UNIVERSITÉ Concordia UNIVERSITY FACULTY OF ENGINEERING AND COMPUTER SCIENCE

Project Proposal for INSE 6180
Security and Privacy Implications of Data Mining
Winter 2024

Project Title:
OSINT Analysis on Reddit Comments and Posts for Demographics of Sentiments, Violent Posts, and Content Moderation

Presented to:
Prof. Nizar Bouguila

Made and Presented by:

Name	Student Number
Rahul Hulli	40234542

Abstract:

The goal of this project is to evaluate articles and comments on Reddit by using Open-Source Intelligence (OSINT) methodologies. One of the biggest conversation sites on the internet, Reddit offers a plethora of content created by its users, which makes it a great place to learn about attitudes, communities, and trends in the online community. The main goals are to identify violent posts for content moderation and use sentiment analysis to gauge the emotional tone of discussions to provide a demographic of those expressing suicidal thoughts. The goal is to advance the field of social media analytics and deepen the knowledge of online speech and behavior.

Introduction:

The way individuals connect, and exchange information online has changed dramatically because of the widespread use of social media platforms. Reddit is a well-known website that is well-known for both its wide variety of debate subjects and its varied user base. Researchers can learn a great deal about user feelings, community dynamics, and new trends by examining Reddit posts and comments. Using Open-Source Intelligence (OSINT), this initiative seeks to better understand online behavior by gleaning valuable information from Reddit data.

Problem Statement:

Analysis on Reddit is made easier or harder by the volume of user-generated content there. Finding pertinent insights from the massive amounts of data is one of the main issues. Further endangering user safety and community well-being is the existence of inappropriate or violent content. The development of methods for sentiment analysis, the detection of violent messages, and content moderation on Reddit is, thus, the main problem statement of this project.

Motivation:

This study is driven by the rising realization that, in the current digital era, comprehending online groups and behaviors is crucial. One can learn about user sentiments, preferences, and interactions by examining Reddit data. Maintaining a polite and safe online environment also depends on the ability to recognize and control unwanted content. The goal of this project is to provide methods and instruments that will improve content analysis and advance the area of social media research.

Objectives:

The main objectives of this project are:

1. Conduct sentiment analysis to understand the emotional tone of Reddit comments and posts.
2. Develop methods for identifying violent or aggressive content within Reddit discussions.
3. Evaluate the performance of the developed techniques using appropriate metrics and validation methods.
4. Generate insights and findings from the analysis to contribute to the understanding of online communities and behaviors.

Literature Review:

- **Sentiment Analysis on Social Media:**

Because social media sites like Facebook and Twitter have large user bases and a wealth of textual data, researchers have been paying close attention to these platforms for sentiment analysis. To analyze sentiment in social media content, numerous studies have investigated a variety of natural language processing (NLP) strategies (Pak & Paroubek, 2010 [6]). These techniques range from lexicon-based methods to complex machine-learning algorithms and deep-learning models [7]. However, there are particular difficulties with sentiment analysis on social media, such as the usage of slang, emojis, and informal language in addition to context-dependent sentiment expressions. Researchers have suggested creative solutions to these problems, like using context-aware sentiment analysis methods and integrating domain-specific lexicons (Agarwal et al., 2011 [8]).

- **Detection of Violent and Aggressive Content:**

Maintaining user safety and promoting a positive online environment requires the identification of violent or hostile content on online platforms. In this field, research has concentrated on creating automated systems that employ machine learning and natural language processing (NLP) to identify and categorize violent content (Zhang et al., 2018 [9]). Previous research has looked at keyword-based strategies, which use lists of violent terms to find potentially dangerous content. Furthermore, studies have looked into the use of supervised learning algorithms, like Random Forest and K-nearest neighbors (KNN), to categorize violent messages according to textual characteristics (Davidson et al., 2017 [10]).

- **Content Moderation and Hate Speech Detection:**

To stop the spread of dangerous content and encourage polite conversation, content moderation, and hate speech identification are essential components of online platforms. Scholars have utilized diverse natural language processing (NLP) methodologies, such as sentiment analysis, text categorization, and topic modeling, to automatically detect hate speech and objectionable language (Fortuna & Nunes, 2018 [11]). Because they can handle high-dimensional feature spaces and non-linear decision boundaries, Support Vector Machines (SVM) have become a popular option for hate speech detection applications. Online content can be classified as hate speech or non-hate speech using SVM-based models that have been trained on annotated datasets (Sobhani et al., 2019 [12]).

- **Evaluation Metrics for Text Classification:**

It is necessary to employ suitable metrics to evaluate the accuracy, precision, recall, and F1 scores of text categorization models. While choosing assessment measures for text classification tasks, researchers advise considering class imbalances, label noise, and evaluation biases. The model's capacity to accurately classify occurrences of interest, such as violent posts or hate speech, may be understood using metrics like precision, recall, and F1-score. By calculating the trade-off between false positives and false negatives, these metrics help assess how successful content moderation algorithms are in practical applications.

Methodology

Step	Description
Data Collection	Obtain a dataset containing Reddit comments and posts, including relevant attributes.
Data Preprocessing	Clean the data by removing irrelevant information, handling missing values, and preprocessing text data.
Sentiment Analysis	Perform sentiment analysis on comments and posts using NLP techniques and ML algorithms.
Identification of Violent Posts	Create keyword lists containing terms related to violence or aggression. Use ML algorithms to classify posts.
Content Moderation	Utilize NLP techniques to identify inappropriate or offensive language, hate speech, or misinformation.
Evaluation and Validation	Assess the performance of each analysis component using metrics such as accuracy, precision, recall, and F1-score. Validate results against ground truth labels.
Visualization and Reporting	Visualize findings using graphs, charts, and dashboards. Prepare a comprehensive report summarizing methodology, results, and implications.

Proposed Models for Analysis

Analysis	Description	Proposed Model
Sentiment Analysis	Determine the emotional tone of Reddit comments and posts.	- NLP techniques for sentiment analysis (positive, negative, neutral). – SVM for classification.
Identification of Violent Posts	Detect posts containing violent or aggressive content.	- Keyword lists for violence-related terms. - Random Forest, KNN for classification.

Content Moderation	Automatically detect and flag inappropriate or offensive content in Reddit comments and posts.	- NLP techniques to identify offensive language, hate speech, or misinformation. - SVM for classification.

Conclusion

To detect violent content, enable content moderation, and extract insightful information about feelings, this project aims to do an Open-Source Intelligence (OSINT) analysis on Reddit posts and comments. Through utilizing Reddit's enormous user-generated data repository, this initiative aims to gain a deeper comprehension of online groups, trends, and habits. In the end, this study has the potential to make a substantial contribution to our understanding of online conversation and open the door to the development of more successful digital platform content moderation techniques.

References:

[1] Top Reddit Posts and Comments: https://www.kaggle.com/datasets/tushar5harma/topredditcomments
[2] Reddit Sentiment Analysis Application: https://github.com/jalvin99/RedditSentimentAnalyzer
[3] Sentiment Analysis, Part 1 — A friendly guide to Sentiment Analysis: https://medium.com/besedo-engineering/sentiment-analysis-part-1-a-friendly-guide-to-sentiment-analysis-963e09d9fc9b
[4] Natural Language Processing To Analyze Abuse and Domestic Violence Subreddits During COVID-19: https://github.com/amiekong/nlp-reddit-analysis?tab=readme-ov-file
[5] Topic modeling using Latent Dirichlet Allocation(LDA) and Gibbs Sampling explained! https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045
[6] Twitter as a Corpus for Sentiment Analysis and Opinion Mining: https://aclanthology.org/L10-1263/
[7] Starters Guide to Sentiment Analysis using Natural Language Processing: https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/
[8] Context-aware Models for Twitter Sentiment Analysis: https://journals.openedition.org/ijcol/322
[9] Deep learning for sentiment analysis: A survey: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1253
[10] Automated Hate Speech Detection and the Problem of Offensive Language: https://arxiv.org/abs/1703.04009
[11] A Survey on Automatic Detection of Hate Speech in Text: https://dl.acm.org/doi/10.1145/3232676
[12] Emotionally Informed Hate Speech Detection: A Multi-target Perspective: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8236572/