Hate Speech Detection in Romanian Facebook Comments: An AI-Based Approach

Badescu Madalina Mihaela, Constantinescu Andrei Eduard,
Bouruc Liviu Petru

TABLE OF CONTENTS

1

INTRODUCTION

HATE SPEECH DETECTION

Hate speech detection on social media presents a complex and pressing problem, especially in multilingual and multicultural contexts like Romania.

SOCIAL MEDIA PLATFORMS

The widespread use of social media comes the challenge of managing and mitigating harmful content, such as hate speech, which can perpetuate discrimination, prejudice, and violence.

2

THE DATASET

DATA COLLECTION

We utilized the Facebook API to scrape comments from public posts written in Romanian, targeting public and controversial pages such as Cristian Tudor Popescu, Dan Negru, Klaus Iohannis and more

AI MODELS FOR LABELS

We employed four AI models to annotate the comments as hate speech or non-hate speech. Each AI model provided its interpretation of the comments, which served as the basis for our labeled dataset

AI MODEL FOR LABELS

PaLM 1 & PaLM 2

PaLM1 focuses on providing an approachable way to explore and prototype with generative AI applications, while PaLM2 represents a significant improvement in model size, training data, and performance, enabling more accurate and reliable results.

Vertex AI

Provides access to Gemini, a multimodal model capable of understanding diverse inputs and generating outputs across different modalities, including text, images, video, and code.

GPT 3.5

State-of-the-art language generation model developed by OpenAI. It excels in understanding and generating human-like text across various contexts, making it suitable for tasks such as sentiment analysis and classification.

DATASET BALANCING

To address this issue and ensure fair and accurate model training, we employed a technique called Random Under-sampling

DATASET PREPROCESSING

Crucial step in natural language processing (NLP) tasks, aimed at transforming raw text data into a format suitable for machine learning models.

PREPROCESSING STEPS

Lowercasing	Convert all text to lowercase to ensure uniformity in text representation and reduce the complexity of text processing.
Unicode Normalization	Normalize text using the unidecode library to convert accented characters and Unicode characters into their closest ASCII equivalents, simplifying text representation.
Tokenization	Tokenize the text into individual words or tokens using spaCy's language processing pipeline (nlp), enabling granular analysis and manipulation of text at the token level.
Stopword Removal	Remove stopwords, such as common words like "the," "and," "is," etc., using a predefined list of stopwords for the Romanian language. This step helps eliminate noise and irrelevant information from the text
Punctuation Removal	Removes punctuation, such as ",", "!", ".", since it is not important to the overall model.
Spell check	The spelling check aims at improving the quality of the text data before it is used for classification or analysis. Specifically, it attempts to correct misspelled words using a spell-checking library (pspell).
Stemming	Apply stemming using the SnowballStemmer for Romanian ("romanian") to reduce words to their root or base form. Stemming helps standardize word variations and improve the model's ability to generalize across different forms of words.

FEATURE REPRESENTATION

Feature representation is a critical aspect of natural language processing (NLP) and machine learning tasks, as it involves transforming raw text data into a format that can be understood and processed by algorithms

FEATURE REPRESENTATION

TF-IDF

Popular feature representation technique that assigns weights to terms based on their frequency in a document relative to their frequency across all documents in the corpus. It measures the importance of a term within a document while considering its prevalence in the entire corpus.

BAG OF WORDS

Representation is based on the vocabulary of the corpus, where each unique word corresponds to a dimension in the feature space. The value of each dimension represents the frequency of the corresponding word in the document.

3

MODELS & TRAINING

GPT 3.5 Dataset

Run	Tokenization Method	Preprocessing Steps	TF-IDF Results (Accuracy, Weighted Avg)	BoW Results (Accuracy, Weighted Avg)
1	Whitespace	Remove stopwords. Apply stemming	0.63, 0.63	0.63, 0.62
2	Word Tokenization	Remove stopwords. Apply stemming	0.65, 0.65	0.64, 0.63
3	Word Tokenization	Remove stopwords. No stemming	0.63, 0.63	0.64, 0.62
4	SpaCy Tokenization	Remove stopwords. No stemming	0.65, 0.64	0.65, 0.63
5	SpaCy Tokenization	Remove stopwords. Apply stemming.	0.66, 0.66	0.66, 0.65
6	SpaCy Tokenization	Remove stopwords. Remove Punctuation Marks. Apply stemming.	0.65, 0.65	0.65, 0.65
7	SpaCy Tokenization	Remove stopwords. Correct misspelled words. Apply stemming.	0.63, 0.63	0.63, 0.63
8	SpaCy Tokenization	Remove stopwords. Remove Punctuation Marks. Correct misspelled words. Apply Stemming.	0.68, 0.68	0.68, 0.67

Other Datasets

Dataset	Model	Undersampling	TF-IDF Results (Accuracy, Weighted Avg)	BoW Results (Accuracy, Weighted Avg)
PaLM 1	TF-IDF SVM	With	0.82, 0.80	0.81, 0.78
PaLM 1	TF-IDF SVM	Without	0.72, 0.74	0.73, 0.74
PaLM 1	BoW SVM	With	0.81, 0.78	0.73, 0.74
PaLM 1	BoW SVM	Without	0.73, 0.76	0.75, 0.76
PaLM2	TF-IDF SVM	With	0.83, 0.80	0.81, 0.78
PaLM2	TF-IDF SVM	Without	0.72, 0.74	0.75, 0.76
PaLM2	BoW SVM	With	0.81, 0.78	0.73, 0.75
PaLM2	BoW SVM	Without	0.75, 0.76	0.76, 0.77
Vertex	TF-IDF SVM	With	0.84, 0.81	0.81, 0.78
Vertex	TF-IDF SVM	Without	0.71, 0.73	0.73, 0.75
Vertex	BoW SVM	With	0.81, 0.78	0.73, 0.75
Vertex	BoW SVM	Without	0.73, 0.75	0.75, 0.76

4

CONCLUSION

CONCLUSION

Hate speech detection on social media platforms is a critical endeavor in today's digital age, especially considering the harmful consequences it can have on individuals and communities. In this project, we undertook the task of hate speech detection in Romanian Facebook comments using advanced technologies such as Natural Language Processing (NLP) and machine learning.

THANK YOU!

Q&A