# LUSIP PROJECT REPORT

# DESCRIPTIVE ANALYSIS OF HATE SPEECH OF REDDIT COMMENTS

Made by-

- **Mohammed Atif Khan (21UCS130)**
- **Mohammed Abdul Haq (21UCS129)**
- **Ramgopal Reddy(21UCS167)**
- **Divya Bajaj (21UCC036)**

# Abstract

- Social media is a collection of internet-based platforms that allow users to create and interact with content in real-time. It enables individuals, groups, and organizations to connect, communicate, and share information. However, the relationship between social media and free speech is complex and can lead to the spread of hate speech and harmful content.
- As the internet and social media continue to evolve, the problem of hate speech and offensive language on these platforms also evolves. Existing research highlights that hate is prevalent across various platforms, but there is a lack of models for detecting online hate using data from multiple platforms with the maximum accuracy needed.
- To gain a deeper understanding, we done a various descriptive analysis by analyzing the dataset provided by Improving the Detection of Multilingual Online Attacks with Rich Social Media Data from Singapore (Haber et al., ACL 2023) (Further description of the dataset is in the subsequent section of this document). we have conducted a rudimentary analysis to better comprehend Reddit platform dynamics and the prevalence of hate speech in language apart from english. This project aims to contribute to a comprehensive understanding of the challenges and implications of hate speech in the ever-changing landscape of social media.

# Introduction

- According to the United Nations (2019), hate speech is defined as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor." Despite this, multiple definitions of hate speech can exist. The term "Offensive language" is also used concurrently with hate speech. An observation that can be made from the way classifiers operate is that oftentimes hate speech is misclassified as offensive language and vice versa , from ( Davidson et al., 2017; Mozafari et al., 2019 ).
- Hate Speech is defined as language used to express hatred towards a targeted group/individual based on protected attributes such as race or religion, offensive language contains offensive terms but is not targeting any group in particular, from ( Yuan et al., WOAH 2022 ).
- For the purpose of this project, we have used binary classification (hate, offensive / neither). This implies that offensive comments will be labelled as hate as well.

# Hate Keyword Lexicon

- The researchers gathered five Spanish words associated with hateful speech towards women from the website hatebase.org. To expand their dataset, they used a technique called word embedding to find similar words in a large collection of Spanish texts. They excluded some words from the initial list as they were mainly related to animals. Afterwards, they used an online application to find synonymous words that were not already included. This helped them create three lexicons: one for misogyny, another for insults, and a third one for xenophobia. The lexicons were later translated into English.
  ( lor-Miriam Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2020 )
- HURTLEX is a lexicon that includes offensive, aggressive, and hateful words from 53 languages. It contains approximately 1000 words manually chosen to cover 17 specific categories. To expand the lexicon, additional words were added using MultiWordNet synsets and Babelnet in a semi-

automatic process. The lexicon was then refined and updated to create the modern version of HURTLEX.
( Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022.)

- All the lexicons mentioned earlier were combined into a single, comprehensive lexicon. To ensure its accuracy and avoid repetition, the lexicon was filtered, and duplicate words were removed. This process resulted in a larger and more refined lexicon.

# SINGAPORE DATASET ANALYSIS

- Citation: Improving the Detection of Multilingual Online Attacks with Rich Social Media Data from Singapore (Haber et al., ACL 2023)
- The data set is a translated version of the original data set to English using the Google Translate API.
- Time: The oldest comment in the labelled dataset is from May 2011, and the most recent from August 31st 2022, which is the end of the sampling period.
- Subreddit: The 15,000 comments in the labelled datasets are from 26 different subreddits, out of 104 subreddits initially selected for data collection. A large majority of 12,561 comments (83.7%) is from the r/indonesia subreddit, followed by 1,389 comments (9.3%) from r/malaysia, 272 comments (1.8%) from r/malaygonewild and 239 comments (1.6%) from r/Singapore. (Some duplicate comments were recognised after data preprocessing and the final data set contained **14983 unique comments**).
- 3 sampling strategies were used to get the data set:
  - Keyword sampling (9000 comments)
  - Active learning (4000 comments)
  - Random sampling (2000 comments)
- Language: 12,212 comments (81.4%) were majority-labelled as Indonesian, followed by 1,635 comments (10.9%) labelled as Malay and 218 comments (1.5%) were majority-labelled as Singlish. The remaining 688 (4.6%) comments were marked as containing one of dozens of other languages spoken in or around Singapore, such as Javenese and Hokkien Chinese, and code-mixed combinations thereof.

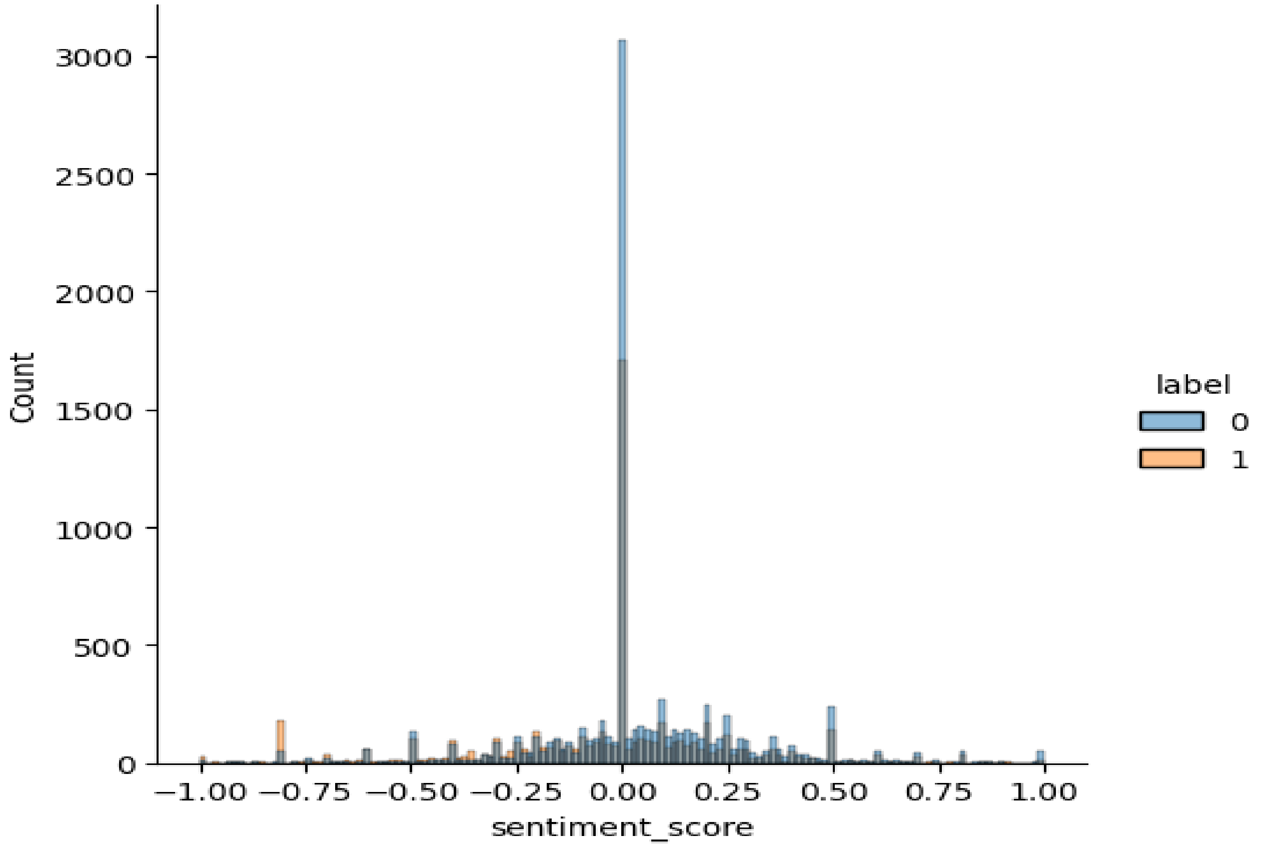
| | id | text | label | subreddit | created_utc | date | original_language(s) |
|---|---|---|---|---|---|---|---|
| **0** | sample_11379 | Oohh noo, Vicky Prasetyo is rumored to want to... | 0 | indonesia | 1641812360 | 10-01-2022 | Indonesian |
| **1** | sample_4296 | In my case, my family and friends never asked ... | 1 | indonesia | 1543136805 | 25-11-2018 | Indonesian |
| **2** | sample_6019 | When it was booming, I used to spend more than... | 1 | indonesia | 1580290245 | 29-01-2020 | Indonesian |
| **3** | sample_5254 | > In a response, he stated: "If there is a hou... | 0 | malaysia | 1642219040 | 15-01-2022 | Malay |
| **4** | sample_4120 | HEH ELU YES, GO FOR IT NGABBB BANDUNG COLD PER... | 0 | indonesia | 1640355228 | 24-12-2021 | Indonesian |

- Initially the data set contained columns namely, id, text, label, subreddit, created_utc, date, and original_language(s), which was cut down to get only text and label (**a label of 0 imply non-hate speech and a label of 1 imply a hate-speech**).
- Next, the text was analysed using 'Text-Blob' (a Python library that provides a simple API for common natural language processing (NLP) tasks, including sentiment analysis)
    - For sentiment analysis, TextBlob provides a pre-trained sentiment analyzer that can classify text into two main sentiment categories: positive and negative.
    - The sentiment analysis in TextBlob is based on a lexicon of words and their associated **sentiment scores**.
    - Each word in the text is assigned a polarity score (ranging from -1 to 1), which indicates whether the word is positive or negative.
    - The overall sentiment of the text is then calculated as the average of all the individual word polarity scores.
    - **We annotated a comment to be 'positive' if the sentiment score is greater than 0, 'negative' if the sentiment score was less than 0 and 'neutral' otherwise.**

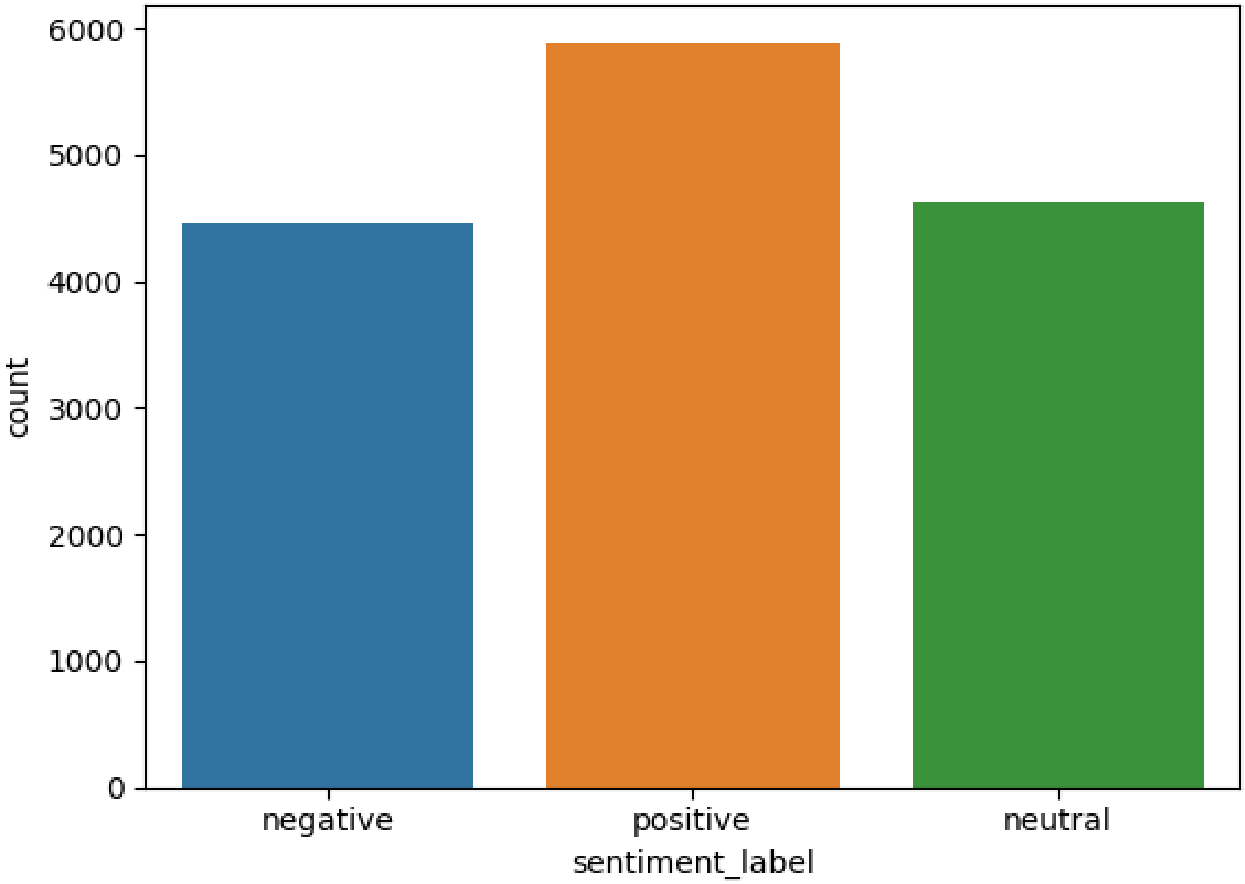
| | text | label | sentiment_score | sentiment_label |
|---|---|---|---|---|
| **0** | Oohh noo, Vicky Prasetyo is rumored to want to... | 0 | -0.750000 | negative |
| **1** | In my case, my family and friends never asked ... | 1 | -0.105000 | negative |
| **2** | When it was booming, I used to spend more than... | 1 | -0.075000 | negative |
| **3** | > In a response, he stated: "If there is a hou... | 0 | 0.085185 | positive |
| **4** | HEH ELU YES, GO FOR IT NGABBB BANDUNG COLD PER... | 0 | 0.200000 | positive |

- The hate-keyword lexicon had only a single column namely, 'Hate_keywords' and contained **1528 unique hate keywords**.
- Both the lexicon and the comment dataset were then pre-processed using 'nltk' library and then data analysis was done to get the number of hate keywords present in a particular comment along with the name of the keyword being used in that comment.
- **Total number of comments in the dataset labelled as 0 were 8811 and total number of comments labelled as 1 were 6172.**
- Using sentiment analysis for classification, the following results were obtained:

| Sentiment label | No. of Comments |
|---|---|
| positive | 5893 |
| neutral | 4632 |
| negative | 4458 |



**Sentiment score vs count of comments**
**(0 => non-hate comment, 1 => hate comment)**



| | Predicted 0 | Predicted 1 |
|---|---|---|
| **Actual 0** | 6648 | 2163 |
| **Actual 1** | 3877 | 2295 |

**Confusion Matrix**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| **0** | 0.63 | 0.75 | 0.69 | 8811 |
| **1** | 0.51 | 0.37 | 0.43 | 6172 |
| **accuracy** | | | 0.60 | 14983 |
| **macro avg** | 0.57 | 0.56 | 0.56 | 14983 |
| **weighted avg** | 0.58 | 0.60 | 0.58 | 14983 |

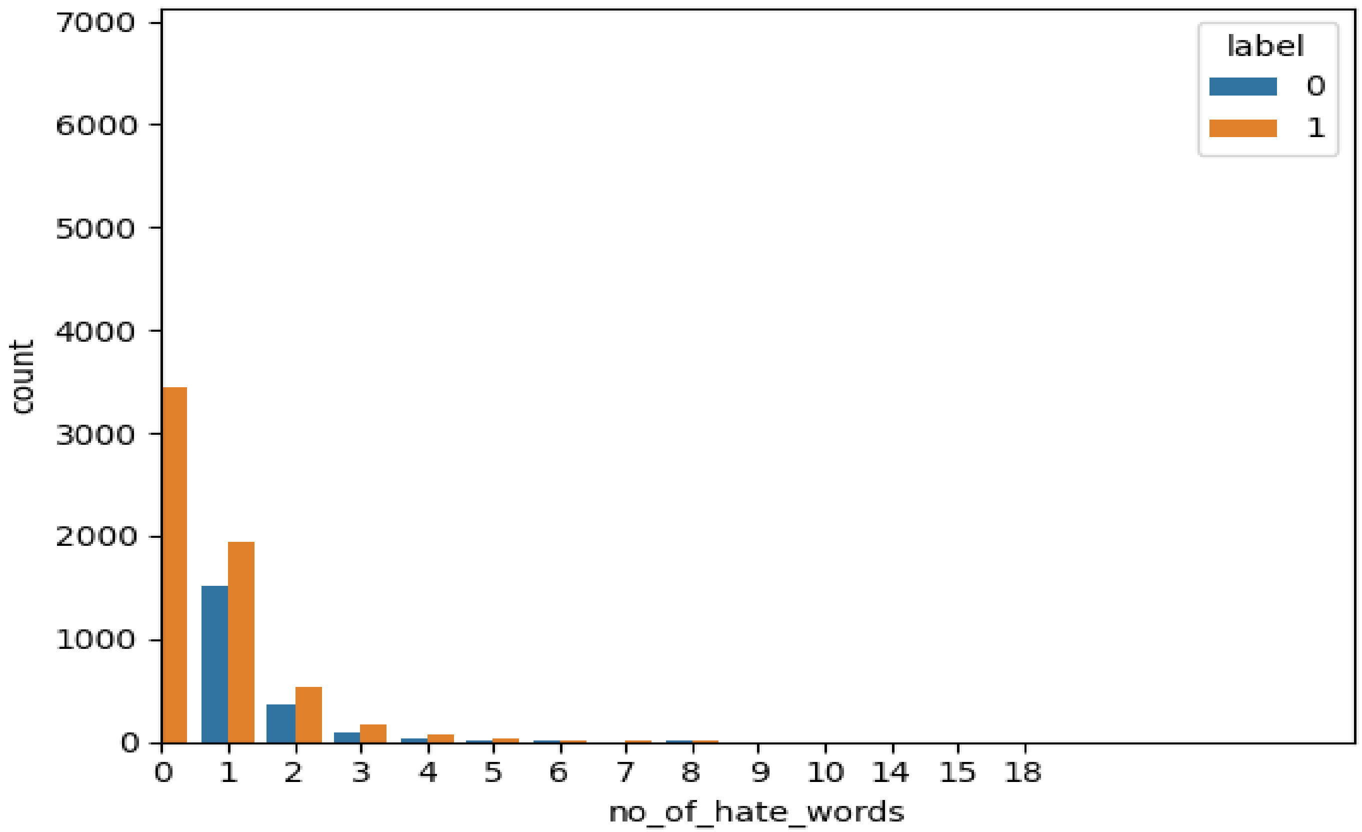
**Classification Report of labelling using sentiment Analysis**

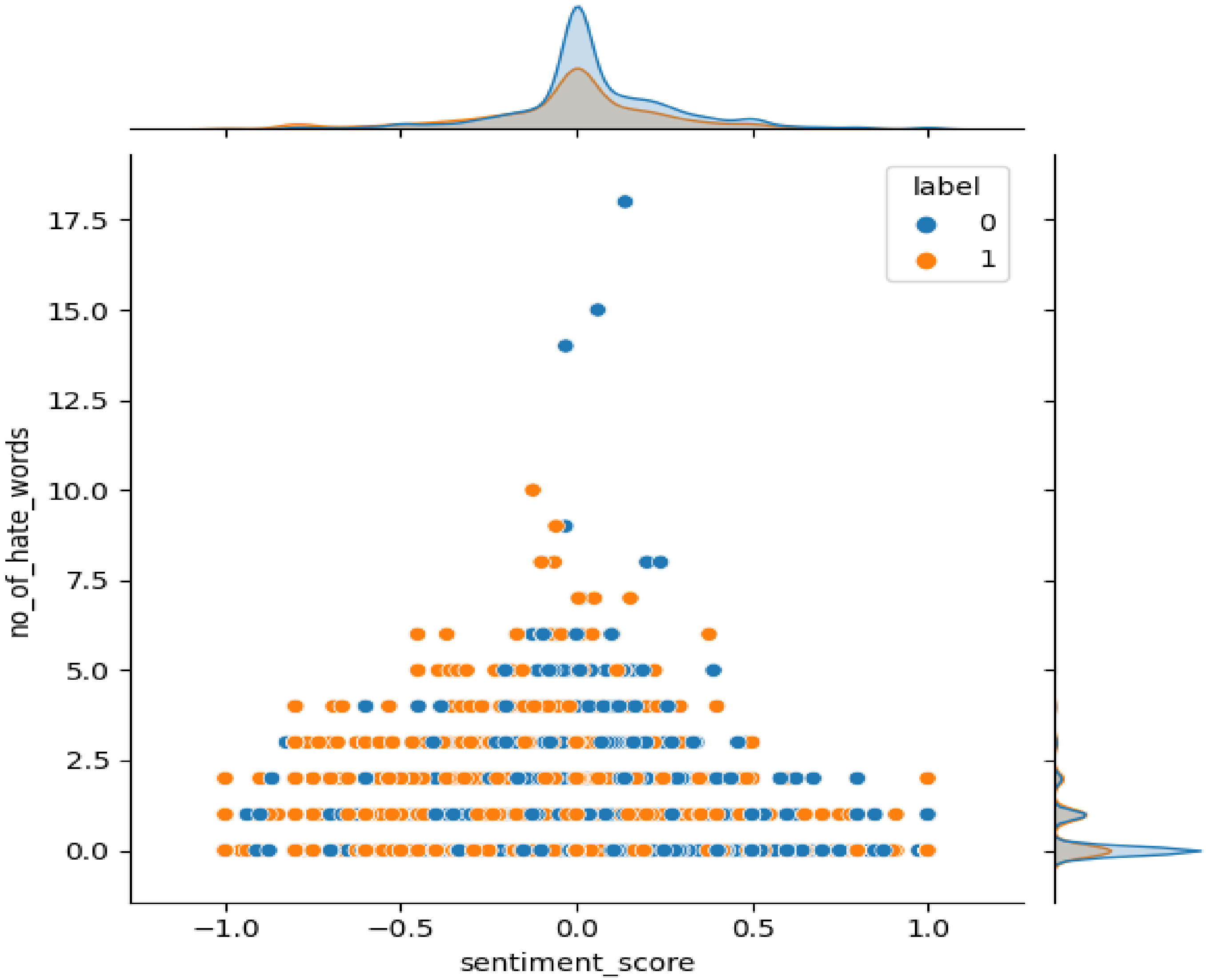
**(using TextBlob)**

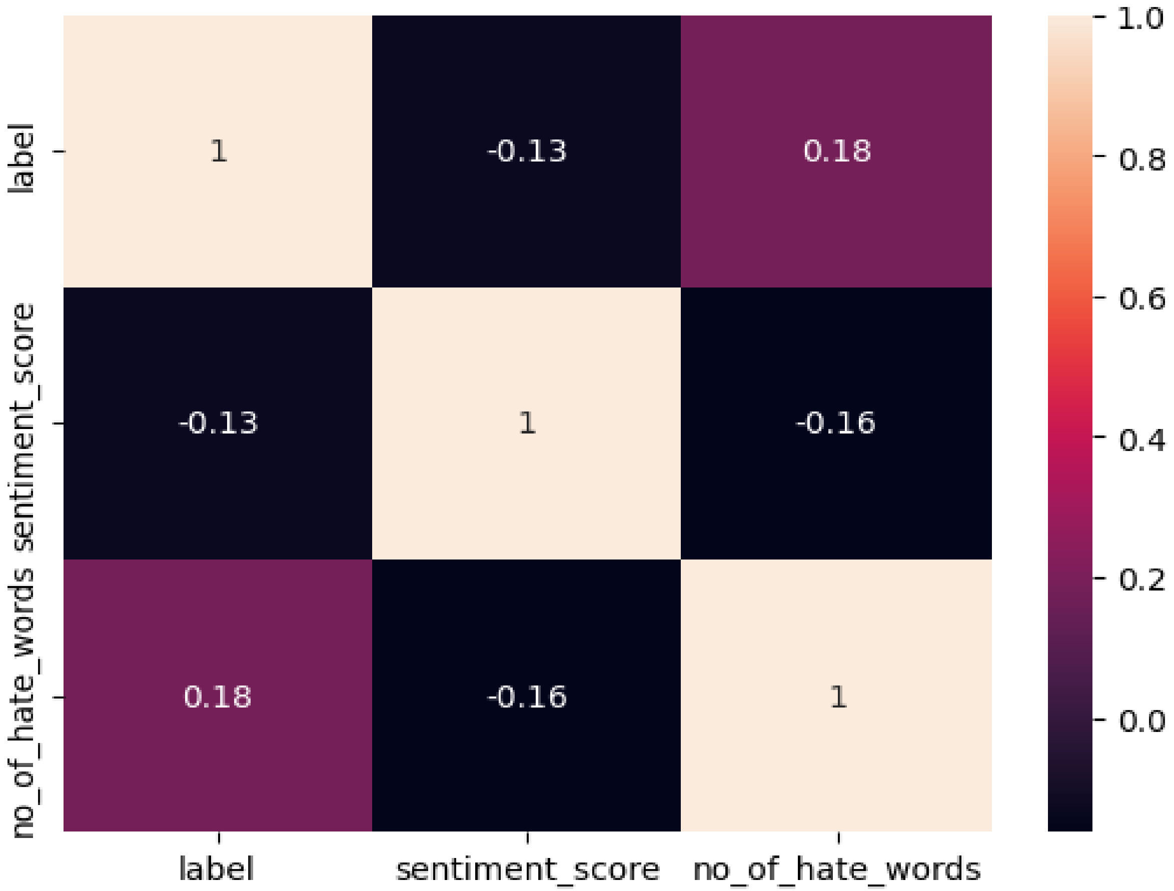
**NB: Log Loss value for classification using Sentiment Analysis = 14.530045149186916**

- Using number of hate keywords present in a particular comment for classification (no. of hate keywords present greater than zero implies the comment being hateful and being equal to zero implies non-hateful), the following results were obtained:

| Label | No. of hate words per comment | No. of comments |
|---|---|---|
| **0** | 0 | 6789 |
| | 1 | 1510 |
| | 2 | 351 |
| | 3 | 92 |
| | 4 | 37 |
| | 5 | 19 |
| | 6 | 6 |
| | 7 | 1 |
| | 8 | 2 |
| | 9 | 1 |
| | 14 | 1 |
| | 15 | 1 |
| | 18 | 1 |
| **1** | 0 | 3441 |
| | 1 | 1937 |
| | 2 | 538 |
| | 3 | 158 |
| | 4 | 60 |
| | 5 | 21 |
| | 6 | 10 |
| | 7 | 3 |
| | 8 | 2 |
| | 9 | 1 |
| | 10 | 1 |


label
0
1
count
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9
10
14
15
18
no_of_hate_words


label
0
1
no_of_hate_words
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−1.0
−0.5
0.0
0.5
1.0
sentiment_score



| | Predicted 0 | Predicted 1 |
|---|---|---|
| **Actual 0** | 6648 | 2163 |
| **Actual 1** | 3877 | 2295 |

**Confusion Matrix**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| **0** | 0.66 | 0.77 | 0.71 | 8811 |
| **1** | 0.57 | 0.44 | 0.50 | 6172 |
| **accuracy** | | | 0.64 | 14983 |
| **macro avg** | 0.62 | 0.61 | 0.61 | 14983 |
| **weighted avg** | 0.63 | 0.64 | 0.63 | 14983 |

**Classification Report of labelling using hate_keywords**

**NB: Log Loss value for classification using Sentiment Analysis = 13.141992822849028**

- At the end we also produced a table containing the distribution of a particular hate keyword into positive, neutral and negative sentiments as assigned using 'TextBlob' that is, if a particular comment is assigned to be negative then the hate keyword present in the comment is tracked and the frequency of the same keyword in 'negative' column increases. Similarly, any hate keyword present in a positive or neutral sentiment comment is traced and increment in frequency is done accordingly. Below is a sample showing some hate keywords being used more often.

| hate_keywords | negative | neutral | positive | Total |
|---|---|---|---|---|
| **No hate keywords** | 2228 | 3855 | 4147 | 10230 |
| **stupid** | 771 | 3 | 113 | 887 |
| **fuck** | 281 | 7 | 75 | 363 |
| **eat** | 127 | 59 | 176 | 362 |
| **crazi** | 209 | 2 | 62 | 273 |
| **bastard** | 77 | 91 | 100 | 268 |
| **idiot** | 220 | 0 | 25 | 245 |
| **chines** | 90 | 30 | 80 | 200 |
| **god** | 42 | 25 | 100 | 167 |
| **shit** | 93 | 1 | 60 | 154 |

# Discussions & Results

- One of the noticeable fact which can be inferred from the above table is that even if a particular hate keyword is present in a negative sentiment (annotated by 'TextBlob') there are some considerable positive comments which accounts for the same keyword as well. This can be explained with an example, "I hate racist people". The example is having a positive sentiment but contains a hate keyword 'racist' which explains the frequency of 'racist' keyword in positive sentiment category. Similar, is the case with neutral sentiment as well.
- Considering the data set to be uniformly distributed, both the F1 score and accuracy of predicting a comment as 0 (non-hateful / non-offensive) or 1 (hateful / offensive) was greater if we use hate keyword dataset to identify the class rather than using sentiment analysis provided by 'TextBlob'.
- And the fact to be considered is the accuracy and F1 score may increase if the size of the lexicon containing hate keywords is increased considerably (current size being 1528 unique keywords). There is an equal chance of it being decreased because of the above said example (hate keywords present in positive sentiment comment). So, we can infer that using hate keyword to detect hate speech depends only on the word under consideration that is being used in a comment and hence may give highly inaccurate results because of absence of the context of the comment (refer above example).