Title: Analysis of women safety in India using machine learning

Objective: Women and girls have been experiencing a lot of violence and harassment in public places in various cities starting from stalking and leading to sexual harassment or sexual assault. Our main focus in this project is the role of social media in promoting the safety of women in Indian cities with special reference to social media platform Twitter. Twitter and other Twitter handles which include hashtag messages that are widely spread across the whole globe sir as a platform for women to express their views about how they feel while we go out for work or travel in a public transport and what is the state of their mind when they are surrounded by unknown men and whether these women feel safe or not?.

Literature Review

Paper 1: Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python

<https://pdfs.semanticscholar.org/c114/7f3d9b46ff0a0c7c43b668123cb15a26120d.pdf>

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then check its accuracy so that we can use this model for future use according to the results. It comprises steps like data collection, text preprocessing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching an efficiency of almost 85%-90%. But it still lacks the dimension of diversity in the data. Along with this, it has a lot of application issues with the slang used and the short forms of words. Many analyzers don’t perform well when the number of classes is increased. Also, it’s still not tested how accurate the model will be for topics other than the one in consideration. Hence sentiment analysis has a very bright scope of development in the future.

Paper 2: Sentiment Analysis of Twitter Data: A Survey of Techniques

In this paper, the authors provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with cross-domain and cross-lingual methods and some evaluation metrics. The authors also observed that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few efforts in the human-labeled document. They also studied the effects of various features on the classifier. And the conclusion was that more cleaner data, more accurate results can be obtained. The use of the bigram model provides better sentiment accuracies as compared to other models.

And the study of combining the machine learning method into opinion lexicon method in order to improve the accuracy of sentiment classification and adaptive capacity to variety of domains and different languages.

Paper 3:Classifying Sentiment in Microblogs: Is Brevity an Advantage?

The results of the experiments, on the whole, are encouraging for the task of analyzing sentiment in microblogs. They achieved an accuracy of 74.85% for binary classification for a diverse set of topics indicating we can classify microblog documents with a moderate degree of confidence. In short-form corpora, they find it difficult to improve performance by extending a unigram feature representation. This is contrary to the long-form corpora which responds favorably to enriched feature representations. All datasets speculate that engineering features based on deeper linguistic representations such as dependencies and parse trees may work for microblogs. In analyzing discriminative features, they found that a significant role is played by punctuation.

Paper 4:Robust Sentiment Detection on Twitter from Biased and Noisy Data

In this paper, the authors propose an approach to automatically detect sentiments on Twitter messages (tweets) that explores some characteristics of how tweets are written and meta-information of the words that compose these messages. They have presented an effective and robust sentiment detection approach for Twitter messages, which uses biased and noisy labels as input to build its models. The performance of the classifier is due to the fact that: their approach creates a more abstract representation of these messages, instead of using a raw word representation of them as some previous approaches; and although noisy and biased, the data sources provide labels of reasonable quality and, since they have different bias, combining them also brought some benefits.

Paper 5:Sentiment Analysis on Twitter Data using KNN and SVM

In this paper, the authors are trying to find out the positive and negative sentiment on Twitter data. They designed the classifier with only a few features like n0 0.5 1 100 500 1000 1500 2000 2500 3000 KNN with normalization (4 features) KNN with normalization and keyword base( 5 features) SVM with (4 features) SVM with normalization (4 features) SVM with normalization and keyword base (5 features)gram feature, pattern feature, punctuation feature, keyword-based feature, and word feature. They also, use machine learning algorithm SVM (Support Vector Machine). They also, use the KNN classifier and calculate the accuracy of all algorithms. In this paper, they are focusing on dividing the tweets into positive and negative sentiment. And it is observed that the sentiment classifier algorithm (SCA) performs better than SVM.

Paper 6:Sentiment Analysis of Twitter Data

<https://pdfs.semanticscholar.org/ffe0/fa5f2ce6709ff6b1750f9bbc9e31929b25b2.pdf>

The authors experiment with three types of models: a unigram model, a feature-based model, and a tree kernel-based model. For the feature-based model, they use some of the features proposed in past literature and propose new features. For the tree kernel-based model they design a new tree representation for tweets. They use a unigram model, as their baseline. Their experiments show that a unigram model is indeed a hard baseline achieving over 20% over the chance baseline for both classification tasks. Their feature-based model that uses only 100 features achieves similar accuracy as the unigram model that uses over 10,000 features. And the tree kernel-based model outperforms both these models by a significant margin.