**LITERATURE SURVEY**

**1) Automated hate speech detection and the problem of offensive language**

**T. Davidson, D. Warmsley, M. Macy, and I. Weber:** A key challenge for automatic hate-speech detection on social media is the separation of hate speech from other instances of offensive language. Lexical detection methods tend to have low precision because they classify all messages containing particular terms as hate speech and previous work using supervised learning has failed to distinguish between the two categories. We used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. We use crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. We train a multi-class classifier to distinguish between these different categories. Close analysis of the predictions and the errors shows when we can reliably separate hate speech from other offensive language and when this differentiation is more difficult. We find that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. Tweets without explicit hate keywords are also more difficult to classify.

**2) Identifying and categorizing profane words in hate speech**

**P. L. Teh, C. Bin Cheng, and W. M. Chee:** This study attempts to explore the different types of Hate Speech appearing in social media by identifying profane words used in hate speech. This study also compares the profane words used in different generations to assist in identifying the user's profile. Five-hundred (500) comments posted on YouTube on the abusive topics were collected. Profane words are classified into eight different types of hate speech. The finding shows 35% of profane words found in our sample are words related to sexual orientation. Comparison of the terms between 1970 and 2017 also show a high percentage of profane words are sexual orientation. Though the results are found based on only 500 comments collected from YouTube link in the current study, they are useful in establishing the list of profane words which will serve as the base for automatic hate speech identification in our future study. The originality of this research is the development of a training list of profane words for each category and comparison of the type of the words used in 1970 century with today's social media platform.

**3) A survey on automatic detection of hate speech in text**

**P. Fortuna and S. Nunes:** The scientific study of hate speech, from a computer science point of view, is recent. This survey organizes and describes the current state of the field, providing a structured overview of previous approaches, including core algorithms, methods, and main features used. This work also discusses the complexity of the concept of hate speech, defined in many platforms and contexts, and provides a unifying definition. This area has an unquestionable potential for societal impact, particularly in online communities and digital media platforms. The development and systematization of shared resources, such as guidelines, annotated datasets in multiple languages, and algorithms, is a crucial step in advancing the automatic detection of hate speech.

**4) Detecting hate speech and offensive language on twitter using machine learning: An N-gram and TF IDF based approach**

**A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat:** Toxic online content has become a major issue in today's world due to an exponential increase in the use of internet by people of different cultures and educational background. Differentiating hate speech and offensive language is a key challenge in automatic detection of toxic text content. In this paper, we propose an approach to automatically classify tweets on Twitter into three classes: hateful, offensive and clean. Using Twitter dataset, we perform experiments considering n-grams as features and passing their term frequency-inverse document frequency (TFIDF) values to multiple machine learning models. We perform comparative analysis of the models considering several values of n in n-grams and TFIDF normalization methods. After tuning the model giving the best results, we achieve 95.6% accuracy upon evaluating it on test data. We also create a module which serves as an intermediate between user and Twitter.

**5) Hate me, hate me not: Hate speech detection on Facebook**

**F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi:** While favoring communications and easing information sharing, Social Network Sites are also used to launch harmful campaigns against specific groups and individuals. Cyberbullism, incitement to self-harm practices, sexual predation are just some of the severe effects of massive online offensives. Moreover, attacks can be carried out against groups of victims and can degenerate in physical violence. In this work, we aim at containing and preventing the alarming diffusion of such hate campaigns. Using Facebook as a benchmark, we consider the textual content of comments appeared on a set of public Italian pages. We first propose a variety of hate categories to distinguish the kind of hate. Crawled comments are then annotated by up to five distinct human annotators, according to the defined taxonomy. Leveraging morpho-syntactical features, sentiment polarity and word embedding lexicons, we design and implement two classifiers for the Italian language, based on different learning algorithms: the first based on Support Vector Machines (SVM) and the second on a particular Recurrent Neural Network named Long Short Term Memory (LSTM). We test these two learning algorithms in order to verify their classification performances on the task of hate speech recognition. The results show the effectiveness of the two classification approaches tested over the first manually annotated Italian Hate Speech Corpus of social media text