**CHAPTER 1**  
**INTRODUCTION AND**  
**LITERATURE SURVEY**

**1.1 INTRODUCTION**

In this chapter we are going to focus on introductions on various aspects of our thesis .We are going to explain about Sentiment Analysis , Twitter Sentiment Analysis, various tools of Python which are useful for Sentiment Analysis and Natural Language Toolkit (NLTK). Then we will describe the objective of our thesis. Finally we will analyze the importance of Sentiment Analysis and its various application in different fields of interest .

**1.1.2 Sentiment analysis**  
Sentiment analysis is that the domain of study of examine peoples thoughts, sentiments, evaluations, mind-set, and feeling from written language. Sentiment analysis systems are utilized in almost each content as a result of thoughts are vital to the majority human activities. They’re key influencers of our behaviours .Sentiment analysis uses tongue process and text analysis to spot and extract data from a few specific area of interest. Attributable to massive use of the social media such as blogs and social networking sites like Twitter the interest in sentiment analysis has raised to the next extent. There are several problems in Sentiment analysis. The first is that an opinion word that’s thought-about to be positive in one situation could also be taken negative in another scenario. The second challenge is that folks don’t continually reveal their opinions in the same method.

**1.1.3. Twitter Sentiment analysis**

Twitter Sentiment Analysis is the process of determining whether a tweet is positive, negative or neutral [1]. It can be used to identify people's opinion towards a brand or public action through the use of variables such as context, tone, emotion, etc. Researcher can use sentiment analysis to find out public opinion of any epidemic or health related issues . Health department can also use this analysis to gather critical knowledge of awareness level in public with respect to any particular epidemic (Zika Virus in this case)

**1.1.4. Python**

We have used programming language Python for our thesis which is high level and dynamic programming language . Python libraries have grown significantly in last 10 years and some of which are specifically designed for data analysis [2]. We have used Python 3.6 version. There are various open source libraries are available which is compatible with Python 3.6 .

Python is a simple programming language but its simplicity does not limit its versatility. There are other programming languages for data analysis such as 'R' and 'MATLAB' but they are not as flexible as python .

**1.1.4. Natural Language Toolkit (NLTK)**

Natural Language Toolkit (NLTK) is a Python library which is distributed under the GPL open source license and it has been rewritten multiple times in order to take advantage of recent development in Python language [3]. NLTK is a group of multiple python scripts which is used for data classification , text processing and tokenization. This library plays a major role to obtain sentiment from text data.

There are many functions which are very useful for data pre-processing .These functions are used to preprocess the twitter data to make them fit for extracting features. NLTK works well with various machine learning algorithm which are used for Sentiment classification.

We used Python as main programming language in which we have written our script for fetching tweets as well as sentiment analysis. NLTK is the library which does the most important task to classify text into either positive or negative or neutral classes.

**1.1.5. Naive Bayes Classifier**

Task of supervised machine learning is to infer a function from tagged trainning samples. We are going to use Naive Bayes classifier which is a classifier with probabilistic output. In such type of classifiers with probabilistic output we have option to reject with a probability distribution if we are not sure about prediction result and hence we can pass it for manual check up [4].

Suppose , x1 to xn is a dependent vectors and there is a class variable y. So according to Bayes' probability theorem:

P(y| x1,......, xn)=

According to independence assumption :

**formula**

For each 'i' , this becomes

**formula**

Since P(x1,......, xn) is constant on provided input , so we have classification rule as:

**formula**

To evaluate we can use MAP (Maximum A Posterior) estimation P(y) and P(xi|y) ; the P(y) of class 'y' is relative frequency in sample [5].

**1.1.6. Zika Virus**

Zika virus comes from virus family 'Flaviviridae' which is spread by Aedes mosquitoes during day time [6]. Zika virus is named after Zika forest situated in Uganda, the place where the virus was first isolated in 1947 [7]. Initially no major symptoms are identified in Zika fever affected person [6]. Zika virus can also be spread through various other modes such as from mother to her infant, through sexual contact and through blood transfer from a Zika affected person [8]. There is no vaccine available for Zika virus [9].