**Introduction:**

Social Media has become a very powerful tool for information exchange as it allows users to not only consume information but also share and discuss various aspects of their

interest. Nevertheless, online social platforms are beset with hateful speech - content that expresses hatred for a person or group of people. Such content can frighten, intimidate, or silence platform users, and some of it can incite other users to commit violence. Furthermore, social media gives users the freedom to express their thoughts in text without following traditional language grammars, thereby making it difficult to mine social media for insights. Despite widespread recognition of the problems posed by social media content, reliable solutions even for detecting hateful speech are lacking. The main goal of this study is to develop a reliable tool for detection of hate tweets. This paper develops an approach for detecting and classifying hateful speech that uses content produced by self-identifying hateful communities from Twitter.

**1.1 Description:.**

If we conflate hate speech and offensive language then we erroneously consider many people to be hate speakers and fail differentiate between commonplace offensive language and serious hate speech. Given the legal and moral implications of hate speech it is important that we are able to accurately distinguish between the two. Lexical methods are effective ways to identify potentially offensive terms but are inaccurate at identifying hate speech; only a small percentage of tweets flagged by the Hatebase lexicon were considered hate speech by human coders. While automated classification methods can achieve relatively high accuracy at differentiating between these different classes, close analysis of the results shows that the presence or absence of particular offensive or hateful terms can both help and hinder accurate classification. Results from experiments showed Naive Bayes classifier achieved significantly better performance than existing methods in hate speech detection algorithms with precision, recall, and accuracy values of 58%,62%,and67.47%,respectively.

**1.2 Problem Formulation**

The aim of the study is to evaluate the performance for sentiment classification in terms of accuracy, precision and recall. Apart from the system’s ability to predict for a given tweet whether it is hateful or not, the system also generates a list of users who frequently post such content. This provides us with an interesting insight into the

usage pattern of hate-mongers in terms of how they express bigotry, racism and propaganda. The experimental results show that the classifiers yielded better results for the hate tweets review with the Naïve Bayes’ approach giving above 80%

accuracies and outperforming other algorithms.

**1.3 Motivation:**

In recent years, Twitter has become one of the most popular micro-blogging social-media platforms, providing a platform for millions of people to share their daily opinions/thoughts using real-time status updates. Due to high reachability and popularity of social media websites worldwide, organizations also use these websites for planning and mobilizing events for protests and public demonstration.Twitter is a famous platform for opinion and information sharing and this platform is mostly used before, during and after live events. An important and elusive form of such language is hateful speech: content that expresses hatred of a group in society. Hateful speech has become a major problem for every kind of online platform .Therefore, we are aiming at developing a reliable tool for detection of hate tweets.

**1.4 Proposed Solution:**

The proposed solution to this overwhelming problem is the building of such a model with the help of Machine learning that is able to differentiate between offensive language and hateful speech. The main goal of this study is to develop a reliable tool for detection of hate tweets.

**1.5 Scope of the Project:**

Online spaces are often exploited and misused to spread content that can be degrading, abusive, or otherwise harmful to people. Twitter prohibits users to post violent threats,

harassment, and hateful contents. However, there are still tons of users who disobey the rules and use their Twitter account to spread hate speech and negative words