**Regional Language Toxic comment classification**

**Technical Report**

Yashkumar Parikh

parikhy@lakeheadu.ca

Student ID No.: 1138765

# **Abstract**

**Social media sites are gaining popularity day by day. They are best for communication, business, entertainment, and many other things. After more than a decade, social media have become very influential. On the flip side, fake news, hate speech, and online trolls are the biggest concerns because of social media. So, a solution to curb this issue is needed, especially in regional languages. Many social media platforms support regional languages. This paper will provide a machine learning-based solution to this problem. The focus of this paper is to classify comments written in regional languages. Firstly, a dataset has been created in Gujarati, Hindi, English, Marathi, and Punjabi languages. After that, different machine learning and deep learning models are applied to the multilingual dataset. At last, a comparison of all model performances was made.**

# **Steps to Execute Project**

1. Unzip the project folder
2. Upload the project folder (Data and Code files) to google drive
3. Open ipynb file and change ‘dir\_path’ variable to your local path
4. Execute the code cells

# **Dataset**

I have gathered data from various social media platforms in Indic languages like Gujarati, Hindi, Marathi, Punjabi, and English. The shape of the data is (14995, 2). It has two columns comment\_text and toxic. The comment\_text column contains actual text, and the toxic column contains '1' for toxic and '0' for non-toxic comments. The pre-processing step removes URLs, hashtags, mentions, punctuations, and extra white spaces from the comments. Plus, removed rows with empty or null values.

# **Code Editor**

**Google Colab**

Colaboratory, sometimes called Colab, is a Google Research product. Colab is particularly well suited to machine learning, data analysis, and education. It enables anyone to create and execute arbitrary Python code through the browser. Technically speaking, Colab is a hosted Jupyter notebook service that offers free access to computer resources, including GPUs, and requires no setup [2].

# **Code**

In code section I have used five different machine learning algorithms.

1. **Logistic Regression**

**Graphical user interface, application

Description automatically generatedGraphical user interface, text

Description automatically generated with medium confidence**

**Graphical user interface, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface

Description automatically generated with medium confidenceGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated**

1. **Support Vector Machine**

**Chart

Description automatically generated**

**Graphical user interface, table

Description automatically generated with medium confidenceGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, text, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated**

1. **LSTM**

**Diagram

Description automatically generated**

**Text

Description automatically generated with low confidenceText

Description automatically generated with medium confidenceGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generatedA picture containing graphical user interface

Description automatically generatedGraphical user interface

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated**

1. **CNN**

Diagram

Description automatically generated

Graphical user interface, text, application

Description automatically generatedTable

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface

Description automatically generated with medium confidenceGraphical user interface

Description automatically generated with low confidenceGraphical user interface

Description automatically generated with medium confidenceGraphical user interface, text, application, email

Description automatically generated

1. **DistilBERT**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedText

Description automatically generatedGraphical user interface, text, application

Description automatically generatedText

Description automatically generatedText, email

Description automatically generatedTable

Description automatically generated with low confidenceA picture containing graphical user interface

Description automatically generatedGraphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedText

Description automatically generated with low confidence**

# **Libraries**

1. **TensorFlow**

TensorFlow is an end-to-end open-source platform for machine learning. Researchers can advance the state-of-the-art in machine learning thanks to its extensive, adaptable ecosystem of tools, libraries, and community resources. At the same time, developers can create and deploy ML-powered applications [4].

To undertake machine learning and deep neural network research, the Google Brain team, a group of researchers and engineers within Google's Machine Intelligence Research division, created TensorFlow. The system is broad enough to work in several additional domains.

1. **Keras**

Keras is an open-source high-level Neural Network library, and it is written in Python. Francois Chollet, a Google developer, created it. It is user-friendly, expandable, and modular, enabling quicker experimentation with deep neural networks. It supports both Convolutional and Recurrent Networks separately as well as in combination. [3]

The sequential method of TensorFlow groups a linear stack of layers into a tf.keras.Model. So, in the below image Sequential method has grouped one embedding layer: one LSTM layer and two Dense layers. Kernel Regularization is used to apply penalties on layer parameters or layer activity during optimization. Because of regularization, the model does not overfit.

The Compile method configures the model for training. The Loss function and optimizers are defined in this method.

Graphical user interface, text, application, email

Description automatically generated

The Early Stopping is a callback, and it stops model training when a monitored metric has stopped improving. The model.fit() method trains the model for a fixed number of epochs. The validation split, epochs, batch size, and callbacks are parameters of the fit method.

Graphical user interface, text

Description automatically generated

The model.save() method is used to save model architecture, weights, and optimizer rates. The model.load() method loads model and all the parameters.

Graphical user interface, text, application

Description automatically generated

In production, model.predict() method is used to predict the results.

Diagram

Description automatically generated with low confidence

1. **SKLearn**

Scikit-learn is the most helpful library for machine learning in Python.  Classification, regression, clustering, and dimensionality reduction are just a few of the practical machine learning and statistical modelling algorithms in the SKLearn toolkit. [5]

The TfidfVectorizer() method of SKLearn converts a collection of raw documents to a matrix of TF-IDF features. These features are input to the model.

Graphical user interface, text, application, email

Description automatically generated

The train\_test\_split() method is used to split the data into random test and train matrics. The LogisticRegressionCV() method trains the model using Logistic Regression algorithm. To evaluate the model, SKLearn provides bunch of methods like score(), accuracy\_score() and classification\_report().

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

1. **Hugging Faces - Transformers**

Transformers provides APIs and tools to Pretrained models that can save the time and resources needed to train a model from scratch while lowering computing expenses and carbon footprint. These models facilitate typical tasks in several modalities, including NLP, CV, and Audio, quickly download and train state-of-the-art pre-trained models.

Graphical user interface

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

1. **NumPy**

NumPy stands for Numerical Python, the standard Python library for working with arrays (vectors and matrices), linear algebra, and other numerical computations. NumPy is written in C. Its arrays are quicker and use less memory than Python lists or arrays. [7]

1. **Pandas**

Pandas is an open-source Python package. It is fast, powerful, flexible, and easy for data science/analysis and machine learning tasks. Moreover, It is built on top of the Python programming language.

Text

Description automatically generated

1. **Matplotlib**

Matplotlib is a comprehensive Python library for creating static, animated, and interactive visualizations. Matplotlib can is used in Python scripts, Python shells, web application servers, and various graphical user interface toolkits.

Chart, pie chart

Description automatically generated

A picture containing text

Description automatically generated

1. **Seaborn**

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. [10] The heatmap () method plots the confusion matrix.

Chart

Description automatically generated

1. **Joblib**

Joblib is a set of tools to provide lightweight pipelining in Python. In particular: transparent disk-caching of functions and lazy re-evaluation is accessible, simple parallel computing. [11]

The dump() method saves all the model and weights into .pkl file and load() method is used to load the model back using .pkl file.

Graphical user interface, application, Word

Description automatically generated

# **References**

1. https://colab.research.google.com
2. https://research.google.com/colaboratory/faq.html
3. <https://keras.io>
4. <https://www.tensorflow.org/learn>
5. https://scikit-learn.org/stable/
6. https://huggingface.co/docs/transformers/main/en/index
7. https://numpy.org/doc/stable/user/quickstart.html
8. https://pandas.pydata.org/
9. https://pypi.org/project/matplotlib/
10. https://seaborn.pydata.org/
11. https://joblib.readthedocs.io/en/latest/