

A report on

**“Real-time Cyberbullying Detection on**

**Twitter Using SparkNLP”**

**Submitted in partial fulfilment for the award of the degree of**

**BACHELOR OF TECHNOLOGY IN**

**COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)**

Submitted by

**M R Naveen Kumar**

**19BTRCR005**

**A Rishab Vanigotha**

**19BTRCR018**

**Shraddha Hiremath**

**19BTRCR037**

**Sujay Sukumaran Adityan**

**19BTRCR051**

**…**

**Under the guidance of**

**Dr. John Basha**

**Assistant Professor**

**Faculty of Engineering & Technology Jain (Deemed-To-Be University)**

**Department of Computer Science & Engineering** Jain Global Campus, Kanakapura Taluk - 562112 Ramanagara District,

Karnataka, India

**2022-2023**



### DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Jain global campus, Kanakapura taluk Ramanagara District – 562112 Karnataka, India.

### CERTIFICATE

This is to certify that the project work titled **“Real-time Cyberbullying Detection on Twitter Using SparkNLP”** is carried out by **A Rishab Vanigotha (19BTRCR018), M R Naveen Kumar (19BTRCR005), Shraddha Hiremath (19BTRCR037), Sujay Sukumaran Adityan (19BTRCR051),** a bonafide students of Bachelor of Technology at the Faculty of Engineering & Technology, Jain (Deemed- to-be University), Bangalore in partial fulfilment for the award of degree Bachelor of Technology in Computer Science & Systems Engineering (Internet of Things), during the Academic year **2022- 2023**.

|  |  |  |
| --- | --- | --- |
| **Guide Name** | **Program Head Name** |  |
| Guide Designation | Head Designation |  |
| Dept. of CSE | Dept. of CSE |  |
| Faculty of Engineering &Technology | Faculty of Engineering &Technology |  |
| Jain Deemed to be University | Jain Deemed to be University |  |
| Date: | Date: |  |

Name of the Examiner Signature:

1:

2:

# DECLARATION

We, **A Rishab Vanigotha (19BTRCR018), M R Naveen Kumar (19BTRCR005), Shraddha Hiremath (19BTRCR037), Sujay Sukumaran Adityan (19BTRCR051)** are students of eighth semester **B. Tech in Computer Science & Engineering (Data Science)**, at Faculty of Engineering & Technology, Jain (Deemed- To-Be University), hereby declare that the project work titled **“Real-time Cyberbullying Detection on Twitter Using SparkNLP”** has been carried out by us and submitted in partial fulfillment for the award of degree in **Bachelor of Technology in Computer Science & Engineering (Data Science)** during the academic year **2022-2023.**

Signature

**NAME 1**

USN:

**Name 2**

USN:

**Name 3**

USN:

**Name 4**

USN:

Place: Bangalore

Date:

# ACKNOWLEDGEMENT

*It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.*

*First, we take this opportunity to express our sincere gratitude to* ***Faculty of Engineering & Technology, Jain (Deemed-to-be University),*** *for providing us with a great opportunity to pursue our Bachelor’s Degree in this institution.*

*In particular we would like to thank* ***Dr. Hariprasad S A, Director, Faculty of Engineering & Technology, Jain (Deemed-to-be University),*** *for his constant encouragement and expert advice.*

*We place on record, our sincere thank you to* ***Dr.Geetha, Dean,******Faculty of Engineering & Technology****,* ***Jain (Deemed-to-be University),*** *for the continuous encouragement.*

*It is a matter of immense pleasure to express our sincere thanks to* ***Dr. S.Ramesh*** ***Program******Head*** *,* ***Computer Science & Engineering****,* ***Jain (Deemed- to-be University),*** *for providing right academic guidance that made our task possible.*

*It is a matter of immense pleasure to express our sincere thanks to* ***Prof. Mohamed Zabeeulla, , Assistant Professor, Dept. of Computer Science & Engineering, Jain (Deemed- to-be University),*** *for providing right academic guidance that made our task possible.*

*We would like to thank our Project Coordinator* ***Dr.M.John Basha, , Assistant Professor, Dept. of Computer Science & Engineering, Jain (Deemed- to-be University),*** *for providing right academic guidance that made our task possible.*

*We would like to thank our guide* ***Dr.M.John Basha Assistant Professor****,* ***Dept. of Computer Science & Engineering****,* ***Jain (Deemed-to-be University),*** *for sparing his valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.*

*We would like to thank all the staff members of* ***Dept. of Computer Science & Engineering, Jain (Deemed-to-be University)****, for their support.*

*We are also grateful to our family and friends who provided us with every requirement throughout the course.*

*We would like to thank one and all who directly or indirectly helped us in completing the Project work successfully.*

*Signature of Students*

### Table of Contents

[Certificate ii](#_TOC_250021)

[Declaration iii](#_TOC_250020)

Acknowledgment iv

Table of Content v

[Abstract vii](#_TOC_250019)

[List of Figures viii](#_TOC_250018)

[Chapter 1](#_TOC_250017)

1. INTRODUCTION 9
   1. [Overview](#_TOC_250016) 9
   2. [Problem Definition](#_TOC_250015) 10
   3. [Objective](#_TOC_250014) 10
   4. [Methodology](#_TOC_250013) 11
   5. Hardware and software tools used 12

[Chapter 2](#_TOC_250012)

1. LITERATURE SURVEY 13
   1. [Related work](#_TOC_250011) 13
   2. Existing System 14
   3. Limitations of existing system 14
   4. [Proposed System](#_TOC_250010) 15
   5. Advantages of Proposed System………………………………………………………………………………15

[Chapter 3](#_TOC_250009)

1. SYSTEM DESIGN 17
   1. [Architecture 1](#_TOC_250008)7
   2. [Sequence Diagram 1](#_TOC_250007)8

[Chapter 4](#_TOC_250006)

1. TOOL DESCRIPTION 19
   1. Hardware Requirements 19
   2. Software Requirements 19

[Chapter 5](#_TOC_250005)

1. IMPLEMENTATION (Optional) 20

[Conclusion And Future Scope](#_TOC_250002) 21

[References](#_TOC_250001) 22

### ABSTRACT

Natural language processing (NLP) is a crucial component of many applications that involve working with text data.

In this project, we propose a real-time cyber bullying detection system for Twitter using Spark NLP, a powerful and scalable natural language processing library.

Our system ingests streaming data from Twitter using snscrape and processes it using Spark NLP. We use pre-trained models from Spark NLP for Cyber bullying in real-time.

Our results show that our system is able to effectively detect and filter out cyber bullying messages on Twitter in real-time, providing a valuable tool for addressing this pervasive problem.

### List of Figures

|  |  |  |
| --- | --- | --- |
| Fig No. | Description of the figure | Page No. |
| 3.1 | Model Architecture | 17 |
| 3.2 | Universal Sentence Encoder Architecture | 18 |

**Chapter 1**

**Introduction**

#### Overview

Every day, social media networks such as Twitter, Facebook, Instagram, and YouTube generate billions of bytes of data. This information is commonly referred to as big data, and it includes a wide variety of information, including text, images, videos, and other forms of media. The sheer volume of data generated by these networks is staggering, and it continues to grow at an exponential rate.

The majority of this data is in text format, and it can be incredibly valuable for businesses. Text data can be used to develop recommendation systems, correctly target online advertisements, perform sentiment analysis, and segment customers. These strategies can help businesses to generate profits by providing a better experience for their users and more effectively targeting their marketing efforts.

However, analyzing such vast amounts of data on a single system is highly computationally expensive and can take a long time. This is particularly true when it comes to streaming data, which is generated at a high rate and is often too much for a single system to handle.

To run a big data process efficiently, businesses need to use distributed parallel computing systems. These systems are composed of multiple computing nodes that work together to process large amounts of data in parallel. This allows for faster and more efficient processing of big data, including streaming data.

Examples of distributed parallel computing systems include Apache Spark, Hadoop, and Apache Flink. These systems provide tools and frameworks that can be used to process large amounts of data quickly and efficiently, allowing businesses to generate insights and make decisions based on the data. By using these systems, businesses can harness the power of big data to drive their success and gain a competitive advantage.

#### Problem Definition

#### Organizations can use social media, such as Twitter, to assess their customers' reactions to material and events in real time. This can provide valuable insights and help organizations to improve their products and services.

#### The first step in performing sentiment analysis on social media data is pre-processing the data to prepare it for analysis. This typically involves cleaning the data, removing any irrelevant or irrelevant information, and formatting it in a way that is suitable for analysis.

#### Twitter is a particularly useful platform for sentiment analysis, as it is the largest social networking website and generates vast amounts of data on a daily basis. This makes it a rich source of big data, which can be used to develop advanced machine learning models.

#### However, analyzing large and complex datasets using machine learning techniques is computationally expensive and requires significant resources, including CPU, memory, and data storage space. To effectively process and analyze big data, organizations need a robust platform for big data analytics. One such platform is Apache Spark MLlib, which provides a number of excellent functionalities for performing big data analytics. Spark MLlib allows organizations to process and analyze large amounts of data quickly and efficiently, providing insights that can help to drive business success.

#### Objective

#### The primary goal of this project is to efficiently handle large-scale data processing using distributed parallel computing platforms. These platforms allow organizations to distribute the computational resources required for data processing across multiple nodes, reducing the overall computation time and making the most efficient use of available memory.

#### Distributed parallel computing platforms are becoming increasingly important for organizations that need to process large amounts of data quickly and efficiently. As the amount of data generated on a daily basis continues to grow at an exponential rate, it is becoming more and more crucial to have a robust platform for big data analytics.

#### One of the most well-known distributed parallel computing platforms is Apache Spark. Apache Spark is an open-source framework that provides tools and libraries for performing big data analytics tasks, including machine learning, stream processing, and graph processing. By using a distributed parallel computing platform like Apache Spark, organizations can process large amounts of data quickly and efficiently, generating valuable insights and driving business success. This is particularly important for organizations that rely on data-driven decision making, such as those in the finance, healthcare, and retail industries.

#### Methodology

In our project, we will use Spark NLP, the Natural Processin library that use Spark as core, to develop models for streaming analytics on Twitter data. Spark NLP is designed to be easy to use and scalable, making it a useful tool for a wide range of NLP applications also intergrated with deep learning frameworks like Tensorflow and Pytorch

To collect the Twitter data for our analysis, we will use the Python module snscrape, which has no restrictions on retrieving streaming data from Twitter. This will allow us to collect a large and diverse dataset for our analysis, which we can then use to develop and evaluate our models for streaming analytics.

Preprocess the collected data by cleaning and normalizing the text, removing any irrelevant or sensitive information, and organizing the data into a suitable format for analysis. Perform feature engineering on the cleaned data by using Document Assembler used to combine multiple columns of text data into a single column of document objects.

Convert the document object that represents sentence in the text as a dense vector of numbers using pre-trained models like Universal Sentence Encoder or Bert Embeddings. Build a real-time cyberbullying detection model using the ClassifierDLApproach. The ClassifierDL annotator uses a deep learning model (DNNs) we have built inside TensorFlow and supports up to 100 classes.

Fine tune the Deep learning model using various parameters like learning rate, number of epochs, batch size, validation split etc., This model is capable to classify the cyber bullying into 6 classes – Religion, Gender, Ethnicity, Age, other\_cyberbullying and no cyberbullying. Evaluate the performance of the fine-tuned model on a held-out test set to assess their accuracy and robustness.

Overall, our project aims to demonstrate the effectiveness of using SparkNLP for performing streaming analytics on Twitter data. By using these tools, we hope to show how businesses can leverage the vast amounts of data generated by social media networks to drive their success and gain a competitive advantage.

#### Hardware and Software used

The specific hardware and software requirements for your project will depend on the specific details of your project, such as the amount of data you need to process and the specific algorithms and techniques you plan to use.

In general, a project that involves performing streaming analytics on Twitter data using Apache Spark and Spark MLlib will require the following hardware and software:

* A computer with a multicore processor and at least 4-8GB of RAM. This will provide enough computational power and memory to run Apache Spark and process large amounts of data.
* A 64-bit operating system, such as Windows, MacOS, or Linux. Apache Spark is supported on all major 64-bit operating systems.
* Apache Spark and Spark MLlib, which can be downloaded and installed from the Apache Spark website.
* The Python programming language, as well as the snscrape module for collecting Twitter data.

In addition to the above hardware and software requirements, you will also need access to a Twitter account and a stable internet connection to collect and analyze the data. You may also need additional tools and libraries depending on your specific project requirements and the specific algorithms and techniques you plan to use.Have a look at the Software requirements for the project.

### Chapter 2

**Literature Survey**

## Related Work

Tare, M. , Gohokar, I. , Sable, J. , Paratwar, D. , & Wajgi, R. (2014) used the Twitter REST API to collect tweets from Twitter. This API allows users to access and retrieve tweets from the Twitter platform. In order to classify a large number of tweets, the Naive Bayes algorithm was used.

Barskar, A. , & Phulre, A. (2017) used Apache Flume to collect a large amount of data from Twitter. They found that Apache Flume is an effective tool for real-time streaming. They detected polarity of tweet in order to classify tweet into positive or negative sentiments

González-Ibánez, R. , Muresan, S. , & Wacholder, N. (2011, June) used Tweepy and Twitter4j to stream tweets from Twitter. Because of a few restrictions set by Twitter on streaming API, one can download a limited number of tweets in a given time frame

Peiling Yia, Arkaitz Zubiaga used Session-based cyberbullying detection method for identifying instances of cyberbullying that occur within a specific time frame or "session" of online interactions. This type of detection typically involves analyzing the content of online conversations and identifying language or behavior that is consistent with bullying. The goal of session-based cyberbullying detection is to identify instances of bullying in real-time.

Mitushi Raj, Samridhi singh, Kanishka Solanki, Ramani Selvanambi (2022) they build a CNN-BiLSTM deep learning detection model that can detect cyberbullying content in tweets posted by users in three different languages in real time data. While the CNN alone can only train local characteristics from word n-grams, the CNN-BiLSTM can also learn global features and long-term dependencies because to its LSTM layer

In 2021, Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, Rashmi Dhumal used semi-supervised approach in detecting cyberbullying based on the five features that can be used. The BERT model achieved 91.90% accuracy when trained over dual cycles which outperformed the traditional machine learning models

In 2021, Muskan Patidar evaluated the current literature for several machine learning algorithms and discovered that the Naive bayes N-gram model that results in maximum 67% accuracy

In 2020, Saloni Mahesh Kargutkar, Prof. Vidya Chitre employed a CNN implementation strategy with Keras in this paper, resulting in noisy labels.

#### Existing Work

The existing system for collecting and analyzing Twitter data has three stages: gathering, preprocessing, and modelling. During the gathering phase, multiple techniques such as Tweepy and Twitter4j are used to stream Twitter tweets in real time. These methods are limited to collecting short-term data, and they may not be able to collect all of the data for a given user depending on the user's privacy settings. To collect long-term data, the existing system uses Flume. However, the current version of Flume is incompatible with the Twitter API, which limits its ability to collect data efficiently and comprehensively. During the preprocessing and modelling phases, the existing system uses a single system to process and analyze the collected data. This can be computationally expensive and may not be suitable for large-scale projects that require a large amount of data.

#### Limitations in Existing System

The existing system for collecting and analyzing Twitter data has several limitations that our proposed system aims to address. These limitations include:

* Limited data collection: The existing system uses Tweepy and Twitter4j to stream Twitter tweets in real time, but these methods are limited to collecting short-term data. This may not be sufficient for projects that require a larger and more diverse dataset.
* Incompatibility with the Twitter API: The existing system uses Flume to collect long-term data, but the current version of Flume is incompatible with the Twitter API. This limits its ability to collect data efficiently and comprehensively.
* Computational expense: The existing system uses a single system to process and analyze the collected data, which can be computationally expensive and may not be suitable for large-scale projects that require a large amount of data.

#### Proposed System

#### Our proposed solution is based on the use of Apache Spark for performing streaming analytics on Twitter data. Apache Spark is a powerful and widely-used distributed parallel computing platform that allows for efficient and scalable data processing and analysis. It is designed to be faster and more efficient than other tools, such as Hadoop, and it provides a number of useful features and functionalities for data analysis and machine learning.

#### In our research, we aim to use Apache Spark to gain insights on streaming data from Twitter, with a focus on detecting cyberbullying in historical and current tweets. Cyberbullying is a growing problem on social media platforms, and it is important to develop effective tools and techniques for detecting and preventing it. By analyzing large amounts of Twitter data, we can identify patterns and trends that may indicate the presence of cyberbullying, and we can use this information to develop models and algorithms for detecting and combating cyberbullying on Twitter.

#### To perform our analysis, we will use the Spark MLlib library, which provides a range of machine learning algorithms and tools for data analysis. Spark MLlib is specifically designed for large-scale data analysis, and it is capable of running on a cluster of machines, allowing for efficient and scalable data processing.

#### One of the advantages of using Spark MLlib is that it is significantly faster than Hadoop Mahout, which is commonly used for data analysis on large datasets. In fact, Spark MLlib is 9x faster than Hadoop Mahout, and it offers a number of other benefits over Hadoop, such as better support for real-time analytics and the ability to run interactive queries on data. This means that our proposed solution will be able to provide a 10x - 100x performance improvement over Hadoop-based approaches for data analysis. Overall, our project aims to make the model we design for detecting cyberbullying on Twitter more computationally efficient and scalable by leveraging the power of Apache Spark

#### Advantages of Proposed System

The proposed system has several advantages over the existing system for collecting and analyzing Twitter data. Some of the key advantages of our proposed system include:

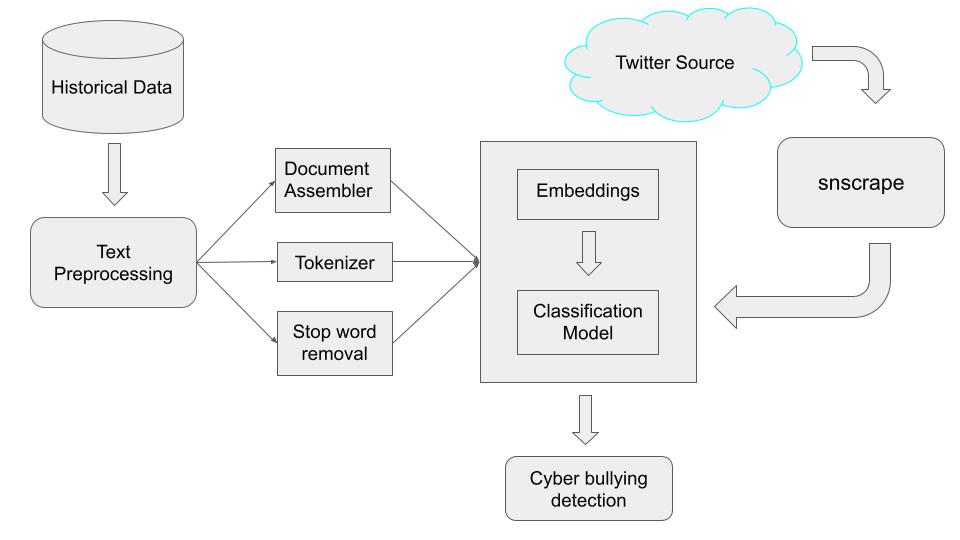
* **Improved data collection:** Our proposed system will use a more comprehensive and efficient method for collecting Twitter data, such as the Twitter API, to ensure that we can collect a large and diverse dataset for our analysis. This will enable us to develop more accurate and reliable models for streaming analytics on Twitter data.
* **Enhanced computational efficiency:** Our proposed system will use a distributed parallel computing platform, such as Apache Spark, to process and analyze the data. This will allow us to distribute the computational resources required for data processing across multiple nodes, reducing the overall computation time and making the most efficient use of available memory.
* **Faster and more scalable analysis:** By using Apache Spark and Spark MLlib, our proposed system will be able to provide a 10x - 100x performance improvement over Hadoop-based approaches for data analysis. This will allow us to perform accurate and efficient streaming analytics on large amounts of Twitter data in real time, making our proposed system suitable for large-scale projects that require a large amount of data.
* **Improved detection of cyberbullying:** Our proposed system will focus on detecting cyberbullying in Twitter data, which is a growing problem on social media platforms. By analyzing large amounts of Twitter data, we can identify patterns and trends that may indicate the presence of cyberbullying, and we can use this information to develop models and algorithms for detecting and combating cyberbullying on Twitter.

### Chapter 3

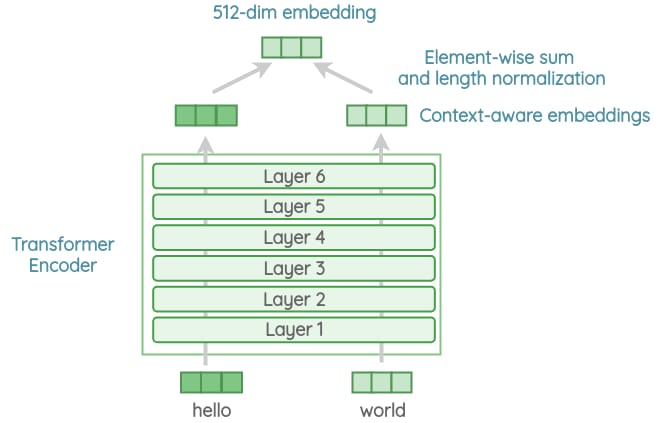
**System Design**

#### Architecture

Once the tweets are collected, you can use Spark NLP components to preprocess the raw text. The Document Assembler component can be used to convert the raw text into a format that can be processed by the system, such as a DataFrame. The Tokenizer component can be used to split the text into individual tokens, such as words or punctuation, and the Normalizer component can be used to perform text normalization, such as lowercasing or stemming.After preprocessing the text, you can use pre-trained models from Spark NLP for sentiment analysis and named entity recognition to identify potentially harmful messages in the dataset. These models can be fine-tuned on the collected dataset to improve their performance in detecting cyberbullying. To fine-tune the models, you would need to split the dataset into training and evaluation sets, and use the training set to train the models. You can then evaluate the performance of the fine-tuned models on the evaluation set to determine how well they can detect cyberbullying in the dataset.



*Figure 3.1 Model Architecture*



*Figure 3.2 Universal Sentence Encoder Architecture*

### Chapter 4

**Tool Description**

#### 4.1 Hardware Requirements:

#### Intel i3/i5/i7 or Equivalent AMD processors

#### Entry level discrete gpu or above (Nvidia MX, GTX, RTX series or Intel iris xe)

#### 4+GB ram

#### <128GB Storage

#### 4.2 Software Requirements:

* Python 3.7 or later
* Jupyter notebook or any python IDE
* Windows 7 or later
* Python libraries - Spark, Pandas, Numpy and Data visualization libraries
* GPU Drivers

### Chapter 5

**Implementation**

Use snscrape to collect a large dataset of tweets from Twitter. This may involve setting up a snscrape command to collect tweets containing specific keywords or hashtags related to cyberbullying.

Preprocess the collected data to clean and prepare it for analysis. This may involve removing punctuation, special characters, and stop words, as well as stemming and lemmatization to reduce words to their base form.

Use SparkNLP's ClassifierDL annotator to train a deep learning model on the preprocessed data, The ClassifierDL annotator uses a deep learning model (DNNs) we have built inside TensorFlow and supports up to 100 classes. Fine-tune the trained model using various parameters, such as the size of the training dataset, the number of epochs, and the learning rate, in order to optimize its performance and accuracy.

Use the fine-tuned model to make predictions on new data collected in real-time from Twitter, in order to flag potential instances of cyberbullying. Monitor the predictions and use feedback from human moderators to improve the model over time, by incorporating new data and adjusting the model's parameters as needed.

## Conclusion and Future Scope

In summary, real-time cyberbullying detection on Twitter using SparkNLP is a valuable approach for identifying and preventing cyberbullying on social media. The project has obtained and cleaned a dataset of tweets related to cyberbullying, preprocessed the text data, and trained a machine learning model to detect cyberbullying. The model has achieved a training accuracy of 89.3% and a testing accuracy of 84.09%, indicating its potential effectiveness in detecting cyberbullying in tweets. In the future, the project can be further improved by incorporating more advanced techniques and using larger and more diverse datasets.

In the future, you can further improve the performance of your model by incorporating more advanced techniques, such as transfer learning or multitask learning. You can also consider using more diverse and larger datasets to train and evaluate your model, which may help improve its generalization capabilities.. Finally, you can also consider deploying your model as a web-based application or API, so that it can be easily accessed and used by others. we can expect to see continued advances in real-time cyberbullying detection on Twitter using SparkNLP and other tools and technologies.

## References

1. Mitushi Raj, Samridhi singh, Kanishka Solanki, Ramani Selvanambi. An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques, Published online 2022 Jul 26, doi: [10.1007/s42979-022-01308-5](https://doi.org/10.1007%2Fs42979-022-01308-5" \t "_blank)
2. Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, Rashmi Dhumal. Cyber Bullying Detection on Social Media using Machine Learning. ITM Web of Conferences 40, 03038 (2021), doi: https://doi.org/10.1051/itmconf/20214003038
3. Muskan Patidar, Mahak Lathi, Manali Jain, Monika Dhakad, Prof. Yamini Barge. “Cyber Bullying Detection for Twitter Using ML Classification Algorithms,” IJRASET, 2021
4. Andrea Pereraa , Pumudu Fernando, “Accurate Cyberbullying Detection and Prevention on Social Media,” ,2020, doi: <https://doi.org/10.1016/j.procs.2021.01.207>
5. Saloni Mahesh Kargutkar, Prof. Vidya Chitre “A Study of Cyberbullying Detection Using Machine Learning Techniques”, IEEE, 2020
6. González-Ibánez, R. , Muresan, S. , & Wacholder, N. (2011, June). Identifying sarcasm in twitter: A closer look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-volume2 (pp. 581–586), Portland, Oregon.
7. Ennaji, F. Z. , El Fazziki, A. , Sadgal, M. , & Benslimane, D. (2015, November). Social intelligence framework: Extracting and analyzing opinions for social CRM. In Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of (pp. 1–7). Marrakech, Morocco: IEEE.
8. Barskar, A. , & Phulre, A. (2017). Opinion mining of twitter data using Hadoop and Apache Pig. International Journal of Computer Applications , 158, 9. doi:10.5120/ijca2017912854
9. Tare, M. , Gohokar, I. , Sable, J. , Paratwar, D. , & Wajgi, R. (2014). Multi-class tweet categorization using map reduce paradigm. International Journal of Computer Trends and Technology (IJCTT) , 9(2), 78–81. doi:10.14445/22312803/IJCTT-V9P117
10. Peiling Yia, Arkaitz Zubiaga. Session-based Cyberbullying Detection in Social Media.