**Toxic Comment Classification with Hybrid Embedding Algorithms**

**Abstract**

Increased social media usage over the past ten years has contributed to an upsurge in hate speech. As more online communication tools have become available for people to express themselves, the spread of offensive language along with various undesirable behaviors which are not tolerated in a society has increased. A major problem in today's society is the rise of antisocial behavior in internet settings. The proliferation of online social platforms has led to an increased need for effective identification and management of toxic comments to ensure positive online interactions. Automatically detecting and identifying such behavior is essential. To solve this problem, fresh approaches must be used to filter out this kind of offensive content. Numerous experiments have been done to automate manual filtering because it is a challenging operation. The current thesis aims to provide a powerful method for classifying toxic comments with the help of a BiLSTM neural network and a hybrid embedding strategy that combines GloVe and FastText word embedding. The global context embedding from GloVe and the capacity to capture sub word information from FastText are complimentary qualities that are combined in our hybrid embedding methodology. The binary categorization of comments (toxic and nontoxic) using BILSTM with Hybrid embedding (combination of glove+ fast-text embedding) was found to have an acceptable accuracy of 95% with an F1-score of 0.95. The current research attempts to address the issue of accurately categorizing comments by relating models to bigger text corpora (high-quality word embedding) rather than just training data.

Keywords: social media platforms, toxic, deep learning, LSTM, Glove, Fast-text, comments, automatic, hate speech

1. **Introduction**

The COVID-19 epidemic posed difficulties for the entire planet (Wang, et al., 2020). Nearly the whole global population has seen changes in their daily lives over the last two years as a result of these shifts. Nowadays, people conduct their business, learn, shop, and socialize online. The requirement for physical separation has also impacted how people express their feelings. Social media platforms have emerged as one of the primary venues for individuals to express, among other things, their opinions, feelings, and other such things, in relation to the pandemic scenario. Additionally, according to recent studies (Daniel Allington, 2021), especially during the ongoing pandemic crisis, social media has been a significant source of incorrect information. In addition, social media platforms were taken into consideration and exploited by the pertinent Public Health Authorities to disseminate information to a larger audience (A. K. M. Chan, 2020).

The internet today provides users with a setting where they can generate and exchange content with essentially no limitations. Most users behave properly and productively when using the internet. Nevertheless, there is a subset of users that engage in behaviors that can be characterized as anti-social. There are several definitions of antisocial behavior now (Cheng, et al., 2015), but there are primarily two sorts of it:

Spreading false information: This type of action typically involves the development and dissemination of false content in many different forms, such as hoaxes, biased and false news and others.

User responses: This kind of behavior, which often prevails in conversations of the user, can take on a wide variety of shapes, including forum manipulating others, bullying on the internet, vitriol, slander, spamming, and other behaviors.

Both antisocial behavioral patterns pose a big problem because their effects might be seen in the real world as well. Real-time conversations between online users are common, and they frequently involve a sizable number of individuals. Modern technologies that provide partial anonymity through such widespread communication also create new hazards in the form of improper user responses. Abuse of words in online debates frequently results in hostile user responses. (Burney, 2013)

Finding the precise explanation of this phenomenon can be challenging because there are so many alternative definitions for this conduct, which is made even more complex in the internet world. Nevertheless, toxic comments during an online conversation can often be described as an aggressive response that compels the aggrieved participants to leave it (e.g., verbal bullying, personal attacks). (Julian Risch, 2020)

Traditional methods for classifying harmful comments frequently used manual involvement, rule-based algorithms, and straightforward lexical analysis. However, the complexity and variety of poisonous language inherent in online dialogues is beyond the scope of these techniques. This constraint has increased interest in creating sophisticated models that are capable of automatically detecting hazardous comments by combining the power of DL and NLP approaches. In this current research approach, we present a thorough investigation on improving toxic comment classification by combining hybrid GloVe and FastText embeddings with BiLSTM. We describe the architecture and design of our model, go into detail about the generation of hybrid embeddings, and talk about the training and evaluation approach. Our findings demonstrate that, when compared with current the most advanced models, our special method produces great accuracy in identifying hurtful remarks.

* 1. **Research Question**

How does dual embedding help in classifying toxic comments on social media with improved accuracy?

* 1. **Research Objectives**
* To develop a BiLSTM with dual embedded algorithm model for classification of toxic comments
* To study, compare and evaluate the results and accuracy generated by the BiLSTM model using the proposed dual embedding technique with the algorithm using the traditional word embedding models.

Further, the thesis is bifurcated into section 2 which describes systematic literature review in the area of toxic comment classification. The methodological framework, highlighting our hybrid embedding approach and the conceptual framework of the BiLSTM model is described in the 3rd section while 4th section provides the experimental setup, dataset, and evaluation metrics. The results and contrasts our approach with existing methods are mentioned in the 5th section; and 6th section concludes the paper by highlighting the significance of our findings and potential applications.

1. **Related Work**

For the purpose of better understanding, the literature review has been divided into the following parts:

* Review on word embedding
* Review on Social Media and Toxic/The Comments
* Review on Models for Automatic Comment Classification
* Gap analysis
  1. **Review on Word Embedding:**

The term ‘word vector’ converts natural language words in a format that computers can comprehend and recognize, making it possible for it to be translated by various NLP algorithms. In NLP methods, word vector representations are often utilized. (Barkan & Koenigstein, 2015). A quick binary approach to express word vectors is one-hot encoding. However, due to its encoding, the vector dimension will be too big to parse, losing the text's semantic information (Ying Xiong, 2020). It is suggested to assign text to low-dimensional space using dispersed word vector representation using predetermined vector dimensions (Hinton, 1986). Additionally, this vector space is useful as a tool for extraction of text features because it can instantly determine how similar two words are.

The Word2vec and Global Vectors (GloVe) word embedding algorithms are both significant. Several word vector models were proposed by (Quoc Le, 2014), including NNLM, RNNLM, and others.. The two word2vec model structures, CBOW and Skip-gram, must be proposed in particular. For phrase classification using CNN, (Kim, 2014) employed as dual-channel inputs static and non-static trained word vectors.

The Word2vec paradigm from natural language processing (NLP) was expanded by (Barkan & Koenigstein, 2015) in any field that has the ability to produce sequences, such as search, advertising, and recommendations. To train the emotional tendency of the text, (Yequan Wang, 2016) integrated aspect embedding and GloVe word embedding as the parameters of the LSTM model. The pre-trained GloVe and Word2vec embedding vectors were utilized by (Naderalvojoud & Sezer, 2020)in their DL model, which increased the precision of sentiment classification. To check the accuracy of sentiment prediction, (Seyed Mahdi Rezaeinia, 2015) concatenating a number of word vector formats, including Word2vec, GloVe, POS2Vec, Lexicon2Vec, and Word-position2Vec.

* 1. **Review on Social Media and Toxic/Hate Comments:**

Cyberbullying (Hamlett, et al., 2022), trolling (Hardaker, 2009), and the emergence of online toxic comments, which are "rapid releases of large quantities of negatively charged, frequently extremely emotional posts in the social media environment" (Hauser, et al., 2017), where users target other groups and organizations, are just a few examples of how online toxicity, Hate speech is characterized as statements that are likely to turn listeners away from a discussion. (Wulczyn, et al., 2017), can take many different forms. As per (Patton, et al., 2016), online toxicity should be handled seriously because it might lead to violent behavior both online and offline. That that participants can frequently post anonymous and won't be held responsible for their behavior in the same manner as in offline encounters is frequently perceived as enhancing toxic behavior in online venues (Hardaker, 2009). Online networks for underprivileged or disadvantaged populations are especially vulnerable to online toxicity because discriminating behaviors, such sexism and racism, can endure and worsen online. (Herring, et al., 2002).

Social networking platforms enable users to swiftly and easily create material on multiple topics. The anonymity and ease of content distribution on social media might lead to more damaging content being shared. According to (Anastasia Giachanou, 2020), various information categories, such as misinformation, disinformation, and mal-information, can be harmful either consciously or unconsciously. Misinformation is false or fictitious information that is prepared and disseminated without respect for the intended purpose. Disinformation, such as fake news, is purposefully produced to deceive the target audience (Nasir, et al., 2021). Mal-information, such as hate speech (Davidson, et al., 2017); (Anastasia Giachanou, 2020), is intended to instigate or hurt. In this poll, our main concern is identifying hate speech.

(LIM, 2020), who looked at the growing component of toxicity on social media, asserts that as technology tolerance and communication skills advance, user response to online toxic comments is on a rise. The finding of his study compels mass media specialists to further explore this phenomenon and confirms that societal concerns and media digitalization are publicly discussed. The categorization of text methodology was utilized by (Shervin Malmasi, 2017) to discover defamatory comments on social media. The motive of the study was to distinguish between offensive language, hate speech, and other sorts of social media content. The results of the investigation revealed that there isn't much of a difference between swearing and expressing hatred, and that doing so can be quite difficult. More research in this area is also suggested by the researcher in order to establish a clear grasp of the differences between profanity and hate speech.

Hatred can be shown in many different ways, whether it is intended or not. Although there aren't many phrases that can be used purposefully to disparage someone or something, other people might use those terms inadvertently. Twitter users have the potential to use exceedingly offensive language in their tweets and comments to actively rile up other users. (Vilenchik, 2019). However, a user could post a comment that people following them interpret in a different way, resulting in unexpected responses. While a post that incites hate tends to take the owner's tone in most circumstances, it occasionally may go in a direction that is not what the owner had in mind. According to (Vilenchik, 2019)research, the user's action is generally not associated with the feedback they get about that activity. Because of this, even though simple statistics can be used to monitor hatred in social networking sites on situations where certain terms are targeted, they sometimes can't be tracked effectively.

Hate speech is defined by Facebook's protected features as an assault on a person's dignity, particularly their race, ethnicity, or place of origin. Twitter's rules state that users cannot use tweets to harm or abuse people based on their ethnic background, sexual orientation, faith, any other characteristic. Along with content that is blocked based on age, caste, and disability, YouTube filters content that promotes violence or hatred against specific individuals or groups. Toxic comments has also been discussed in different contexts, however it is typically researched in relation to online radicalization or criminal action (Pantelidou, et al., 2021).

According to research, hearing or eating toxic statements can have a number of adverse consequences on people, particularly in terms of changes in behaviors, attitudes, and emotions. These alterations could activity with tangible result in suffering from various types of harm and a rise in the likelihood of participating in hate speech or hate-related behaviors, such as violent radicalization. From this vantage point, hate speech isn't just a collection of statements; it's also an effects. (Salminen & , 2020)

According to (Saha, 2019), study participants' expressions of stress were directly associated to their exposure to online hatred. According to earlier research by (Edgar Pacheco, 2018), this kind of content might lead to unfavorable emotions like "angry," "anxious," "sad," or "humiliated." Additionally, (Edgar Pacheco, 2018) discovered that exposure to hate speech online can result in behavioral changes, such as a desire to stay home, difficulty sleeping, an extreme fear of approaching strangers, and social isolation. In addition to finding evidence of a substantial association among being exposed to toxic comments online and perpetration, (Blaya & Audrin, 2019) also uncovered evidence of a causal connection between cyber hate victimization and perpetration. The social and psychological trust among those who have encountered hate speech was examined by (Nasi, 2015). They discovered that seeing bad writings and photos decreases social confidence in neighbors, people in general, individuals one has only encountered online, as well as coworkers or classmates.

* 1. **Review on models for Automatic Comment Classification:**

Text classification challenge that involves toxic comment classification. The majority of past studies treated Toxic comment classification as an issue of binary classification that distinguished one specific kind of hazardous comments from all the others (Nunes., 2018). Later studies define the classification of harmful comments as either multi-class classification tasks, in which a comment will be assigned to one of the numerous types of toxic comment, or multi-label tasks, in which a comment may be assigned to none, at least one (and maybe more) types of toxic comment.

Every day, millions of user-generated posts are created. As data volumes increase, accurately classifying and manipulating the data in accordance with a given formula is more and more difficult. (Van den Brakel, 2017)used associational terms as the classification features in a rule-based system to categorize hostility in tweets and Twitter comments. To strengthen their technique even further, their study included expressions that accuse or attribute anything after an interesting or disruptive to social order incident, to an individual or an entity. Because of this, their strategy significantly improved on the accepted techniques and methodologies for determining the degree of hatred. As a result, it is feasible to assess the intensity of hatred on social media with a high degree of accuracy.

The numerous social media networks' content formats could make it difficult to create a system that can reliably identify hateful emotions. While screening criteria for text-based microblogging sites may be reasonably simple, those based on other types of media, such videos, may be more difficult to identify. A framework was created by (Doring, 2020) for spotting atrocity and radicalism in YouTube videos. In order to categorize the contents, they used a variety of classification algorithms as well as philological, syntactic, and content-based elements from videos and comments that users had left on posts and other user-generated content. According to their research, videos are more intensely perceived than other types of content. As a result, their impact should be determined not only by their absolute values but also by their real-world applications.

SVM, NB, and DT are examples of statistical ML techniques that have historically been employed for text classification (Wiegand., 2017). DNN-based models, including CNN, RNN, BiLSTM, and hybrid neural networks that mix several DNN configurations, have been more popular in research since 2010 (Fabio Del Vigna, 2017) (Wiegand., 2017) (Ziqi Zhang, 2018)

The most popular DL algorithms for detecting cyberbullying are RNN and CNN methods. (Themeli, et al., 2021) used both traditional ML and DNN models to detect hate speech. According to their experimental findings, Bag of Words (BoW) outperforms GloVe and N-gram graphs when combined with a logistic regression model and a three-layered artificial neural network. A unique ensemble approach was proposed by Bu and Cho that employs two DL models for knowledge transfer: a CNN for character-level syntactic feature extraction from the text and a Long-term Recurrent Convolutional Network (LRCN) for semantic feature extraction (Bu & Cho, 2018). In their research on deep learning models, (Agrawal & Awekar, 2018) used CNN, LSTM, Bi-LSTM, and Bi-LSTM with attention together with random, GloVe, and Sentiment-Specific Word Embedding (SSWE) to transfer information across domains. When detecting aggression, (Aroyehun & Gelbukh, 2018) compared the results with those of traditional machine learning techniques, adding additional data and pseudo-labeling for the same, and employed seven varying combinations of CNN, LSTM, and Bi-LSTM models.

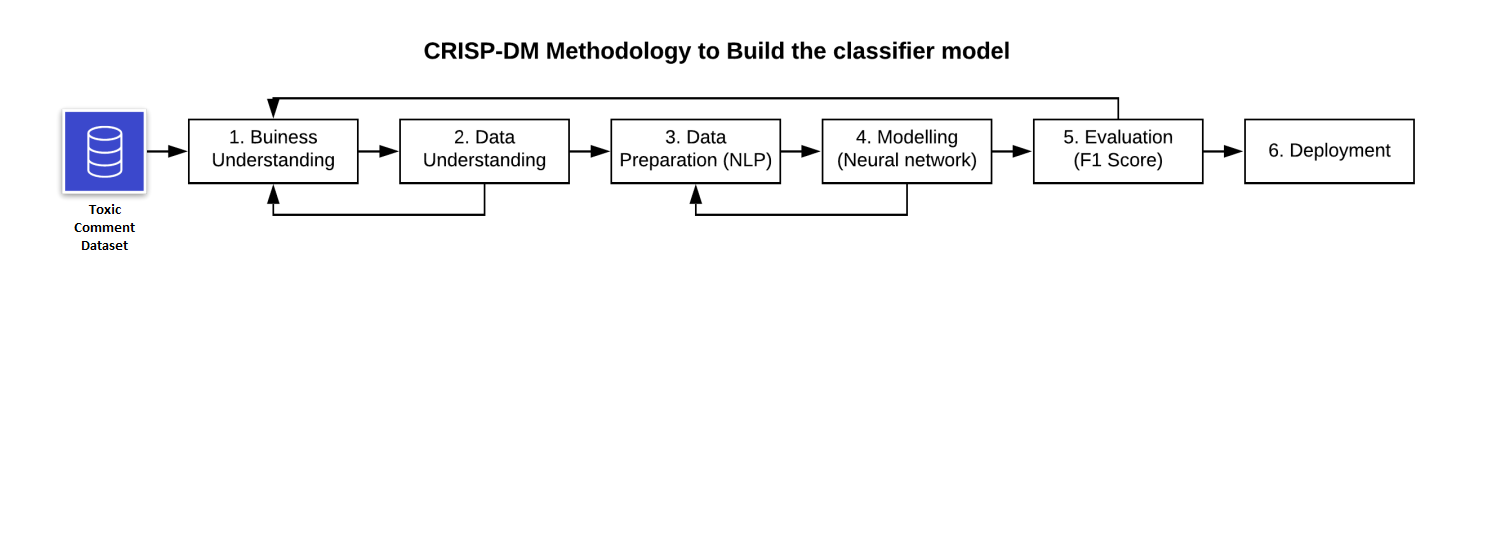
* 1. **Gap identified**

Although a lot of the research in the literature review focus on the material that users post on various social media like Twitter, Facebook are also quite important. The above papers reviewed, examines various methods to determine what kind of information might be regarded as toxic comments and how to contextualize these kinds of comments. Some text being analyzed uses overtly objectionable language, while in other situations it uses associative terminology that are considered to be hate speech. To be able to separate such enormous amounts of data analysis for analysis, many frameworks and techniques based on the study's objective and the volume of data, have been developed. However, the efficacy of the previous models are less due to the use of single algorithms and combinations of algorithms which provide lesser accuracy as compared to the proposed model which uses BiLSTM with dual embedding of Glove and Fast Text. This model generates accuracy more than the earlier models researched and also helps to analyze comments and its classification in a broader perspective.

1. **Research Methodology**

In this study, using NLP and Deep Learning algorithms, we evaluate texts, ascertain the deeper importance of comments, and identify hatred to provide a novel way for identifying harmful comments on social networking sites. The objective is to develop a model that is more accurate and computationally efficient than current methods. This chapter examines the preferred approach for implementing systems and gives a brief summary of it. The Cross-Industry Process for Data Mining methodological approach (CRISP-DM) was chosen as the overall methodology in this research.

* 1. **CRISP-DM Modules**



**Figure no. 1: CRISP DM Model**

Figure 1 depicts the cross-industry procedure for data mining methods to develop a classification model. It has six sequential phases which are as below:

**Phase 1: Business Understanding:**

Gaining a thorough grasp of the business challenge is the initial step in the CRISP-DM process. Understanding the issue arising with toxic comments and classification of the same is the main aim of this phase of the CRISP DM Model. The comments posted on platforms are categorized as toxic since they might reflect a remark disparaging a community and their semantic meaning might suggest something incorrect. With the help of the proposed model, we aim at classification of comments on social media as toxic or non-toxic and it is done to stop people from making potentially harmful comments, to promote more civil conversations, and to track the toxicity of everyone else's responses.

**Phase 2: Data Understanding:**

The data requirements for the project are fully understood by business and data analysts during the second stage of CRISP-DM. The review of the business questions identified in business understanding and the documentation of the data needs come first in this phase. The dataset we utilized for this study was made accessible by Google Jigsaw in December 2017 as part of the Kaggle "Toxic-Comment-Classification-Challenge." It is a publicly available repository called Kaggle that hosts 159,571 tagged reader comments that were posted from Wikipedia-talk pages. These statements were rated as toxic, threatening, vulgar, and disrespectful by social rating organizations. This dataset contains a column called "comment text" that functions as a variable that is independent and a column called "toxic" that functions as a variable that is dependent. A binary categorization that allows us to determine whether or not a statement is toxic serves as the dependent variable. A comment has a value of 1 if it is toxic, while a comment has a value of 0 if it is not toxic. Glove and Fast-Text embedding layers are produced for word embeddings and are downloaded from Stanford and Facebook's repositories.

**Phase 3: Data Preparation:**

The following stage involved pre-processing the data to provide useful information that would be fed to the Deep-Learning system. Making sure that the data being gathered is clear and devoid of noise is crucial because this could jeopardise the validity of the conclusion. Among other things, the comments data may comprise ordinary English terms, abbreviations, URLs, special symbols, slang terms, white-space, and emojis. To remove stop-words, URLs, meta tags, and symbols, a function using the "re" – regular expression library was developed. With 60 percent of the information being used for training, 20 percent for testing, and the remaining 20 percent for validation, the model was split into testing, validation, and training sets. We may evaluate the effectiveness of our model by using this approach.

3.1 **Word Embedding:**

Along with neural networks, word-embedding plays a key role in the Natural Language Process (NLP). Glove and FastText are the pre trained word embedding models are used to classify the comments.

3.1.1 **Glove Embedding:**

The GloVe model is the global word vector text representation model. Glove embedding generates matrix co-occurrence (word \* context). Thus, a massive matrix is generated, and factorization is performed to represent the matrix in a lower dimension matrix.

3.1.2 **FastText Embedding**

When compared to Glove, the FastText word-embedding layer is considerably different. In contrast to previous embedding techniques, it develops embedding for uncommon words on n-grams supplied to the layer.

3.1.3 **Hybrid Embedding**

Proposed dual embedding techniques combining the Glove & FastText model together will act as a enrichment for the word embedding approach. To develop deep learning models, the Glove and FastText embedding layers are employed together. These models have already been trained, therefore many of the terms in the comments will be present in the pre-trained layers. The key used to represent words has a length of 500 and a value of a word vector.

**Phase 4: Modelling:**

The suggested model is based on the Bidirectional Long Short Memory (BiLSTM) technique, which combines fast text embeddings with glove embeddings for applications dependent on natural language processing. It is chosen as a combination of different embeddings. These embeddings then allow the model to capture different features of words.

The following models were developed, trained and validated in order to develop the ideal proposed model:

1. BILSTM with Glove
2. BILSTM with Fast-text
3. BILSTM with Hybrid embedding (combination of glove+fast text embedding)

All of the models' results must be compared, and the simulation having the greatest degree of accuracy shall be chosen.

**Phase 5: Evaluation:**

The outcomes of the implementation demonstrated that the recommended model could quickly and accurately categorise words as harmful or non-toxic. Based to the comparative stage, the suggested model outperformed the previous models in terms of accuracy and several characteristics like F1-score, recall, and precision. We evaluated our model using a confusion matrix that showed the frequency of false positives and false negatives as well as the length and sequence of model training.

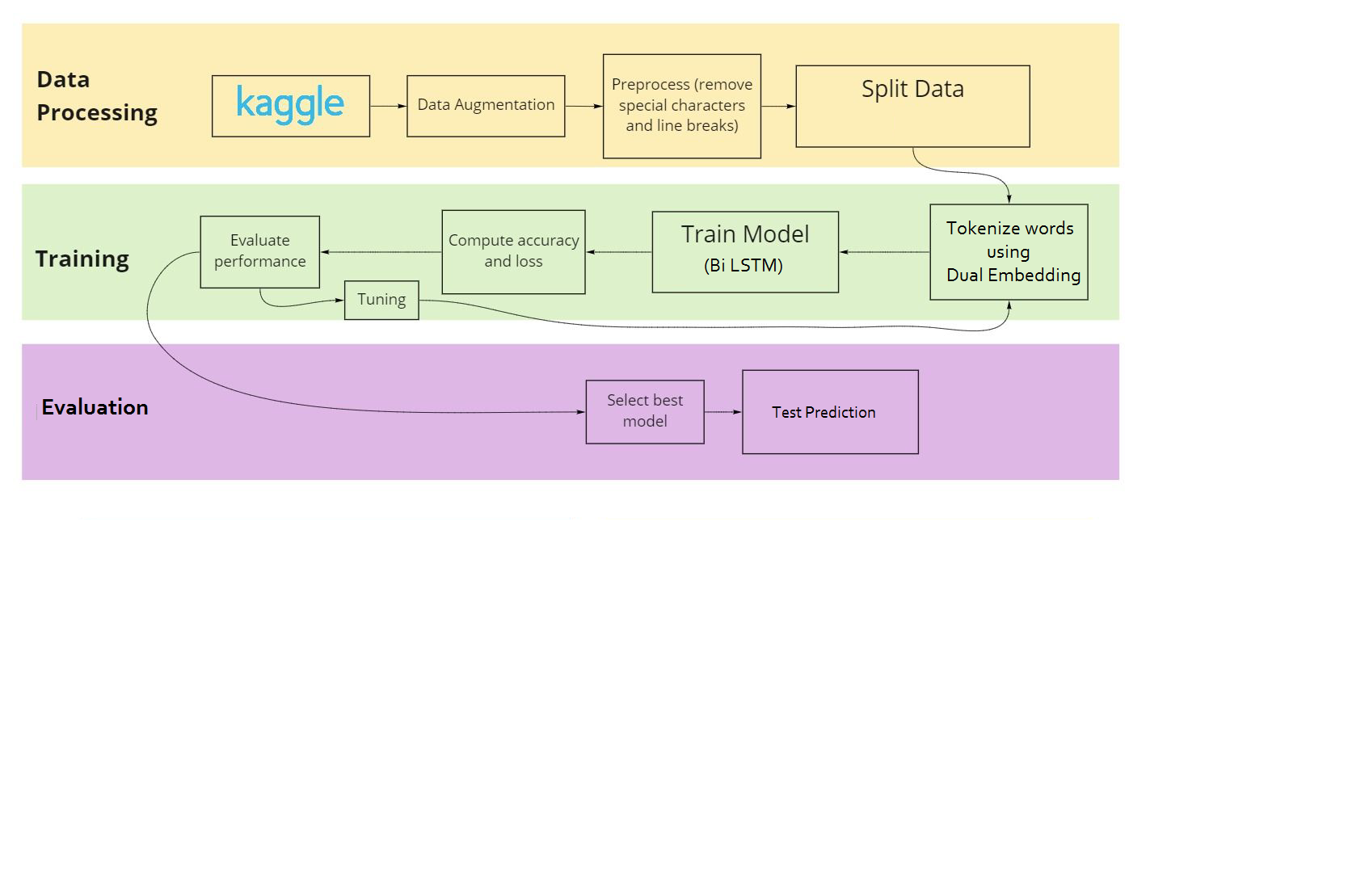
**Phase 6: Model Deployment:**

We have built the BiLstm DL model with hybrid dual embedding, which is efficient with the best performance parameters in terms of F1, accuracy. The model can be further deployed in real-time platforms for toxicity detection

1. **Design Specification**

This section covers the architecture used to develop a revolutionary proposed method for categorizing toxic remarks. Various BiLSTM model versions were built utilizing the following word embeddings: GloVE, Fast-text, and Dual Embedding. In the evaluation section further, the results of all models will be compared, accompanied by a visual representation of the final model's findings.

The details of the tools, characteristics, and features required to develop the suggested model are covered in the next section.

****

**Figure no. 2: Proposed Algorithm Framework**

1. **Implementation**

In this chapter, the actions necessary for data implementation are highlighted.

**The configuration of the proposed model is described as below:**

|  |  |
| --- | --- |
| **IDE** | **Google Colab Notebook** |
| **Program Language** | **Python** |
| **Computation** | **Tesla T4 GPU** |
| **Visualization Library** | **Matplotlib, WordCloud, Seaborn,** |
| **Modelling Library** | **Keras, tensorflow, sklearn** |

Table 1: Configuration Table

* 1. **EDA that is Exploratory Data Analysis:**

A Kaggle notebook with the GPU turned on is loaded with the Kaggle toxic comment categorization dataset. For EDA and preparation, Python libraries are employed.

The EDA is further carried out as follows:

* Count plot of target class
* Word Cloud

The counts of the data in each category bin are displayed using a count plot, which uses bars.

When representing text data, word clouds, a sort of data visualization in which the size of each word represents its frequency or importance, are utilized.

The word cloud for each class text is defined as shown as below:

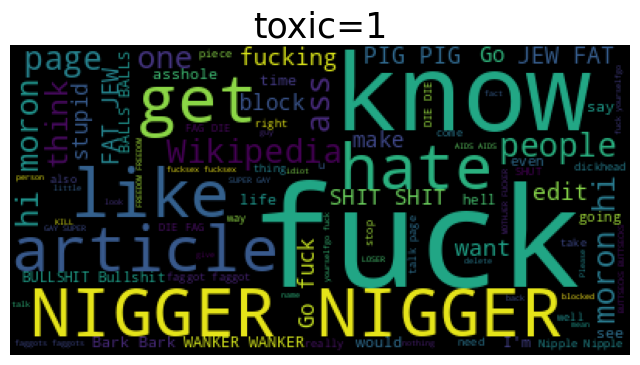


Figure no. 3: Toxic Comment Classification

Toxic =1 signifies a toxic comment

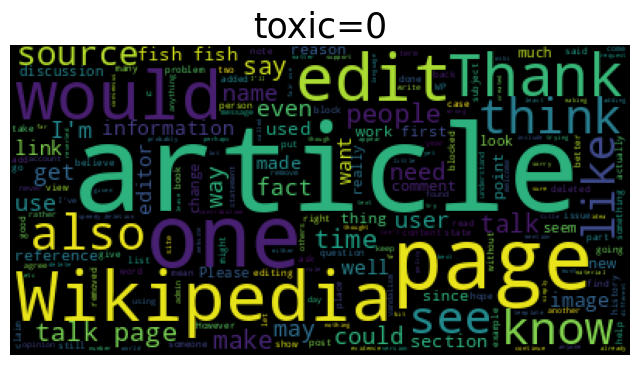
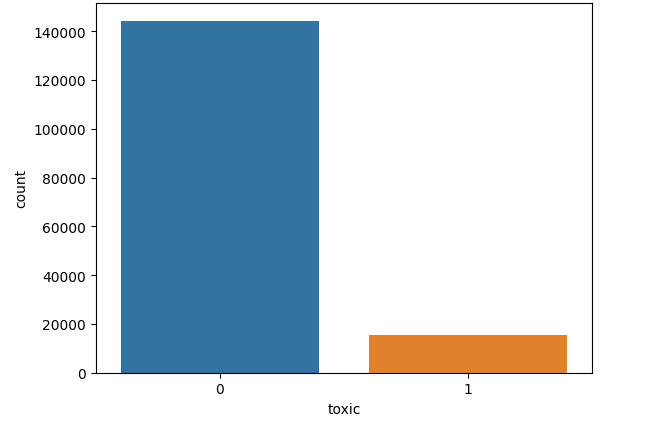


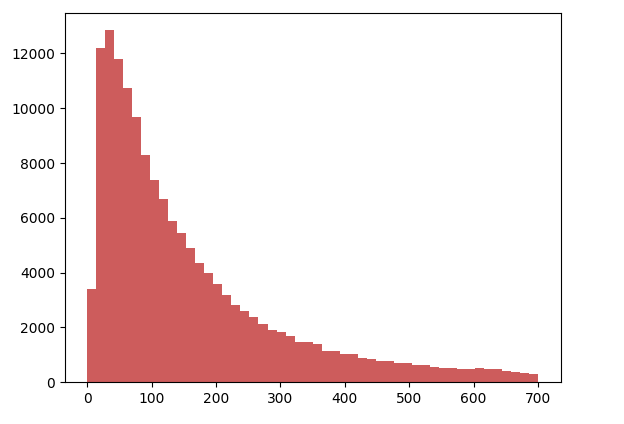
Figure no. 4: toxic Comment Classification

Toxic=0 signifies a non-toxic comment



**Figure no.5: Count for Toxic Comments**

Figure 5 shows a count plots graph for toxic comments, with a total of 15,294 toxic comments (0) and 15,294 non-toxic comments (1).



**Figure no. 6: Character Length Count Graph**

Figure 6 uses the amount of character in the training data to depict the total word count of each comment text. Most of the comments in the graph are less than 700 characters with many of the characters going up by more than 12,000 characters.

* 1. **Data Pre Processing and Cleaning**

We first convert the tweets CSV file into a Pandas data frame. Following are the steps we will perform for the preprocessing the data using the NLTK:

* Remove HTML entities
* Substitute @mentions, URLs, etc. with whitespace using regular expressions
* Substitute any non-alphabetic whitespace.
* All the words in lower case.
* Removing stop words.
* Punctuations
  1. **Model Training Evaluation:**

For the purpose of evaluation of the proposed model, the following evaluation metrics are considered:

* The metric of ‘Accuracy’ displays the percentage of cases in the testing data that were correctly identified (including true positives and true negatives).
* The percentage of true positives to all occurrences that the model predicted as positive is measured by ‘Precision’ metric. It displays the ratio of positive incidents that really occurred vs those that were anticipated. ‘Recall’, also referred to as sensitivity or true positive rate, is the ratio of true positives to all other instances of real positive data. It evaluates the model's ability to account for every occurrence of success.
* The ‘F1 score’ is the mean of recall and precision. Both false positives and false negatives are taken into account in order to provide a fair assessment. Since the metric takes both false positives and false negatives into account, it is very helpful when working with imbalanced datasets.

**Evaluation**

To ascertain the accuracy of the proposed model, we train, test, validate and compare the results on toxic comment classification of two other models with the proposed model. These models are

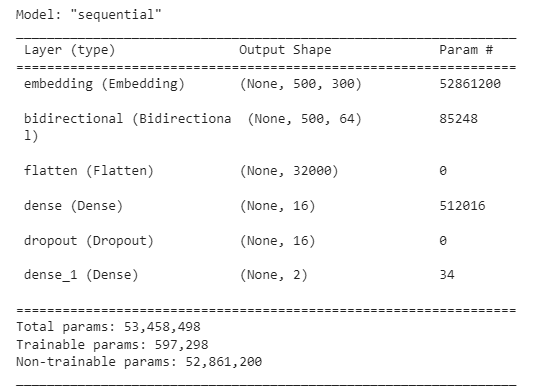
1. BiLSTM with Glove Embedding
2. BiLSTM with Text Embedding
3. BiLSTM with hybrid Glove and Fast Text Embedding

The evaluation and results for the Deep Learning Models is as below:

**Model 1: BiLSTM with Glove Embedding:**

A BiLSTM model with GloVe embedding is a popular approach for NLP tasks such as machine translation, named entity identification, sentiment analysis, and text categorization. In this system, words are represented as continuous vector spaces using pre-trained word embeddings, and the past and present contexts of the input sequence are mined for contextual information using BiLSTMs. To implement a BiLSTM model with GloVe embeddings, preprocess text data, load GloVe word embeddings, prepare the embedding matrix, build the model using a deep learning framework, compile, train, evaluate, and fine-tune the embeddings if dataset and task differ significantly from the pre-training corpus.

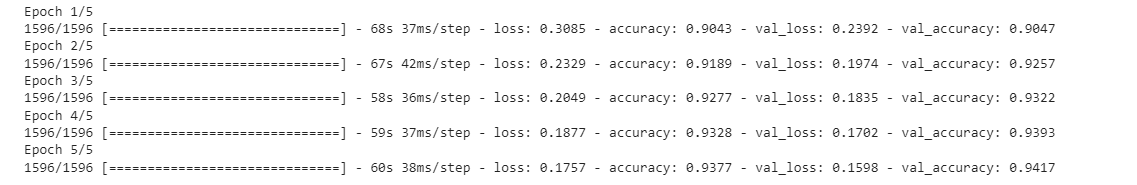
**Model Summary:**



Snippet 1: Model 1 with BiLSTM and glove embedding

In the above snippet, the embedding layer transforms tokenized words into 300-dimensional vectors, while the bidirectional LSTM layer captures data from past and future tokens. The output is transformed into a 2D tensor, and the dense layer, with 16 neurons, interprets the input and generates a 16-dimensional output. The dropout layer prevents overfitting by applying dropout regularization to the previous dense layer. The final dense layer maps the 16-dimensional output to a 2-dimensional output, representing the probabilities of the two classes in the text classification job.

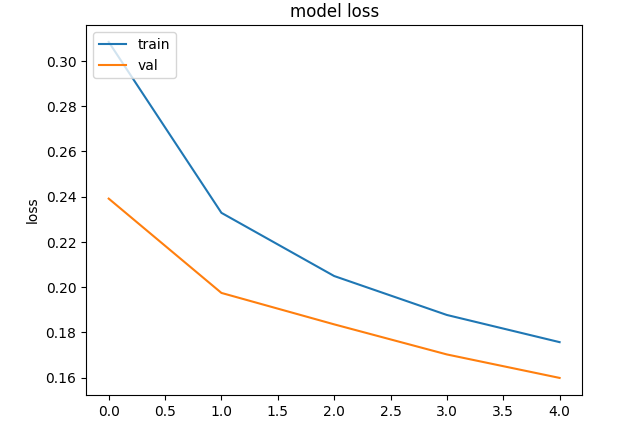
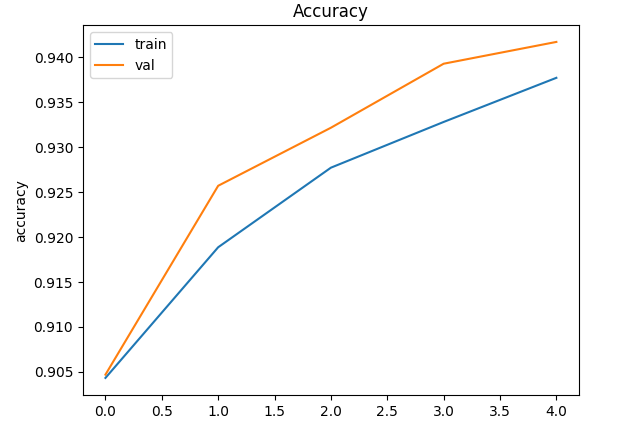
**Model Training:**



Snippet 2: Model 1 Testing Phase

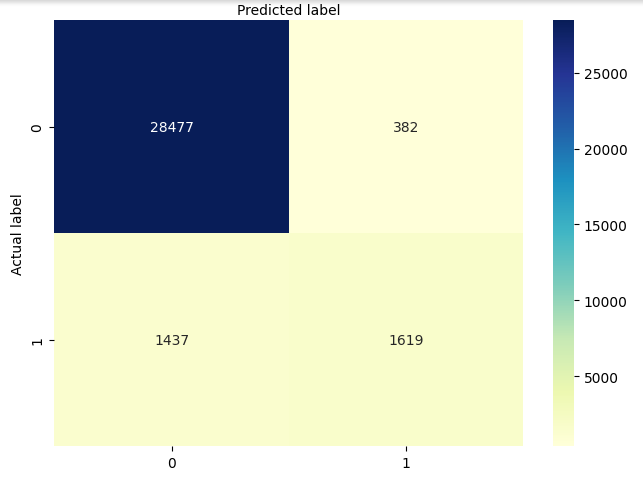
Across epochs, the model is performing better, as seen by lowering loss values and rising accuracy values on training as well as validation data. This indicates that it is improving its ability to generalise to new data. Additionally, included are the overall runtimes for the epoch and each batch. While the loss function measures the difference between expected and actual values, the accuracy measures the manner in which the model has performed based on the training data. The fifth epoch, which is the one being reported, displays an accuracy of 94.17%.

**Accuracy Loss Graph:**



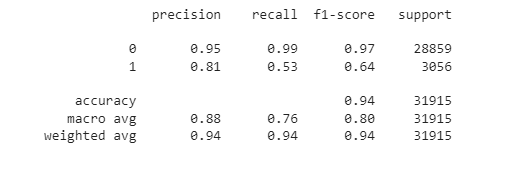
Snippet 3: Accuracy and Model 1 Loss Graph

**Confusion Matrix:**



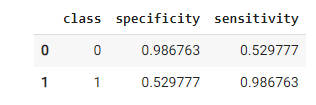
Snippet 4: Confusion Matrix

**Classification Report:**



Snippet 5: Classification Report for Model 1

**Specificity & Sensitivity:**

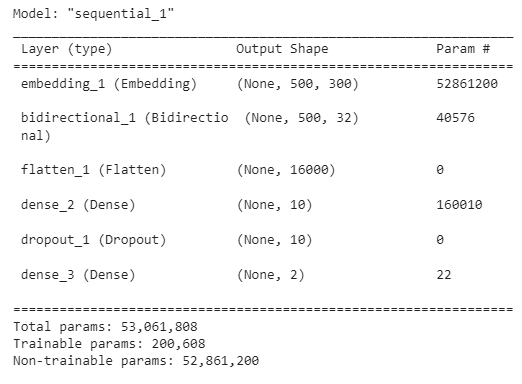


Snippet 6: Model 1 Specificity & Sensitivity

**Model 2: BiLSTM with Fast-Text Embedding:**

A BiLSTM model with text embeddings is a powerful approach for NLP tasks, capturing semantic and syntactic information in a continuous vector space. To implement a BiLSTM model with text embeddings, preprocess the text data, build a vocabulary, and convert it to integer sequences. Create an embedding layer, either random or pre-trained, and build the BiLSTM model in a deep learning framework. Add layers like the Embedding layer, BiLSTM layer, and Dense layer for final classification or prediction. Compile the model and provide the evaluation metrics, optimizer, and loss function. Preprocessed training data are fed into the model, and evaluation criteria are used to assess the model. Fine-tune the text embeddings for specific tasks during training. The success of a BiLSTM model with text embeddings depends on the dataset quality, task complexity, and hyper parameters chosen during training.

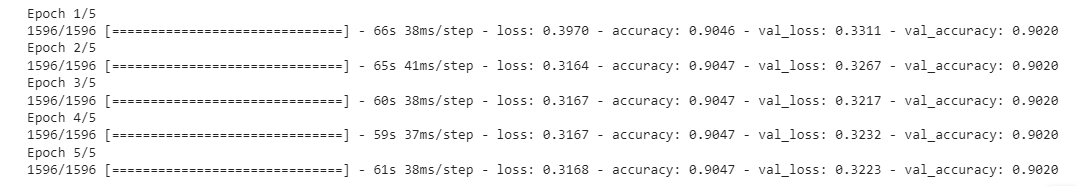
**Model Summary:**



Snippet 7: Model with biLSTM and Fast Text embedding

In the above snippet, the embedding layer transforms tokenized words into 300-dimensional vectors, while the bidirectional LSTM layer captures data from past and future tokens. The output is transformed into a 2D tensor, and the dense layer, with 10 neurons, interprets the input and generates a 10-dimensional output. The dropout layer prevents overfitting by applying dropout regularization to the previous dense layer. The final dense layer maps the 10-dimensional output to a 2-dimensional output, representing the probabilities of the two classes in the text classification job.

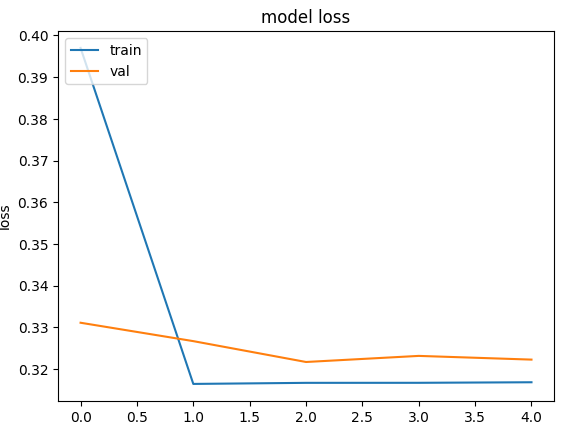
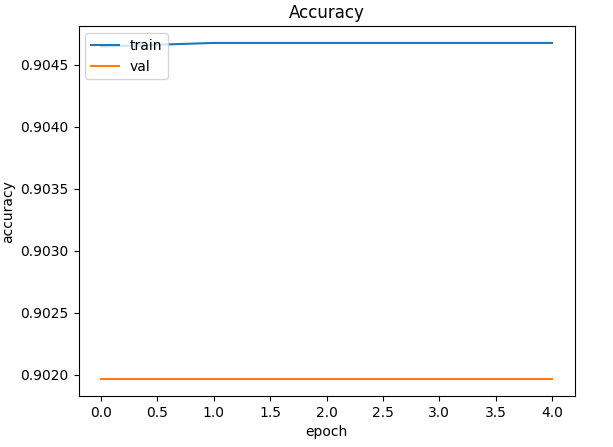
**Model Training:**



Snippet 8: Model 2 Testing Phase

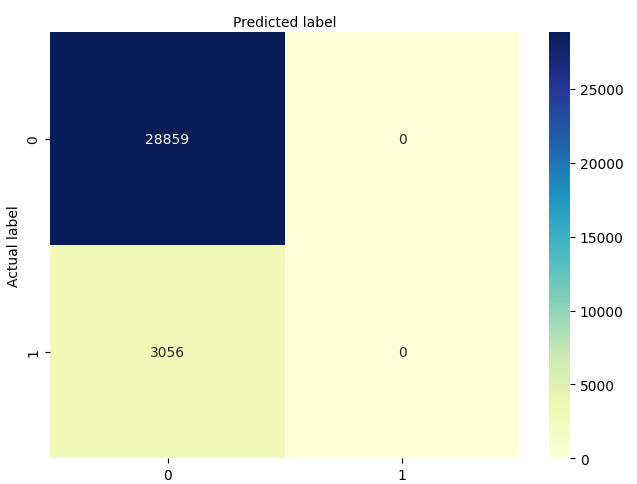
The framework is performing better with time as seen by decreasing loss values and increasing values for accuracy on training and validation data. It shows that the algorithm is getting better at generalising to new data. The total runtimes for the epoch and each batch are also supplied. While the loss function measures the difference between expected and actual values, the accuracy measures the way the model has performed based on the training data. The fifth epoch, which is the one being reported, displays an accuracy of 90.20%

**Accuracy & Graph Loss:**



Snippet 9: Accuracy and Model 2 Loss Graph

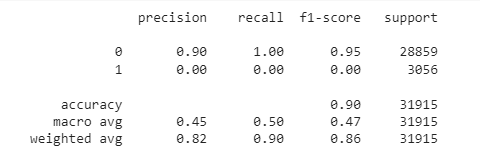
**Confusion Matrix:**



Snippet 10: Confusion Matrix

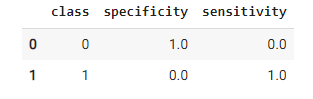
The above confusion matrix shows that the model is an overfitting for class (1) that is toxic comment classification.

**Classification Report:**



Snippet 11: Classification Report for Model 2

**Specificity & Sensitivity:**

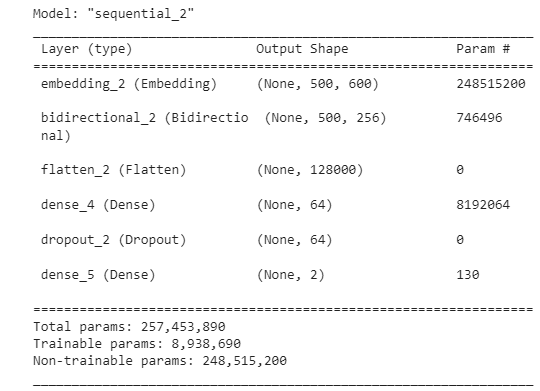


Snippet 12: Model 2 Specificity & sensitivity

Model 3: BiLSTM with Hybrid Embedding if Glove and Fast-Text:

Combining GloVe and FastText embeddings in a BiLSTM model is a powerful approach for NLP tasks, as it leverages the complementary strengths of both techniques. GloVe embeddings capture global word co-occurrence statistics, while FastText embeddings capture subword information and are more robust for handling out-of-vocabulary words. To implement a BiLSTM model with GloVe and FastText embeddings, preprocess text data, load GloVe word embeddings, prepare GloVe embedding matrices, and FastText embedding matrices. Build the BiLSTM model by adding layers 1 and 2, concatenating outputs, and a Bidirectional LSTM layer. Compile the model, train using preprocessed data and labels, and evaluate its performance using chosen metrics.

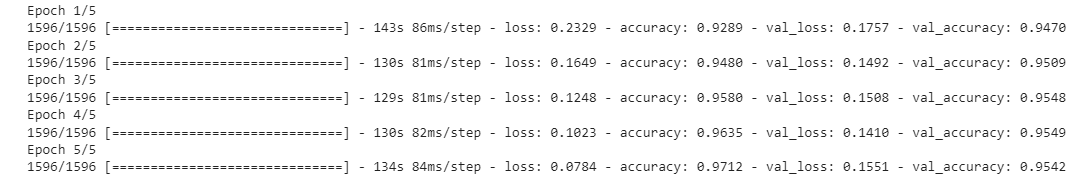
**Model Summary:**



Snippet 13: BiLSTM with dual embedding (Glove + Fast Text) Model

In the above snippet, the embedding layer transforms tokenized words into 600-dimensional vectors, while the bidirectional LSTM layer captures data from past and future tokens. The output is transformed into a 2D tensor, and the dense layer, with 64 neurons, interprets the input and generates a 64-dimensional output. The dropout layer prevents overfitting by applying dropout regularization to the previous dense layer. The final dense layer maps the 64-dimensional output to a 2-dimensional output, representing the probabilities of the two classes in the text classification job.

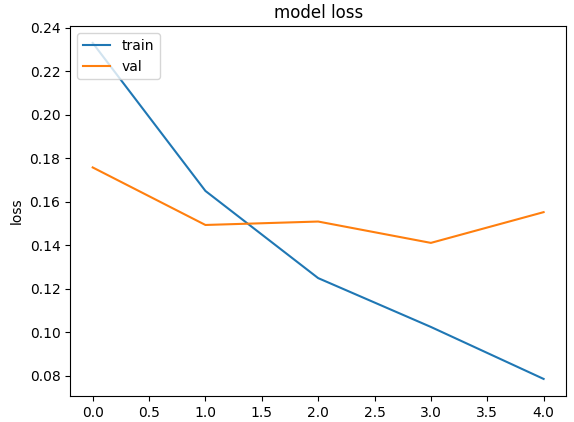
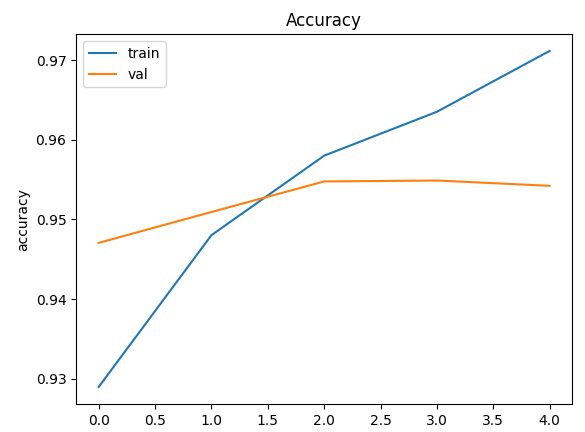
**Model Training:**



Snippet 14: Model 3 or Proposed System Testing Phase

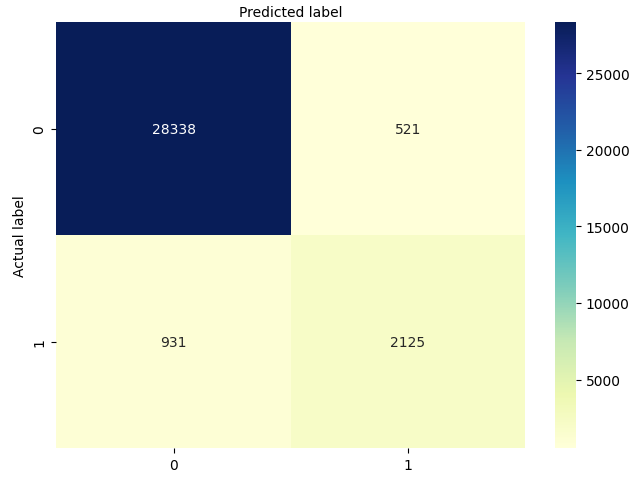
Across epochs, the model is performing better, as seen by lowering loss values and rising accuracy values on training as well as validation data. It indicates that the model is improving its ability to generalise to new data. Additionally included are the overall runtimes for the epoch and each batch. The accuracy reflects how well the model performed on the training data, whereas the loss function quantifies the discrepancy between anticipated and actual values. The fifth epoch, which is the one being reported, displays an accuracy of 95.42 %.

**Accuracy & Loss Graph:**



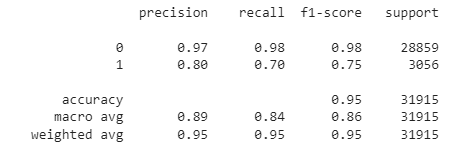
Snippet 15: Accuracy and Model 3 Loss Graph

Confusion Matrix:



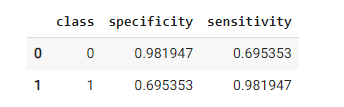
Snippet 16: Confusion Matrix for Proposed Model

**Classification Report:**



Snippet 17: Classification Report for Proposed System

**Specificity & Sensitivity**:



Snippet 18: Specificity & sensitivity of Model 3

1. **Conclusion & future Work**

The "Toxic Comment Classification with Glove and FastText Embedding" project aimed to develop a robust model for toxic comment classification using a BiLSTM with Glove and FastText Embedding. The study achieved excellent performance on the test dataset, combining BiLSTM models with GloVe and FastText embeddings. The approach demonstrated strengths in global word semantics and handling out-of-vocabulary words and character-level information. It also demonstrated robustness to imbalanced data, interpretability, and reasonable training time and model size. The study advances the classification of toxic comments and sets the groundwork for more advanced content moderation methods to be implemented in order to promote a safer and more diverse online community.

As per the three models tested for toxic comments above, the BiLSTM model with hybrid embedding using Glove and Fast Text is the best model as it provides an accuracy of 84 percent more than the other two models. The accuracy offered by each of the three models, as well as their evaluation criteria for poisonous remarks, are stated in the table below. The table indicates data for toxic comments classified only:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Precision** | | **Recall** | | **F1 Score** | |
|  | **toxic** | **non toxic** | **toxic** | **non toxic** | **toxic** | **non toxic** |
| **BiLSTM with Glove Embedding** | 0.81 | 0.95 | 0.53 | 0.99 | 0.64 | 0.97 |
| **Accuracy** | 94% | | | | | |
| **BiLSTM with Fast Text Embedding** | 0.00 | 0.90 | 0.00 | 1.00 | 0.00 | 0.95 |
| **Accuracy** | 90% | | | | | |
| **BiLSTM with Hybrid Embedding (Glove + Fast Text)** | 0.80 | 0.0.97 | 0.70 | 0.98 | 0.75 | 0.98 |
| **Accuracy** | 95% | | | | | |

`

Table 2: Accuracy from Classification Report

As per the results above, it is evident that the last model of BiLSTM with hybrid embedding of Glove and Fast Text is the most desired model with an accuracy of 95 %.

Future work may involve further fine-tuning the model's hyper parameters, exploring other word embedding techniques, and investigating additional ensemble strategies to enhance performance even further. Additionally, applying the model to real-world platforms and monitoring its performance in production settings would be valuable for assessing its practical effectiveness.

Overall, our research advances the classification of toxic comments and sets the groundwork for the implementation of more advanced and efficient moderation of content systems to promote a safer and more diverse online community.

**Reference**

A. K. M. Chan, C. P. N. J. W. R. A. L. G. M. J., 2020. "Social media for rapid knowledge dissemination: early experience from the COVID‐19 pandemic". *Anaesthesia, PMC Covid Collection*, December, pp. 1579-1582.

Agrawal, S. & Awekar, A., 2018. *Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms..* Springer: Cham, Switzerland, In Advances in Information Retrieval.

Anastasia Giachanou, P. R., 2020. *The Battle Against Online Harmful Information: The Cases of Fake News and Hate Speech.* Ireland, CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event.

Aroyehun, S. & Gelbukh, A., 2018. *Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling..* Santa Fe, NM, USA, In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying.

Aswani R., K. A. K. I. P. V., 2019. Experience: managing misinformation in social mediainsights for policymakers from twitter analytics.. *Journal of Data and Information Quality,* 12(1), pp. 1-18.

Barkan, O. & Koenigstein, N., 2015. *Item2Vec: Neural Item Embedding for Collaborative Filtering.* s.l., In: 2016 IEEE26th International Workshop on Machine Learning for Signal Processing (MLSP), Italy. IEEE, pp 1–6.

Blaya, C. & Audrin, C., 2019. *Toward an understanding of the characteristics of secondary school cyberhate perpetrators.,* s.l.: Frontiers in Education.

Burney, E., 2013. *Making People Behave: Anti-Social Behaviour, Politics and Policy, Second Edition.* 2nd ed. Cullompton, UK: Willan Publishing.

Bu, S. & Cho, S., 2018. *A hybrid deep learning system of CNN and LRCN to detect cyberbullying from SNS comments.* Oviedo, Spain, In Proceedings of the International Conference on Hybrid Artificial Intelligence Systems.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C. & Leskovec, J., 2015. *Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions.* Oxford,UK, Proceedings of the International AAAI Conference on Web and Social Media.

Daniel Allington, B. D. ,. S. W. ,. N. D. ,. J. R., 2021. "Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency". *Psychol Med, PubMed*, July, pp. 1763-1769.

Davidson, T., Warmsley, D., Macy, M. & Weber, I., 2017. *Automated Hate Speech Detection and the Problems of Offensive Language.* s.l., Proceedings of the International AAAI Conference on Web and Social Media 11(1).

Doring, N. &. M. M. R., 2020. *Gendered hate speech in YouTube and YouNow comments: Results of two content analyses,* s.l.: SCM Studies in Communication and Media.

Edgar Pacheco, N. M., 2018. *Online hate speech: A survey on personal experiences and exposure among adult New Zealanders.,* s.l.: SSRN.

Fabio Del Vigna, A. C. F. D. M. P. a., 2017. *Hate me, hate me not: Hate speech detection on facebook..* s.l., In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). 86ś95..

Hamlett, M., Powell, G., Silva, Y. N. & Hall, D., 2022. *A Labeled Dataset for Investigating Cyberbullying Content.* Atlanta, GA, USA, The 16th International AAAI Conference on Web and Social Media (ICWSM).

Hardaker, C., 2009. *Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions..* Lancaster University, Linguistic Impoliteness and Rudeness II conference, pp. 215-242.

Hauser, F., Julia Hautz, K. H. & Fülle, J., 2017. Firestorms: Modeling conflict diffusion and management strategies in online communities.. *The Journal of Strategic Information Systems.,* Volume 26, pp. 285-321.

Herring, S., Job-Sluder, K., Scheckler, R. & Barab, S. A., 2002. *Searching for Safety Online: Managing "Trolling" in a Feminist Forum,* s.l.: The Information Society.

Hinton, G. E., 1986. *Learning Distributed Representations of Concepts..* s.l., In Proceedings of the Eighth Annual Conference of the Cognitive Science Society. Amherst, MA, 1–12..

Julian Risch, R. K., 2020. *Toxic Comment Detection in Online Discussions.* Singapore: Springer.

Kim, Y., 2014. *Convolutional Neural Networks for Sentence Classification.* s.l., Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Association for Computational Linguistics.

LIM, S. S., 2020. *Manufacturing hate 4.0: Can media studies rise to the challenge?,* s.l.: Singapore Management University.

Naderalvojoud, B. & Sezer, E. A., 2020. Sentiment aware word embeddings using refinement and senti-contextualized learning approach. *Neurocomputing,* Volume 405, pp. 149-160.

Nasi, M. R. P. H. J. H. E. &. O., 2015. Exposure to online hate material and social trust among Finnish youth.. *Information Technology & People*, pp. 607-622.

Nasir, J. A., Khan, O. S. & Varlamis, I., 2021. Fake News Detection: A hybrid CNN-RNN based Deep Learning (DL) Approach.. *International Journal of Information Management Data Insights,* 1(1).

Nunes., P. F. a. S., 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR).*

Pantelidou, K. et al., 2021. *Change point detection in terrorism-related online content using deep learning derived indicators,* s.l.: Information.

Patton, D. U., Eschmann, R. D., Elsaesser, C. & Bocanegra, E., 2016. Sticks, stones and Facebook accounts: What violence outreach workers know about social media and urban-based gang violence in Chicago. In: s.l.:Computers in Human Behavior, pp. 591-600.

Quoc Le, T. M., 2014. *Distributed Representations of Sentences and Documents.* s.l., Proceedings of the 31st International Conference on Machine Learning.

Saha, K. C. E. &. D. C. M., 2019. *Prevalence and psychological effects of hateful speech in online college communities..* s.l., Proceedings of the 10th ACM Conference on Web Science.

Salminen, J. & M. H. S. A. C. S.-g. J. H. A. &. B. J. J., 2020. Developing an online hate classifier for multiple social media platforms. *Springer Link.*

Seyed Mahdi Rezaeinia, R. R. ,. A. G. ,. S. M. R. ,. R. R. ,. A. G. ,., 2015. Sentiment Analysis based on Improved Pre-trained Word Embeddings. *Expert Systems With Applications,* pp. 139-147.

Shervin Malmasi, M. Z., 2017. *Detecting Hate Speech on Social Media.* Varna, Bulgaria, Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017.

Themeli, C., Giannakopoulos, G. & Pittaras, N., 2021. *A study of text representations in Hate Speech Detection,* s.l.: s.n.

Van den Brakel, J. S. E. D. P. &. B. B., 2017. Social media as a data source for official statistics;the Dutch Consumer Confidence Index. *Survey Methodology*, pp. 183-210.

Vilenchik, D., 2019. *Simple statistics are sometimes too simple: A case study in social media data..* s.l., IEEE Transactions on Knowledge and Data Engineering.

Wang, T., Lu, K., Chow, K. & Zhu, Q., 2020. *COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model.,* s.l.: IEEE.

Wiegand., A. S. a. M., 2017. *A survey on hate speech detection using natural language processing..* s.l., In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. .

Wulczyn, E., Thain, N. & Dixon, L., 2017. *Ex Machina: Personal Attacks Seen at Scale.* Perth, Austrailia, Proceedings of the 26th International Conference on World Wide Web, WWW 2017.

Yequan Wang, M. H. X. Z. L. Z., 2016. *Attention-based LSTM for aspect-level sentiment classification..* Austin, Texas, In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas..

Ying Xiong, S. C. ,. H. Q. ,. H. C. ,. Y. S. ,. X. W. ,. Q. C. ,. J. Y. ,. B. T., 2020. *Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity,* s.l.: BMC Medical Informatics and Decision Making 20(S1).

Ziqi Zhang, D. R. a. J. T., 2018. *Detecting hate speech on twitter using a convolution-gru based deep neural network.* s.l., In European semantic web conference. Springer, 745ś760.