Content Filtering of Twitter Posts Using Multi Class Text Classification

Report 1

|  |  |
| --- | --- |
| Advanced Big Data Tools | Kislitsin Aleksei |

Table of contents

[1. Introduction 3](#_Toc149540553)

[2. Problem Statement 3](#_Toc149540554)

[3. Related Works 3](#_Toc149540555)

[4. Proposal method 4](#_Toc149540556)

[Phase 1: Data Preparation 4](#_Toc149540557)

[Phase 2: Model Selection 5](#_Toc149540558)

[Phase 3: Model Training and Optimization 5](#_Toc149540559)

# 1. Introduction

In the digital age, social media platforms, such as Twitter, have become a primary source of information, communication, and engagement. However, the unfiltered and diverse nature of these platforms can lead to exposure to sensitive or undesirable content, including topics like religion, violence, pornography, terrorism, politics, and more. This project aims to address this issue by developing a content filtering system that uses a multi-class text classification machine learning model pretrained on a set of categories. Users can select and exclude these sensitive topics to curate their Twitter feeds to align with their preferences and avoid undesirable content.

# 2. Problem Statement

Twitter users often encounter sensitive or undesirable content, such as religious discussions, violent imagery, pornography, terrorist propaganda, and polarized political discourse, in their feeds. This content can lead to discomfort, offense, or disengagement with the platform.

Issues:

* Political Content: According to Hootsuite website[[1]](#footnote-1), approximately one-third, or 33%, of all Twits are political in nature. Many people who use Twitter are not interested in politics, so these tweets will be irrelevant to them.
* Exposure to Inappropriate Content: Users may be exposed to explicit or offensive content, including erotic material or violent images, which can be distressing and unsuitable for a diverse audience.
* Religious Sensitivity: Sensitive discussions surrounding religion can sometimes escalate into heated debates and controversies, leading to discomfort and polarization among users.
* Bullying and Harassment: Bullying and harassment on Twitter can cause emotional distress and discomfort for users, affecting their mental well-being and discouraging open expression of opinions.
* Loss of Engagement: Users who find their Twitter feeds inundated with content they find sensitive, offensive, or disrespectful may disengage from the platform. This disengagement can result in a loss of user engagement and potential revenue for Twitter. For now, the average organic Twitter engagement rate is 0.05%, based on statistics of Hootsuite’s article[[2]](#footnote-2)

# 3. Related Works

These papers offer insights and techniques that are directly relevant to my research on Twitter text classification and content filtering. They can serve as valuable references, providing guidance on data processing, model selection, and strategies for improving the effectiveness of content filtering system.

1. **"****Detecting Rumors from Microblogs with Recurrent Neural Networks"** by Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie[[3]](#footnote-3)
2. **"****Twitter as a Corpus for Sentiment Analysis and Opinion Mining"** by Alexander Pak and Patrick Paroubek[[4]](#footnote-4)
3. **"****A Survey on Hate Speech Detection using Natural Language Processing"** by Yigitcan Kaya and Onur Varol[[5]](#footnote-5)
4. **"****Twitter Political Classification and Named Entity Recognition"** by Leon Derczynski, Diana Maynard, Giorgio Orsi, and Kalina Bontcheva[[6]](#footnote-6)
5. **"****Automatic Sentiment Analysis of Twitter Messages"** by Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas[[7]](#footnote-7)

Despite the substantial volume of research papers related to the semantic analysis of tweets, it is evident that there is a noticeable scarcity of studies specifically focused on multi-class text classification. While numerous works have explored various aspects of natural language processing and sentiment analysis on Twitter, the focus on the intricate task of categorizing tweets into multiple distinct classes or categories remains relatively limited.

# 4. Proposal method

The project comprises three distinct phases, each designed to build upon the previous one and ultimately achieve the maximum possible accuracy in content filtering on Twitter. Here is an overview of the proposed methodology for each of these phases:

## Phase 1: Data Preparation

In the initial phase, we will focus on preparing the data for subsequent modeling. This phase includes several key stages:

* Choosing the Dataset: The first step is to select a suitable dataset that encompasses a diverse range of Twitter content, including sensitive and non-sensitive categories. The dataset should be large enough to be representative but manageable for processing.
* Labeling Data: We will categorize the tweets in the dataset into distinct sensitivity and relevance classes. Labeling is essential for supervised machine learning, enabling the model to learn the characteristics of different categories.
* Balancing Classes: To ensure the model's ability to learn from all categories without bias, we will balance the classes. This might involve oversampling, undersampling, or using other techniques to achieve class balance.
* Clean Raw Text: Twitter data is often noisy and may contain special characters, URLs, and other non-essential elements. Preprocessing the data with PySpark will be performed to clean and standardize the text. This includes tasks such as removing special characters, handling hashtags, and tokenizing the text.

## Phase 2: Model Selection

The second phase involves testing different models to determine which one is best suited for our content filtering task. We will utilize the same training and testing splits for consistency. Model selection workflow diagram represents step by step execution of this phase. The key steps in this phase include:

* Testing Model Groups: We will explore three groups of models for comparison. The first group comprises traditional machine learning models: Linear SVM, Naive Bayes Classifier, and Logistic Regression. The second group will incorporate Word2Vec or Doc2Vec embeddings with Logistic Regression. Finally, the third group will focus on deep learning using Keras.
* Estimating Results: Each model group will be trained and tested, and their performance metrics, including accuracy, precision, recall, and F1-score, will be estimated. This will help us assess their effectiveness in content classification.
* Selecting the Best Model: After estimating the results, we will identify the model that yields the highest accuracy and overall best performance. This model will be chosen as the basis for further optimization.

## Phase 3: Model Training and Optimization

The final phase concentrates on maximizing the selected model's accuracy through fine-tuning and optimization. The main steps in this phase are:

* Training the Selected Model: The chosen model from Phase 2 will undergo extensive training on the labeled data. This will involve adjusting hyperparameters, determining the optimal architecture (in the case of deep learning), and ensuring the model is well-fitted to the specific content filtering task.
* Hyperparameter Optimization: We will systematically search for the best hyperparameters to enhance model performance. This might involve techniques like grid search or random search, depending on the model's complexity.
* Assessing the Result: The final model will be rigorously evaluated using testing data to ensure it meets the highest standards of accuracy and generalization.

Model selection workflow diagram

A screenshot of a computer screen

Description automatically generated

The proposed methodology is designed to systematically progress from data preparation to model selection and, finally, to model optimization, ultimately leading to the development of an accurate content filtering system for Twitter. Each phase contributes to achieving the maximum possible accuracy in content classification.

1. 29 Twitter Stats That Matter to Marketers in 2023 (2023) - <https://blog.hootsuite.com/twitter-statistics/> [↑](#footnote-ref-1)
2. 29 Twitter Stats That Matter to Marketers in 2023 (2023) - <https://blog.hootsuite.com/twitter-statistics/> [↑](#footnote-ref-2)
3. Detecting Rumors from Microblogs with Recurrent Neural Networks (2016): <https://www.ijcai.org/Proceedings/16/Papers/537.pdf> [↑](#footnote-ref-3)
4. Twitter as a Corpus for Sentiment Analysis and Opinion Mining (2010): <https://aclanthology.org/L10-1263/> [↑](#footnote-ref-4)
5. A Survey on Hate Speech Detection using Natural Language Processing (2017): <https://aclanthology.org/W17-1101/> [↑](#footnote-ref-5)
6. Twitter Political Classification and Named Entity Recognition (2022): <https://arxiv.org/pdf/2209.08110.pdf> [↑](#footnote-ref-6)
7. Automatic Sentiment Analysis of Twitter Messages (2012): <https://ieeexplore.ieee.org/document/6412377> [↑](#footnote-ref-7)