Bayes, SVM, Maximum Entropy, Long-short term memory etc. In this step we label the data and train our model. For the labelling purpose we would like to use established datasets. But if such datasets are not found then we will propose our own labelling method to perform multi-level sentiment analysis. For that case, we will use various sentiment libraries like SenticNet [11] and others. Then finally we test the model against our testing data.

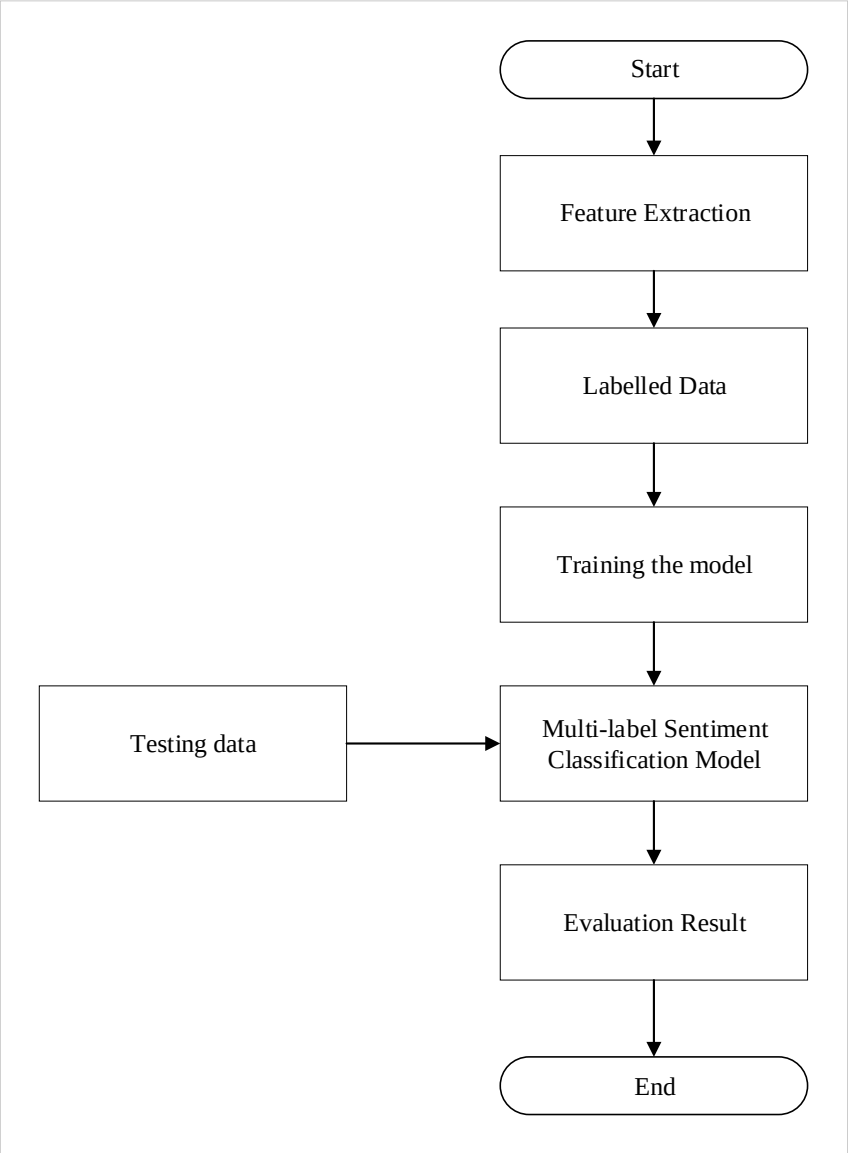

Figure 3: Multi-label Sentiment Classification Model

11. Resources Required to Complete the Work

The resources required to complete the work can be divided into two categories. They are hardwares and software. The requirements are illustrated below:

- **Hardware Components**
  - Personal Computer
- **Software Components**
  - Python 3.7
  - PyCharm IDE for Python

12. Cost Estimate

The cost that will occor to complete our proposed work is given below:

Cost of Materials:

A general laptop	Tk.	50000
Paper	Tk.	500
Total	**Tk.**	**50500**

Typing, Drafting, Binding:

Internet Browsing & Typing	Tk.	2500
Drafting	Tk.	500
Binding	Tk.	500
Total	**Tk.**	**3500**
Grand Total	**Tk.**	**54500**

References

[1] M. Boia, B. Faltings, C.-C. Musat, and P. Pu, "A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets," in *2013 International Conference on Social Computing*. IEEE, 2013, pp. 345–350.

[2] K. Manuel, K. V. Indukuri, and P. R. Krishna, "Analyzing internet slang for sentiment mining," in *2010 Second Vaagdevi International Conference on Information Technology for Real World Problems*. IEEE, 2010, pp. 9–11.

[3] M. A. Cabanlit and K. J. Espinosa, "Optimizing n-gram based text feature selection in sentiment analysis for commercial products in twitter through polarity lexicons," in *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications*. IEEE, 2014, pp. 94–97.

[4] U. R. Hodeghatta, "Sentiment analysis of hollywood movies on twitter," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 1401–1404.

[5] J. M. Soler, F. Cuartero, and M. Roblizo, "Twitter as a tool for predicting elections results," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2012, pp. 1194–1200.

[6] Y. Priyadarshana, K. Gunathunga, K. N. N. Perera, L. Ranathunga, P. Karunaratne, and T. Thanthriwatta, "Sentiment analysis: Measuring sentiment strength of call centre conversations," in *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 2015, pp. 1–9.

[7] R. Srivastava and M. P. S. Bhatia, "Quantifying modified opinion strength: A fuzzy inference system for sentiment analysis," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2013, pp. 1512–1519.

[8] P. K. Bhowmick, A. Basu, and P. Mitra, "Reader perspective emotion analysis in text through ensemble based multi-label classification framework." *Computer and Information Science*, vol. 2, no. 4, pp. 64–74, 2009.

[9] S. M. Liu and J.-H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1083–1093, 2015.

[10] M. Bouazizi and T. Ohtsuki, "Multi-class sentiment analysis on twitter: Classification performance and challenges," *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 181–194, 2019.

[11] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

13. CSE Undergraduate Studies (CUGS) Committee Reference

Meeting No.: **Resolution No.:** **Date:**

14. Number of Undergraduate Students working with the Supervisor at Present: 2

_____

Signature of the Student

_____

Signature of the Supervisor

_____

Signature of the Head of the Department