**Machine Learning Project**

**(2019-2020)**

**Tweets Classification**

**Synopsis**



**Institute of Engineering & Technology**

**Team members:**

**Aditya Srivastava(171500020)**

**Supervised By:**

**Dr. Suresh C. Raikwar**

**Assistant Professor**

**Department of computer engineering & applications**

Sentiment Analysis of Twitter Data

**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  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.

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 interface’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).

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.

II. DESIGN

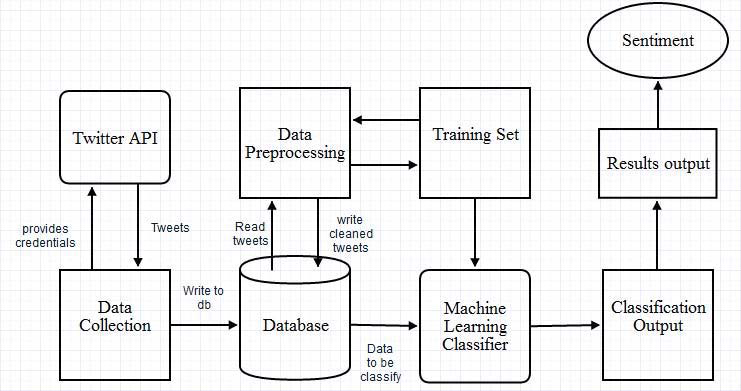
 A) Flow Diagram**

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.

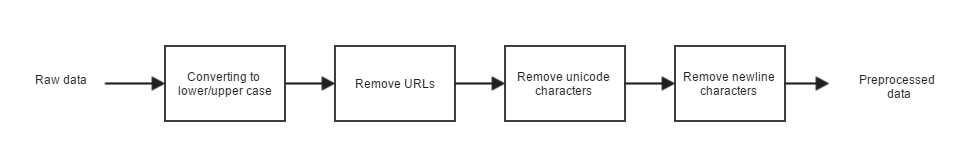
•  Sentiments in graphical representation: The results of the sentiment analysis is provided using pie charts.

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.

Figure 2. Data collection

(B)Data preprocessing. 

•  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 character.

•  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 python programming language. The tweets have been extracted using the “twitter” package which uses the credentials provided by Twitter to access the Twitter API.

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  using python 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 aspect extraction and supervised machine learning algorithms. By the end of the project, we would understand the general sentiment , which aspects  people liked or disliked .