Sentiment Analysis of Twitter Data

Kiruthika M.
Department of Computer Engineering
Fr. C. Rodrigues Institute of Technology, Mumbai University, Vashi, Maharashtra, India

Sanjana Woonna
Department of Computer Engineering
Fr. C. Rodrigues Institute of Technology, Mumbai University, Vashi, Maharashtra, India

Priyanka Giri
Department of Computer Engineering
Fr. C. Rodrigues Institute of Technology, Mumbai University, Vashi, Maharashtra, India

Abstract- "What other people think" has always been an important fact for most of us during the decision-making process. Customers post their experiences and opinion about the various products and services. But, due to massive volume of reviews, customers can't read all reviews. In order to solve this problem, a lot of research is being carried out in Sentiment Analysis/Opinion Mining. Sentiment Analysis is an approach to classify the sentiments of user reviews, documents etc. in terms of positive (good), negative (bad) or neutral (surprise). However, most of the sentiment analysis approaches today provide an overall polarity of the text. But it is desirable to understand the sentiment of each aspect of different entities for deep grained analysis. Hence, we propose a system that would analyze tweets about movies into three categories, which are positive, negative and neutral using supervised learning approach. This paper discusses the problems related to sentiment analysis, literature survey, proposed system, scope, existing systems and workflow of the proposed system.

Keywords – sentiment, review, aspect, hashtag, entity, emoticon

I.INTRODUCTION

Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. It involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets.

Few fields of research are predominant in Sentiment analysis:

- Sentiment classification: It deals with classifying entire documents according to the opinions towards certain objects.
- Feature based Sentiment classification: Feature -based Sentiment classification on the other hand considers the opinions on features of certain objects
- Opinion summarization: Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions.

The rest of the paper is organized as follows. In section II, we discuss different levels of sentiment analysis. In section III, we present the challenges of sentiment analysis. In section IV, we discuss tasks of sentiment analysis on micro-blog data. In section V, we discuss literature survey which includes three papers. We explain our proposed system which includes the information about twitter data and aspect sentiment classification in section VI respectively. In section VII we present the design of our system. In section VIII, we give implementation details of system along with the details about the data collection and data preprocessing. In section IX we present our graphical user interfac's screenshots. We conclude and give future directions of research in section X.

II. Different Levels of Analysis

There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA.

- Document Level: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment.

- Sentence Level: The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. The first step is to identify whether the sentence is subjective or objective.

- Aspect / Entity Level: Both the document level and the sentence level analysis do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion).

III.CHALLENGES IN SENTIMENT ANALYSIS

Sentiment words and phrases are important for sentiment analysis, only using them is far from sufficient. The problem is much more complex. In other words, we can say that sentiment lexicon is necessary but not sufficient for sentiment analysis. Below, we highlight several issues:

- A positive or negative sentiment word may have opposite orientations in different application domains. For example, “suck” usually indicates negative sentiment, e.g., “This movie sucks,” but it can also imply positive sentiment, e.g., “This vacuum cleaner really sucks.”

- A sentence containing sentiment words may not express any sentiment. For example, “Why wasn't @priyankachopra nominated for best actress for #BajiraoMastani this year for @filmfare @jiteshpillaai?”

- Sarcastic sentences with or without sentiment words are hard to deal , e.g., “Finally, ShikharDhawan's stay at the crease has been as short as #BajiraoMastani's at the box office!! #AUSvIND”

- Many sentences without sentiment words can also imply opinions. For example, “Have seen the movie some days back.. And the scenes are still so fresh in my mind.. What have you people done #BajiraoMastani.”

IV. SENTIMENT ANALYSIS TASK

Given a set of opinion documents *D*, sentiment analysis consists of the following four main tasks:

1) Task 1 (entity extraction and categorization):

Extract all entity expressions in D, and categorize or group synonymous entity expressions into entity clusters (or categories). Each entity expression cluster indicates a unique entity e_i .

2) Task 2 (aspect extraction and categorization):

Extract all aspect expressions of the entities, and categorize these aspect expressions into clusters. Each aspect expression cluster of entity e_i represents a unique aspect a_{ij}.

3) Task 3 (aspect sentiment classification):

Determine whether an opinion on an aspect a_{ij} is positive, negative or neutral, or assign a numeric sentiment rating to the aspect.

4) Task 4 (opinion generation):

Produce all opinion quintuples (e_i, a_{ij}, s_{ijkl}) expressed in document d based on the results of the above tasks**.**

V. LITERATURE SURVEY

In "Twitter Sentiment Analysis: The Good The Bad and The OMG! " paper, they have investigated the utility of linguistic features for detecting the sentiment of Twitter messages. They have evaluated the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in micro blogging. A supervised approach has been introduced to solve the problem, but leverage existing hashtags in the Twitter data for building training data.

The paper "Sentiment Analysis of Twitter Data" published in 2012 introduces a machine learning approach to implement sentiment analysis on data. They have performed sentiment classification of Twitter data where, the classes are "positive", "negative", and "neutral". Two kinds of models have been used: tree kernel and feature based models and both these models outperform the unigram baseline. For the feature-based approach, they performed feature analysis, which reveals that the most important features are those that combine the prior polarity of words and their parts-of-speech tags.

The paper "Twitter Sentiment Classification using Distant Supervision" published in 2009 introduces a novel approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative with respect to a query term. The paper describes the preprocessing steps needed in order to achieve high accuracy. The main contribution of this paper is the idea of using tweets with emoticons for distant supervised learning. Different machine learning classifiers and feature extractors have been used along with the usage of unigrams, bigrams, unigrams and bigrams, and parts of speech as features.

VI. PROPOSED SYSTEM

We have proposed a system that performs aspect level sentiment analysis on twitter data or tweets based on movies into three categories:

- Positive
- Negative
- Neutral

A) About Twitter

Twitter is a social networking and micro blogging service that allows users to post real time messages, called tweets. Tweets are short messages, restricted to 140 characters in length. Due to the nature of this micro blogging service (quick and short messages), people use acronyms, make spelling mistakes, use emoticons and other characters that express special meanings.
Following is a brief terminology associated with tweets.

- Emoticons: These are facial expressions pictorially represented using punctuation and letters which express the user's mood.
- Target: Users of Twitter use the "@" symbol to refer to other users on the micro blog which automatically alerts them.
- Hashtags: Users usually use hash tags to mark topics. This is done to increase the visibility of their tweets.

B) Aspect level sentiment classification

Sentence level or document level sentiment classification is insufficient in many applications as it only reflects the overall opinion and does not evaluate all the aspects of an entity. Hence, in order to understand the sentiment of each aspect, we perform aspect-level sentiment analysis or feature-based opinion mining. This paper, proposes to perform sentiment analysis of multiple aspects of various entities related to movies. For example:

"wowwwww awesome reviews on #BajiraoMastani omg cant wait too see going to watch it tomorrow."

In this particular tweet, "*#BajiraoMastani*" is the entity, "*reviews*" is the aspect and "*awesome*" is the sentiment. The opinion here is positive.

"#BajiraoMastani is a soulful artwork."

In this tweet, the movie is evaluated as a whole i.e. a GENERAL aspect of the entity is represented by using "*#BajiraoMastani* ". The opinion on the GENERAL aspect is positive.

VII. DESIGN

A) Flow Diagram

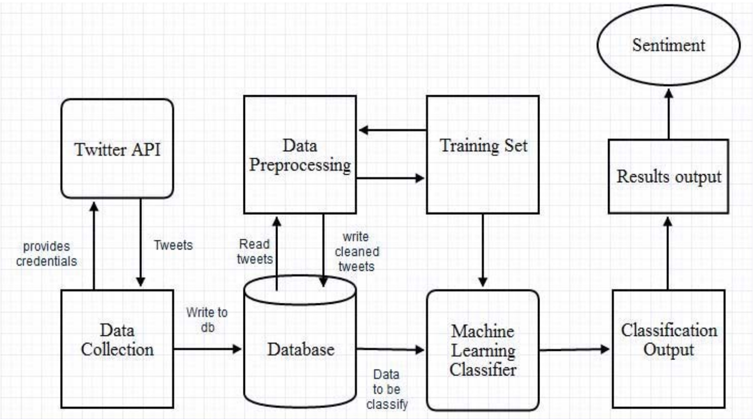


Figure 1. DWT Decomposition model

B) Our system has following steps

- Data Collection using Twitter API: Publically large sets of Twitter data is not available .Hence, we first extract the twitter data from the Twitter API.
- Data Preprocessing: This involves cleaning and simplifying the data by performing spell correction, punctuation handling, stemming etc. so as to remove noise from the data.
- Applying classification algorithms: The classification algorithms are applied on these tweets in order to categorize them . Different models provide different accuracy and we choose the model with the highest accuracy.
- Classified tweets: The result of the above step is classifies tweets which may belong to any of the three categories mentioned.
- Sentiments in graphical representation: The results of the sentiment analysis is provided using pie charts.

VIII. IMPLEMENTATION

A) Data Collection

Large datasets of tweets are publically unavailable, hence data can be extracted using the Twitter API. To access Twitter data, it is required to create an application on the developer site, which provides us with credentials. Using these, we can access the data by providing the search query and the number of tweets. We have followed of the above procedure for data collection.

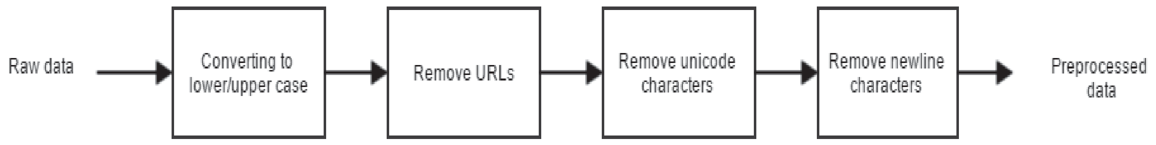
The data collection comprises of tweets about the following movies:

Name	Start date	End date	No of tweets
Bajirao Mastani	18/12/2015	10/01/2016	7008
Dilwale	18/12/2015	10/01/2016	3548
Wazir	08/01/2016	30/01/2016	6197
Krampus	15/12/2015	31/12/2015	5718
The Hunger Games	15/12/2015	31/12/2015	3024
Point Break	30/12/2015	30/01/2016	7254

Figure 2. Data collection

We have also made a data dictionary with the list of positive and negative words ,
Number of Positive words: 5791
Number of Negative words: 3562

B) Data Preprocessing


Raw data
Converting to lower/upper case
Remove URLs
Remove unicode characters
Remove newline characters
Preprocessed data

- Converting to lower/upper case: In order to simplify the process of analysis, we convert the whole text into upper/lower case so that words with different cases but same spelling are not differentiated by the classifier.

- Removing URLs: Hyperlinks in tweets do not play much role in sentiment classification hence they have been removed.

- Removing unicode characters: Unicode characters are used to represent emoticons and many other complex symbols. So to avoid complexity for preprocessing we should remove this characters. For example,

  "@deepikapadukone bae congrats for two best actress nomination in filmfare #Piku #BajiraoMastani <ed><U+00A0><U+00BD><ed><U+00B9><U+008C><U+2764>"

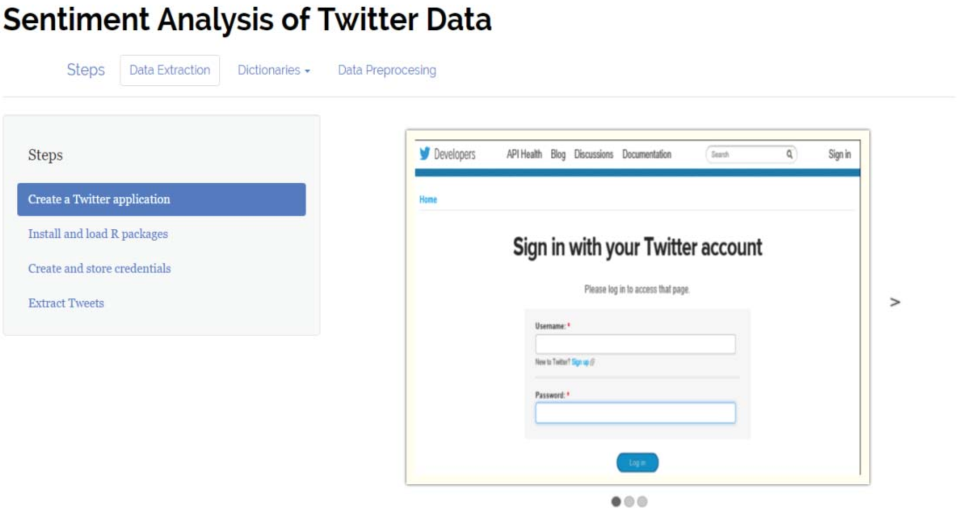
  The above tweet contains unicode characters that doesn't play any role in sentiment detection we removed it, and the after removal tweet will be like this,

  "@deepikapadukone bae congrats for two best actress nomination in filmfare #Piku #BajiraoMastani"

- Removing newline characters: These characters are just to indicate a newline, represented by “\n”, hence not required for sentiment classification.

We have implemented the project using R programming language. The tweets have been extracted using the “twitteR” package which uses the credentials provided by Twitter to access the Twitter API. The raw data extracted contains fields like text, favorited, favoriteCount, replyTOSN, created, truncated, replyToSID, id. The preprocessing was performed using the “tm” and “SnowballC” packages which are used for text mining. The graphical user interface has been created using the “Shiny”, “markdown”, “DT” packages available in R. We have employed the Twitter POS tagger from CMU [14] which is especially designed for Twitter data. Unlike most POS taggers which are based on the Penn Treebank tags, a set of specialized tags is used to annotate the tweets such as E (emotion), @ (at-mention), U (URL), N (noun), V (verb), and A (adjective).

C) Screenshots of the system



Sentiment Analysis of Twitter Data

Steps

Create a Twitter application

Install and load R packages

Create and store credentials

Extract Tweets

Sign in with your Twitter account

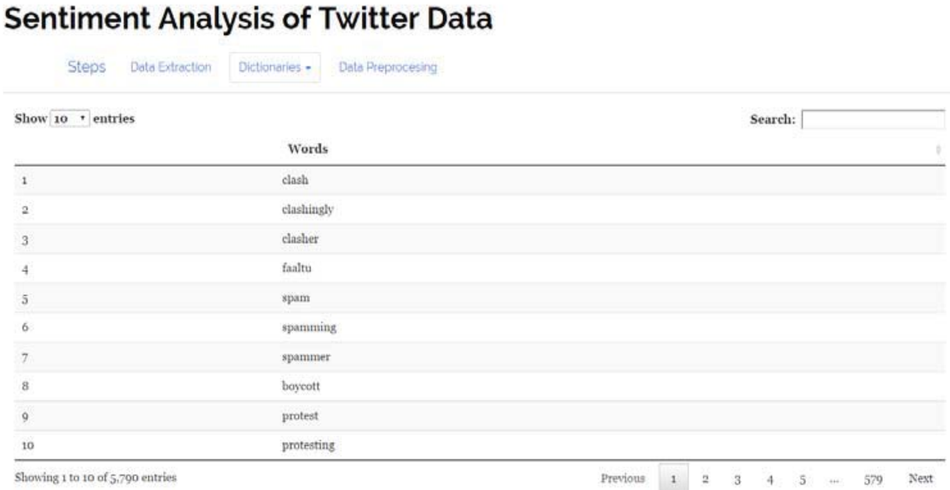
Please log in to access that page.

Username: *

New to Twitter? Sign up

Password: *

Figure 3. Data Extraction



Sentiment Analysis of Twitter Data

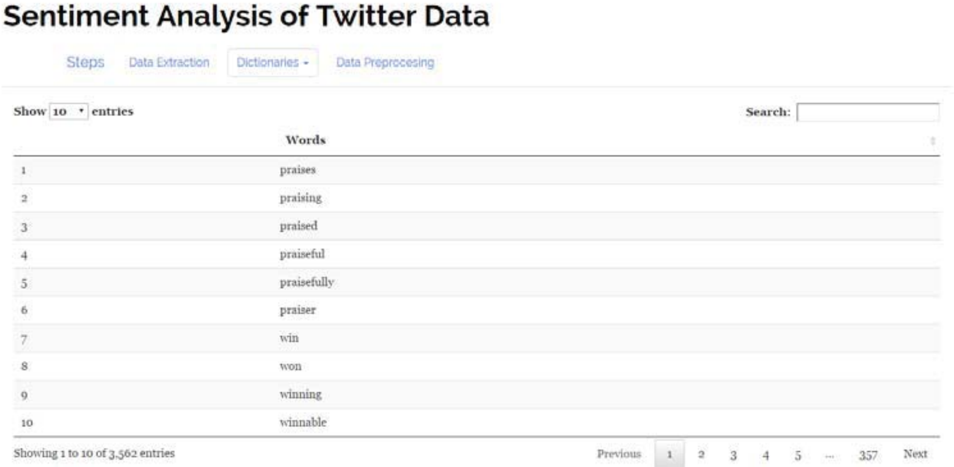
Show 10 entries

Search:

	Words
1	clash
2	clashingly
3	clasher
4	faaltu
5	spam
6	spamming
7	spammer
8	boycott
9	protest
10	protesting

Showing 1 to 10 of 5,790 entries

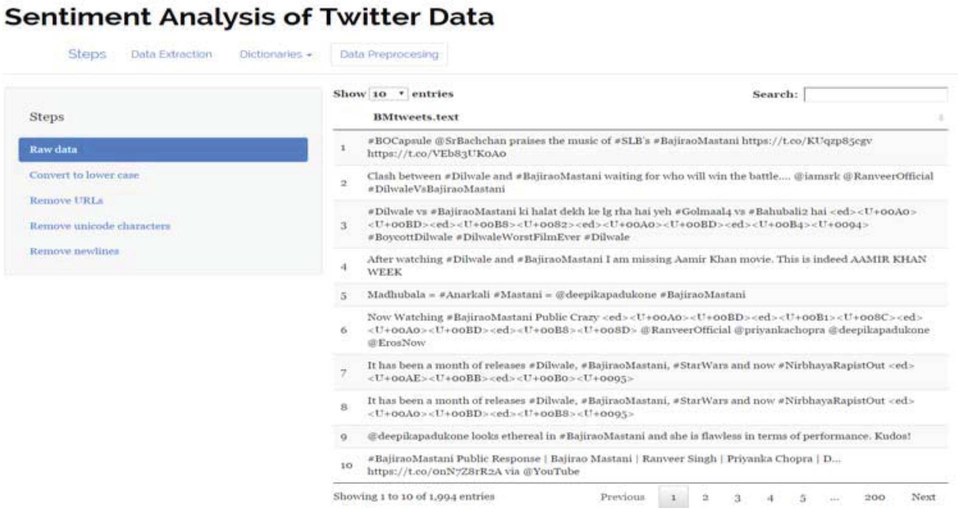
Figure 4(a). Data dictionaries(Negative words)



	Words
1	praises
2	praising
3	praised
4	praiseful
5	praisefully
6	praiser
7	win
8	won
9	winning
10	winnable

Showing 1 to 10 of 3,562 entries

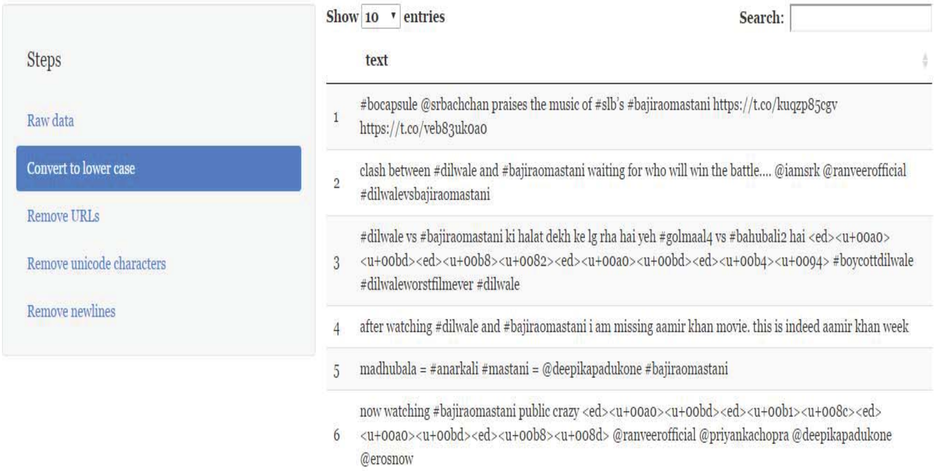
Figure 4(b). Data Dictionary(Positive words)



	BMtweets.text
1	#BOCapsule @SrBachchan praises the music of #SLB's #BajiraoMastani https://t.co/KUqzp85cgv https://t.co/VEb83UKoAo
2	Clash between #Dilwale and #BajiraoMastani waiting for who will win the battle.... @iamsrk @RanveerOfficial #DilwaleVsBajiraoMastani
3	#Dilwale vs #BajiraoMastani ki halat dekh ke lg rha hai yeh #Golmaal4 vs #Bahubali2 hai <ed><U+00A0><U+00BD><ed><U+00B8><U+0082><ed><U+00A0><U+00BD><ed><U+00B4><U+0094> #BoycottDilwale #DilwaleWorstFilmEver #Dilwale
4	After watching #Dilwale and #BajiraoMastani I am missing Aamir Khan movie. This is indeed AAMIR KHAN WEEK
5	Madhubala = #Anarkali #Mastani = @deepikapadukone #BajiraoMastani
6	Now Watching #BajiraoMastani Public Crazy <ed><U+00A0><U+00BD><ed><U+00B1><U+008C><ed><U+00A0><U+00BD><ed><U+00B8><U+008D> @RanveerOfficial @priyankachopra @deepikapadukone @ErosNow
7	It has been a month of releases #Dilwale, #BajiraoMastani, #StarWars and now #NirbhayaRapistOut <ed><U+00AE><U+00BB><ed><U+00B0><U+0095>
8	It has been a month of releases #Dilwale, #BajiraoMastani, #StarWars and now #NirbhayaRapistOut <ed><U+00A0><U+00BD><ed><U+00B8><U+0095>
9	@deepikapadukone looks ethereal in #BajiraoMastani and she is flawless in terms of performance. Kudos!
10	#BajiraoMastani Public Response | Bajirao Mastani | Ranveer Singh | Priyanka Chopra | D... https://t.co/onN7Z8rR2A via @YouTube

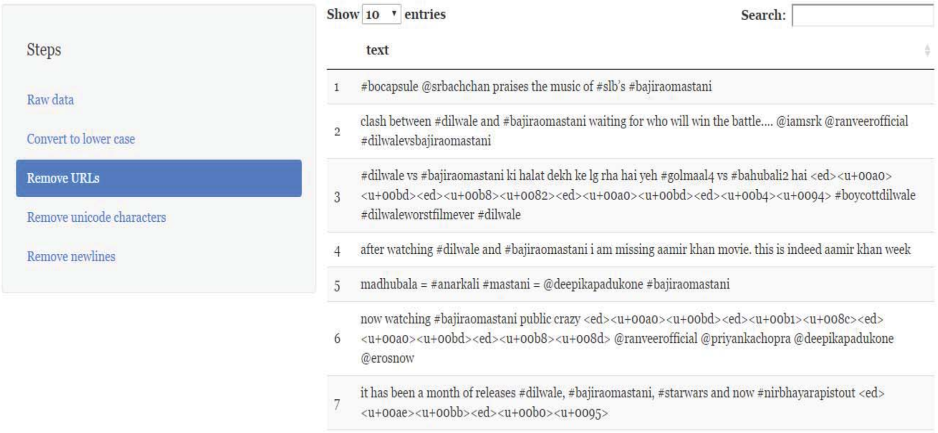
Showing 1 to 10 of 1,994 entries

Figure 5. Raw data



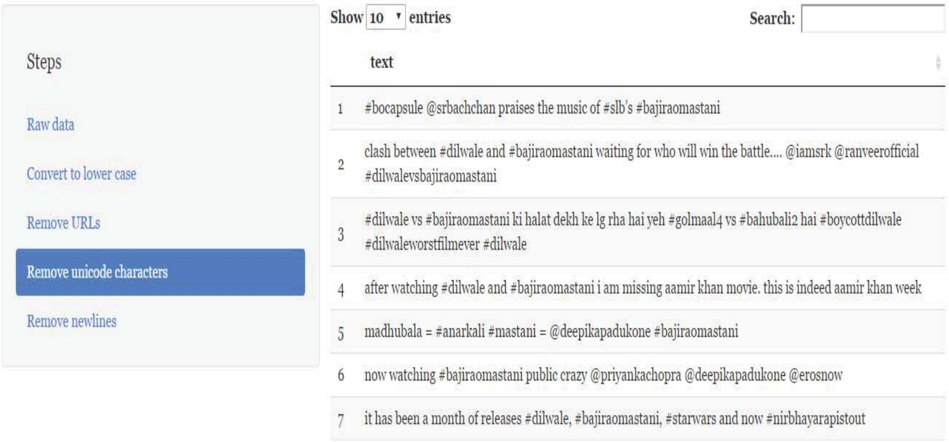
	text
1	#bocapsule @srbachchan praises the music of #slb's #bajiraomastani https://t.co/kuqzp85cgv https://t.co/veb83uk0a0
2	clash between #dilwale and #bajiraomastani waiting for who will win the battle.... @iamsrk @ranveerofficial #dilwalevsbajiraomastani
3	#dilwale vs #bajiraomastani ki halat dekh ke lg rha hai yeh #golmaal4 vs #bahubali2 hai <ed><u+00a0><u+00bd><ed><u+00b8><u+0082><ed><u+00a0><u+00bd><ed><u+00b4><u+0094> #boycottdilwale #dilwaleworstfilmever #dilwale
4	after watching #dilwale and #bajiraomastani i am missing aamir khan movie. this is indeed aamir khan week
5	madhubala = #anarkali #mastani = @deepikapadukone #bajiraomastani
6	now watching #bajiraomastani public crazy <ed><u+00a0><u+00bd><ed><u+00b1><u+008c><ed><u+00a0><u+00bd><ed><u+00b8><u+008d> @ranveerofficial @priyankachopra @deepikapadukone @erosnow

Figure 6.(a) Data Preprocessing(Lower case conversion)



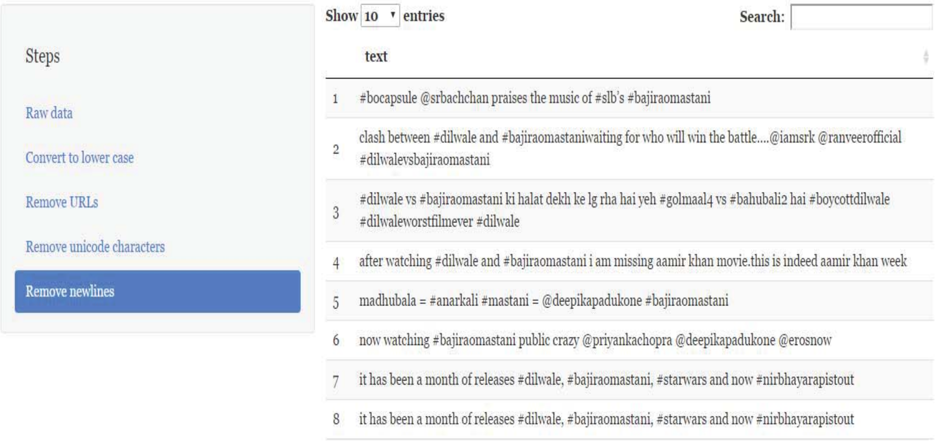
	text
1	#bocapsule @srbachchan praises the music of #slb's #bajiraomastani
2	clash between #dilwale and #bajiraomastani waiting for who will win the battle.... @iamsrk @ranveerofficial #dilwalevsbajiraomastani
3	#dilwale vs #bajiraomastani ki halat dekh ke lg rha hai yeh #golmaal4 vs #bahubali2 hai <ed><u+00a0><u+00bd><ed><u+00b8><u+0082><ed><u+00a0><u+00bd><ed><u+00b4><u+0094> #boycottdilwale #dilwaleworstfilmever #dilwale
4	after watching #dilwale and #bajiraomastani i am missing aamir khan movie. this is indeed aamir khan week
5	madhubala = #anarkali #mastani = @deepikapadukone #bajiraomastani
6	now watching #bajiraomastani public crazy <ed><u+00a0><u+00bd><ed><u+00b1><u+008c><ed><u+00a0><u+00bd><ed><u+00b8><u+008d> @ranveerofficial @priyankachopra @deepikapadukone @erosnow
7	it has been a month of releases #dilwale, #bajiraomastani, #starwars and now #nirbhayarapistout <ed><u+00ae><u+00bb><ed><u+00b0><u+0095>

Figure 6.(a) Data Preprocessing(Removal of URL's)



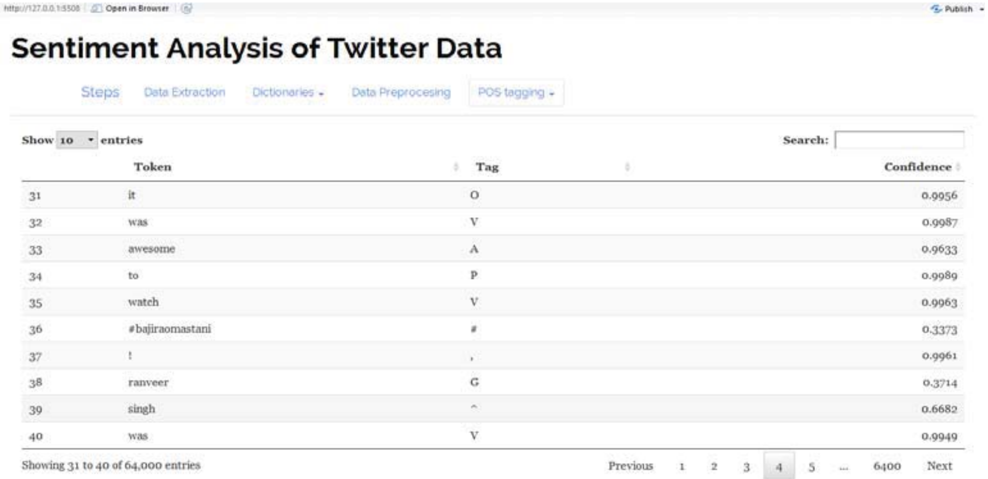
	text
1	#bocapsule @srbachchan praises the music of #slb's #bajiraomastani
2	clash between #dilwale and #bajiraomastani waiting for who will win the battle.... @iamsrk @ranveerofficial #dilwalevsbajiraomastani
3	#dilwale vs #bajiraomastani ki halat dekh ke lg rha hai yeh #golmaal4 vs #bahubali2 hai #boycottdilwale #dilwaleworstfilmever #dilwale
4	after watching #dilwale and #bajiraomastani i am missing aamir khan movie. this is indeed aamir khan week
5	madhubala = #anarkali #mastani = @deepikapadukone #bajiraomastani
6	now watching #bajiraomastani public crazy @priyankachopra @deepikapadukone @erosnow
7	it has been a month of releases #dilwale, #bajiraomastani, #starwars and now #nirbhayarapistout

Figure 6.(a) Data Preprocessing(Removal of Unicode characters)



	text
1	#bocapsule @srbachchan praises the music of #slb's #bajiraomastani
2	clash between #dilwale and #bajiraomastaniwaiting for who will win the battle....@iamsrk @ranveerofficial #dilwalevsbajiraomastani
3	#dilwale vs #bajiraomastani ki halat dekh ke lg rha hai yeh #golmaal4 vs #bahubali2 hai #boycottdilwale #dilwaleworstfilmever #dilwale
4	after watching #dilwale and #bajiraomastani i am missing aamir khan movie.this is indeed aamir khan week
5	madhubala = #anarkali #mastani = @deepikapadukone #bajiraomastani
6	now watching #bajiraomastani public crazy @priyankachopra @deepikapadukone @erosnow
7	it has been a month of releases #dilwale, #bajiraomastani, #starwars and now #nirbhayarapistout
8	it has been a month of releases #dilwale, #bajiraomastani, #starwars and now #nirbhayarapistout

Figure 6.(a) Data Preprocessing(Removal of newline characters)



	Token	Tag	Confidence
31	it	O	0.9956
32	was	V	0.9987
33	awesome	A	0.9633
34	to	P	0.9989
35	watch	V	0.9963
36	#bajiraomastani	#	0.3373
37	!	,	0.9961
38	ranveer	G	0.3714
39	singh	^	0.6682
40	was	V	0.9949

Showing 31 to 40 of 64,000 entries

Figure 7. POS tagged dataset

IX. CONCLUSION

In this paper we present an approach to perform aspect-level sentiment classification for Twitter. Thus far we have collected tweets using Twitter API, applied appropriate preprocessing on the tweets and performed POS tagging using R programming language. As, data retrieved from Twitter is very dirty, it is difficult to perform aspect level sentiment classification. Hence, our classifier will make use of POS tagger, dictionaries, aspect extraction and supervised machine learning algorithms. By the end of the project, we would understand the general sentiment around the movie, which aspects of the movie people liked or disliked and gain insights on how opinions on movies change over a period of time.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed, "Data mining : concepts and techniques "

[2] Mathew A. Russell, Trevor Hastie, Robert Tibshirani, Jerome Friedman, "Mining the Social Web"

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, "Elements of Statistical Learning"

[4] Stuart J. Russell and Peter Norvig, "Artificial Intelligence A Modern Approach Third Edition"

[5] Daniel Jurafsky and James H. Martin, "An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition"

[6] R Programming for Data Science /Roger D. Peng.

[7] Kouloumpis, Efthymios; Wilson, Theresa; Moore, Johanna D, “Twitter Sentiment Analysis: The Good the Bad and the OMG!..”

[8] Councill, I., McDonald, R., Velikovich L., “What's great and what's not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis”

[9] A Pak, P Paroubek,” Twitter as a Corpus for Sentiment Analysis and Opinion Mining.”

[10] A Agarwal, B Xie, I Vovsha, O Rambow, “Sentiment analysis of twitter data”

[11] L Jiang, M Yu, M Zhou, X Liu, T Zhao, “Target-dependent twitter sentiment classification”

[12] A Go, R Bhayani, L Huang, “Twitter sentiment classification using distant supervision”

[13] John Dodd, “Twitter Sentiment Analysis.”

[14] www.socialmediatoday.com/content/who-benefits-sentiment-analysis.

[15] www.quora.com/What-are-the-applications-of-sentiment-analysis-Why-is-it-in-so-much-discussion-and-demand.

[16] www.b-eye-network.com/view/15276

[17] www.sentiment.christopherpotts.net/lingcog.html

[18] www.sciencedirect.com/science/article/pii/S2090447914000550

[19] www.socialmediaexplorer.com/social-media-monitoring/sentiment-analysis/

[20] www.value-scope.com/en/sentiment-analysis/

[21] www.semantria.com/sentiment-analysis

[22] www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf

[23] www.aclweb.org/anthology/C14-1221

[24] www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[25] www.dmi.unict.it/~faro/tesi/sentiment_analysis/SA2.pd