Final_report_team66[2] (1)

by Janav Shetty

Submission date: 06-May-2024 02:50PM (UTC+0530)
Submission ID: 2372147217
File name: Final_report_team66_2_1.docx (2.74M)
Word count: 3997
Character count: 22039

ABSTRACT

Toxic behavior in online conversation is so common in today's world due to an increase in social media users. User also tend to use euphemistic language (using special symbols like * ! @) to pass the existing toxicity detection model. So there is a need to come up with a different model to solve this problem. Our project “Detection of euphemism using BERT Model”, aims to develop a computation solution for automated detection of euphemistic language.

We focus on using advanced Natural Language Processing (NLP) techniques , particularly BERT(Bidirectional Encoder Representations from Transformers) .BERT enables us to achieve our goal in this project better than any existing models.

The project scope would be training the model with a wide variety of datasets and developing a real-time detection system. We aim to implement the user-friendly interface and evaluate the performance using different metrics.

We also strive to solve the possible challenges like language biases and continuous adaptation that way occur during the development of the project .

Our project aims to create an impact to create a respectful online conversation and provide a base for advanced online toxicity detection.

CHAPTER 1

INTRODUCTION

In the age of online communication, online toxicity is a very concerning issue due to its negative impact on a user's mental health. The overall decorum of the discussion is influenced by the use of these words. The huge volume of data on these platforms makes it difficult for the content moderators to control the toxic content on these platforms. The changing of nature of the languages is a major problem. Toxic content can have many forms, which makes it a complex task to detect it.

The process is further complicated due to the people trying to bypass the existing detection models to mask toxic content. The subtle words or visually similar looking symbols used to represent negative words is called euphemism. Euphemism can also be indirect language which does not look toxic at the first look.

Another factor to be considered when determining the toxicity of a sentence is the context of the message. Context of a sentence deals with the understanding of the entire conversation, user history, cultural nuances and the real-time dynamics. The understanding of the context of the entire conversation makes the identification of foul language easier.

In our project, we aim to use the current Natural Language Processing techniques, like BERT and LSTM to address the aforementioned issues. We attempt to automate the process of euphemism detection. We strive to improve the ability of identification and classification of euphemisms using the power of the BERT model, to foster a safer environment in social media platforms.

CHAPTER 2

PROBLEM DEFINITION

This research project aims to identify and analyze the euphemistic toxic language in online communication. It uses techniques of natural language processing focusing primarily on BERT model. Our project aims to find the euphemisms that hide toxic or harmful content using which we provide automated ways to flag such language for further review.

Euphemistic language poses a challenge in its detection .This is because it often includes subtle expressions, alternative words, or symbols that are used to convey a toxic language indirectly. The traditional methods have been struggling to accurately identify and flag problematic language. Thus our research aims to leverage the advanced technologies to improve the detection of euphemisms and enhance the moderation of online communication.

We have chosen the BERT model because of its strong performance in analyzing natural language text. The model receives additional contextual information about conversations in order to improve its ability to detect euphemistic language. By focusing on the broader context, we are aiming that the model can understand the intent better behind a user's words.

Our project will also take audio data in addition to the textual data for a more comprehensive analysis and improved detection of toxic language. By including both data types, the model can provide a more nuanced and thorough analysis.

Technically we plan to use Python based frameworks like Pytorch, scikit learn and Hugging Face transformers. They will be training the model, perform its evaluation and deployment. We are planning to use Streamlit to create an interactive user interface which will allow our users to easily input text and audio data and receive toxicity classification results.

The project focuses on accuracy and efficiency in detecting euphemistic language. It aims to create a system that is effective in flagging harmful content and is user friendly. By leveraging advanced NLP models and innovative data handling, this project aims to contribute to the responsible moderation of online communication and create a safer digital space.

CHAPTER 3

LITERATURE SURVEY

3.1 PAPER 1 [1]

The objective of this paper is to understand the impact of adversarial attack on existing neural toxicity detectores solve the issues in such cases. By understanding the problems in existing toxicity classifiers, the research aims to enhance the robustness of the model for more accurate toxicity detection .

The methodology used creates a lexicon of the toxic token to generate realistic adversarial attacks .These attacks involve character-level perturbation and injections and non-toxic distractors token to classify the toxicity . The models used are ELMo and BERT toxicity detection . It also uses contextual denoising autoencoder (CADE) as an approach to learn robustness in character level and contextual information to improve the resilience of toxicity classifiers against adversarial attacks.

Two approaches are used to solve the problem , first involves training on synthetically noised data to enhance the model ability , second is using CADE to learn robust representation that can classifier denoise the token and improve the classification.

Results show that adversarial attacks can significantly reduce the detection in the present model , with some models experiencing a drop of 50% in recall performance . The metrics used to evaluate are AUC , F1 score and recall .The study focuses on developing defenses that can withstand sophisticated adversarial attacks and improve the overall reliability of a toxicity detection system.

3.2 PAPER 2 [2]

The paper introduces a new method to automatically identify multiword euphemisms that are used by fringe groups and organizations in online forums. Euphemisms are ordinary words with hidden meanings that make content moderation challenging on social media.

The researchers of this paper have divided their work into three main stages. First the quality phrases are extracted using Autophrase which is a data driven phrase mining tool. This helps in creating a list of potential candidates which are high-quality phrases that can be used for further analysis. Then the euphemistic phrase candidates are preselected by calculating the cosine similarities of the word embeddings using word2vec.This helps in filtering out the noisy candidates and identify phrases that are semantically related to the target keywords, such as drugs with context to this paper. The last step is to rank the preselected euphemistic phrase candidates using SpanBERT.

SpanBERT was designed to predict the token spans in text which helps the researchers to analyze the long textual data more effectively. Thus by combining these three stages, we achieve the automated detection of euphemistic phrases.

In addition to the above features, this study also allows the detection of both known and potentially new euphemisms used online. This enhances the efficiency and accuracy of identifying euphemisms online. By using the paper's methods such as phrase mining and SpanBERT , we plan to improve the detection of toxic language phrases in online interactions. Overall, by referencing this paper we hope to improve the content moderation efforts on digital platforms.

3.3 PAPER 3 [3]

This paper talks about "Does Context Really Matter?" as the heading of the paper itself states it. So mainly in this paper the authors have focussed on two questions (a) Does context affect human judgement, and (b) Does conditioning on context improves performance of toxicity detection systems?.

To investigate or see whether considering context improves the performance of toxicity detection systems. Two data sets were created in this paper : (i) CAT-SMALL and (ii) CAT-LARGE.

CAT-SMALL consisted of around 250 comments from Wikipedia Talk Pages, it was divided into two groups. One group had context, the other group did not have context. Each comment was analysed for toxicity by three annotators, and scores were rounded to toxic or non toxic.

CAT-LARGE consisted of around 20,000 comments, half of them were annotated with context and the rest half of them without context. This larger dataset allowed the authors for a more detailed analysis of the impact of context on the toxicity detection.

If we look at the models and approaches used in this paper, the researchers have explored various techniques. They used RNN language models and combined the consecutive comments before using RNN. They also included contextual features for sentiment classification. These methods were used to see how context influenced the toxic comments and how well the toxicity classifiers work.

The results of the analysis revealed that context had a statistically significant effect on the perceived toxicity of the comments. Approximately 5.2% of comments showed changes in toxicity labels when the context was considered. However the research found no evidence that context actually improved the performance of toxicity classifiers.

The study concluded that context can make comments seem more or less toxic, but it did not improve any system performance. So the researchers said we need bigger datasets to understand the context's impact on toxicity analysis better.

In summary, the research paper provides important details about how context affects toxicity detection. It also shows the complexity or challenges in accurately judging the toxicity of online comments and using context to improve automated toxicity detection systems.

3.4 PAPER 4 [4]

This paper talks about using Masked Language Model (MLM) for Euphemism Detection. It introduces the concept of self-supervision into the task of detection and classification of Euphemism. Self supervision of this task helps reduce the manual efforts required to define the euphemisms and also deals with problem of subjectivity that is introduced due to human judgement.

The authors have divided their research to focus mainly on two tasks- euphemism detection and euphemism identification. Euphemism detection is the algorithm that takes a set of forbidden words as input and tries to generate the possible variations of the words, basically euphemisms.These set of words generated may signify the same meaning as the words in the forbidden set, and is called Candidate set. The candidate set of terms is used by the model to find new euphemisms and detect forbidden words. The second task is to take a single word that is considered to be euphemistic and find out its meaning.

Both the tasks are applied in succession, like a pipeline, by the moderators to discover new euphemisms and their meanings. The authors use Masked Language Models and other self-supervised techniques for accomplishing these tasks. In this paper, the approach used is to consider an input sentence and mask the term that is likely to be an euphemism. Based on the context of its occurrence and the rest of the sentence, the meaning of the masked term is predicted.

The classification of a euphemistic term is highly dependent on the context of the sentence as the word considered may or may not be used in a veiled sense. The paper uses Masked Language Model twice- once to filter out the masked sentences and then to detect hidden meanings in the masked terms. For euphemism identification task, it uses two classifiers- coarse classifier, to filter out the sentences and multi-class classifier, to identify the meaning of the masked terms.

The model proposed in this paper can detect previously unknown euphemisms, which makes it more effective for content moderation without manual efforts. It proposes a model with self-supervision but the model needs domain-specific training. The paper considers precision at k as the evaluation metric.

CHAPTER 4

DATA

LARGE DATASET

Context information contains the parent comment, discussion topics . The large dataset is included in the data folder in the form of two CSV files. In this there is a gn.csv file which comprises the output of context annotations and gc.csv comprises the in-content annotation.

H5

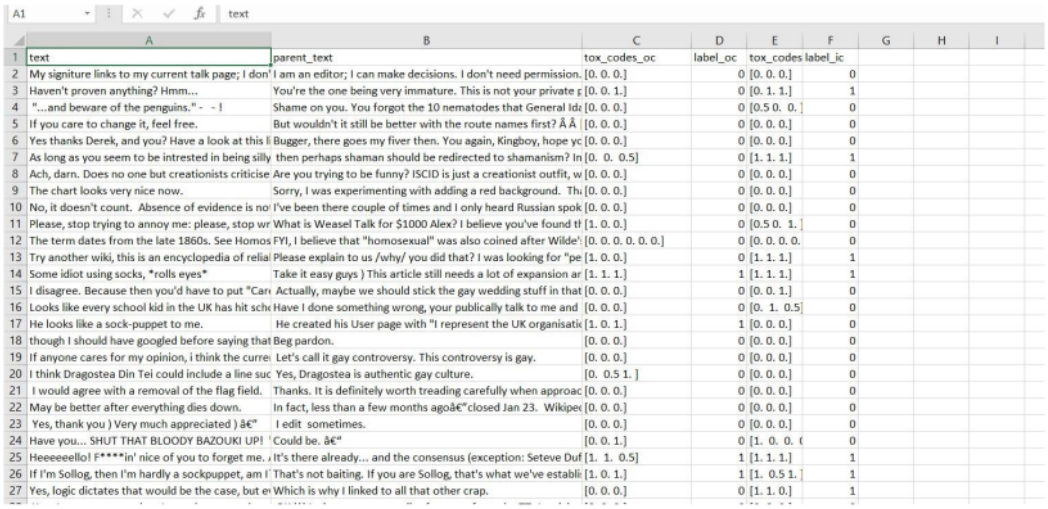
	A	B	C	D
1	id	text	api	parent
2	100012187.76918.76918	I've just seen it under many projects, but not every one. How about templates and	0.0431	
3	100047034.169781.169781	The US numbers were assumed to be 85. That assumption is probably false. But mo	0.1298	
4	100061014.7457.7457	Have you had a look at my proposals at Talk:List of MLB seasons? I'd leave the ML	0.0538	
5	100088289.43091.43091	Then how do we decide which is more commonly used? By how many countries us	0.0695	
6	100094503.94882.94882	Indexed here. Waltzing...#117, p. 339	0.1149	
7	100160006.13934.13934	Or mine to her.) They're due on the 13th ""or"" the 14th but I really hope they're a	0.0212	
8	100193848.79911.79911	Just because you are unaware of something and have not studied it doesn't mean	0.1185	
9	10021404.14154.14154	""Be bold"" is a fine principle, so long as it is invoked in good faith. What you call "	0.1243	
10	100246535.31824.31824	It makes perfect sense to me to remove images at some point. Earlier comments c	0.0725	
11	100319698.17785.17785	I would be glad to help out. Is the "randomized" rotation system okay with you?	0.0349	
12	100563582.5612.5612	You play as American, British, Canadian and Polish.	0.0913	
13	100701769.238494.238494	No, you're thinking straight. This page doesn't even list Anthroposophical Medicine	0.0954	
14	100703467.22360.22360	With the recent experience of List of volcanoes in Indonesia, I should create stubs	0.0443	
15	100884253.3120.3120	Got it, thanks. I've removed Yori's e-mail address from my talk page just in case it v	0.0653	
16	101017694.38900.38900	Since they use so many proxies, AIV is of limited value. Worst part is, I checked the	0.4897	
17	10125255.19992.20012	Opps! Looks like it was. Sorry about that. mav	0.1822	
18	101306002.81517.81517	As long as Apple calls it iPhone, it stays iPhone. Regardless of lawsuits, trademark i	0.0516	
19	101336603.23024.23024	It's ""your POV"" that people who held such a view were by definition "Serbian nati	0.1204	
20	101379818.63218.63218	Why not tie the prophets mentioned to the religions mentioned that they founded	0.0699	
21	101400162.2599.2599	Yup, you're right. I put one version in (I couldn't help thinking of Layla, by Eric Clapt	0.4192	
22	101609294.5232.5232	Good I forgot about Julia/Freddie and Jim/Mary. See you when you get there.	0.0379	
23	101632379.125954.125954	ArbCom makes that choice, not you or 'Elisson'.	0.0727	
24	101677450.31089.31089	Well, there's the trademark section at the bottom of Linux, but I think it's moot sin	0.0384	
25	101952497.3356.3356	Unfortunately, with the rewrite, I have to retract my previous blanket statement ar	0.0646	
26	101961911.2582.2582	The racial and central element of Lloyd's history must be mentioned. She is the ob	0.359	
27	102150304.16058.16058	I believe that newer TA routes were started without franchise.	0.0652	

SMALL DATASET

Civil commenst in Context is dataset derived from the civil comments dataset,where each post is labeled

Label are Non-toxic , toxic , unsure or very toxic

Only 0.07% of posts weer marked as UNSURE by a lable , indicates high confidence in annotations. These posts are excluded for analysis due to low count for generalization overall

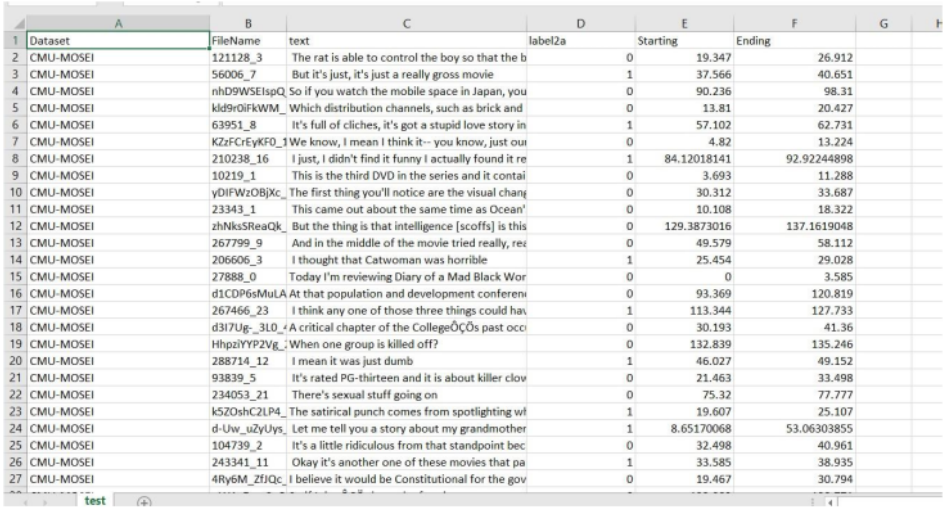


	text	parent_text	tox_codes_oc	label_oc	tox_codes	label_ic
2	My signiture links to my current talk page; I don'	I am an editor; I can make decisions. I don't need permission.	[0. 0. 0.]	0	[0. 0. 0.]	0
3	Haven't proven anything? Hmm...	You're the one being very immature. This is not your private ş	[0. 0. 1.]	0	[0. 1. 1.]	1
4	"...and beware of the penguins." - - !	Shame on you. You forgot the 10 nematodes that General Idı	[0. 0. 0.]	0	[0.5 0. 0.]	0
5	If you care to change it, feel free.	But wouldn't it still be better with the route names first? Ã Å	[0. 0. 0.]	0	[0. 0. 0.]	0
6	Yes thanks Derek, and you? Have a look at this li	Bugger, there goes my fiver then. You again, Kingboy, hope yc	[0. 0. 0.]	0	[0. 0. 0.]	0
7	As long as you seem to be intrested in being silly	then perhaps shaman should be redirected to shamanism? In	[0. 0. 0.5]	0	[1. 1. 1.]	1
8	Ach, darn. Does no one but creationists criticise	Are you trying to be funny? ISCID is just a creationist outfit, w	[0. 0. 0.]	0	[0. 0. 0.]	0
9	The chart looks very nice now.	Sorry, I was experimenting with adding a red background. Thi	[0. 0. 0.]	0	[0. 0. 0.]	0
10	No, it doesn't count. Absence of evidence is no	I've been there couple of times and I only heard Russian spok	[0. 0. 0.]	0	[0. 0. 0.]	0
11	Please, stop trying to annoy me: please, stop wr	What is Weasel Talk for $1000 Alex? I believe you've found th	[1. 0. 0.]	0	[0.5 0. 1.]	0
12	The term dates from the late 1860s. See Homos	FYI, I believe that "homosexual" was also coined after Wilde'	[0. 0. 0. 0. 0. 0.]	0	[0. 0. 0. 0.	0
13	Try another wiki, this is an encyclopedia of relia	Please explain to us /why/ you did that? I was looking for "pe	[1. 0. 0.]	0	[1. 1. 1.]	1
14	Some idiot using socks, *rolls eyes*	Take it easy guys) This article still needs a lot of expansion ar	[1. 1. 1.]	1	[1. 1. 1.]	1
15	I disagree. Because then you'd have to put "Can	Actually, maybe we should stick the gay wedding stuff in that	[0. 0. 0.]	0	[0. 0. 1.]	0
16	Looks like every school kid in the UK has hit sch	Have I done something wrong, your publically talk to me and	[0. 0. 0.]	0	[0. 1. 0.5]	0
17	He looks like a sock-puppet to me.	He created his User page with "I represent the UK organisatic	[1. 0. 1.]	1	[0. 0. 0.]	0
18	though I should have googled before saying that	Beg pardon.	[0. 0. 0.]	0	[0. 0. 0.]	0
19	If anyone cares for my opinion, i think the curre	Let's call it gay controversy. This controversy is gay.	[0. 0. 0.]	0	[0. 0. 0.]	0
20	I think Dragostea Din Tei could include a line suc	Yes, Dragostea is authentic gay culture.	[0. 0.5 1.]	0	[0. 0. 0.]	0
21	I would agree with a removal of the flag field.	Thanks. It is definitely worth treading carefully when approac	[0. 0. 0.]	0	[0. 0. 0.]	0
22	May be better after everything dies down.	In fact, less than a few months agoâ€"closed Jan 23. Wikipe	[0. 0. 0.]	0	[0. 0. 0.]	0
23	Yes, thank you) Very much appreciated) â€"	I edit sometimes.	[0. 0. 0.]	0	[0. 0. 0.]	0
24	Have you... SHUT THAT BLOODY BAZOUKI UP!	Could be. â€"	[0. 0. 1.]	0	[1. 0. 0. (	0
25	Heeeeeello! F****in' nice of you to forget me.	It's there already... and the consensus (exception: Seteve Duf	[1. 1. 0.5]	1	[1. 1. 1.]	1
26	If I'm Sollog, then I'm hardly a sockpuppet, am I	That's not baiting. If you are Sollog, that's what we've establi	[1. 0. 1.]	1	[1. 0.5 1.]	1
27	Yes, logic dictates that would be the case, but e	Which is why I linked to all that other crap.	[0. 0. 0.]	0	[1. 1. 0.]	1

SPOKEN UTTERANCE DATASET

Files included – metadata.csv, test.csv, train.csv, trigger_text.csv , valid.csv

It contains text data from the CMU-MOSEI dataset, with columns like FileName, text, and label2a indicating toxicity. It could be used for training and testing a model to detect social media toxicity using BERT (Bidirectional Encoder Representations from Transformers).



	Dataset	FileName	text	label2a	Starting	Ending
2	CMU-MOSEI	121128_3	The rat is able to control the boy so that the b	0	19.347	26.912
3	CMU-MOSEI	56006_7	But it's just, it's just a really gross movie	1	37.566	40.651
4	CMU-MOSEI	nhD9WSEIspQ	So if you watch the mobile space in Japan, you	0	90.236	98.31
5	CMU-MOSEI	kld9r0iFkWM_	Which distribution channels, such as brick and	0	13.81	20.427
6	CMU-MOSEI	63951_8	It's full of cliches, it's got a stupid love story in	1	57.102	62.731
7	CMU-MOSEI	KZzFCrEyKF0_1	We know, I mean I think it-- you know, just ou	0	4.82	13.224
8	CMU-MOSEI	210238_16	I just, I didn't find it funny I actually found it re	1	84.12018141	92.92244898
9	CMU-MOSEI	10219_1	This is the third DVD in the series and it contai	0	3.693	11.288
10	CMU-MOSEI	yDIFWzOBjXc_	The first thing you'll notice are the visual chang	0	30.312	33.687
11	CMU-MOSEI	23343_1	This came out about the same time as Ocean'	0	10.108	18.322
12	CMU-MOSEI	zhNksSReaQk_	But the thing is that intelligence [scoffs] is this	0	129.3873016	137.1619048
13	CMU-MOSEI	267799_9	And in the middle of the movie tried really, rea	0	49.579	58.112
14	CMU-MOSEI	206606_3	I thought that Catwoman was horrible	1	25.454	29.028
15	CMU-MOSEI	27888_0	Today I'm reviewing Diary of a Mad Black Wor	0	0	3.585
16	CMU-MOSEI	d1CDP6sMuLA	At that population and development conferen	0	93.369	120.819
17	CMU-MOSEI	267466_23	I think any one of those three things could hav	1	113.344	127.733
18	CMU-MOSEI	d3I7Ug-_3L0_4	A critical chapter of the CollegeÕÇÖs past occ	0	30.193	41.36
19	CMU-MOSEI	HhpziYYP2Vg_	When one group is killed off?	0	132.839	135.246
20	CMU-MOSEI	288714_12	I mean it was just dumb	1	46.027	49.152
21	CMU-MOSEI	93839_5	It's rated PG-thirteen and it is about killer clov	0	21.463	33.498
22	CMU-MOSEI	234053_21	There's sexual stuff going on	0	75.32	77.777
23	CMU-MOSEI	k5ZOshC2LP4_	The satirical punch comes from spotlighting wh	1	19.607	25.107
24	CMU-MOSEI	d-Uw_uZyUys_	Let me tell you a story about my grandmother	1	8.65170068	53.06303855
25	CMU-MOSEI	104739_2	It's a little ridiculous from that standpoint bec	0	32.498	40.961
26	CMU-MOSEI	243341_11	Okay it's another one of these movies that pa	1	33.585	38.935
27	CMU-MOSEI	4Ry6M_2fJQc_	I believe it would be Constitutional for the gov	0	19.467	30.794

CHAPTER 5

SYSTEM REQUIREMENT SPECIFICATION

5.1 SOFTWARE REQUIREMENTS

5.1.1 Deployment

- **Name and description**: **STREAMLIT**, a open source python library that allow users to create interactive web applications for data science and machine learning projects with simple python script.
- **Version**: 1.32.0
- **Operating system** : Compatible with major operating system such as windows , mac and linux .
- **Source** : Official documentation **https://docs.streamlit.io/**

5.1.2 Large Language Model

- **Name and description : BERT ,** a pre-trained natural language processing model developed by Google , capable of understanding context and semantics in text through bidirectional training on large amounts of unlabeled data .
- **Version** : BERT Large(LARGE , L-24,H-1024,16-LAYER,340M parameters)
- **Operating system** : compatible with major operating system such as windows , mac and linux .
- **Tools and libraries** : Hugging face transformer , transformers , scikit-learn , NLTK (Natural language toolkit) , Pandas , Matplotlib .

- **Source 1 :** google official documentation https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html
- **Source 2**: hugging face documentation https://huggingface.co/docs/transformers/model_doc/bert

5.1.3 Version Control

- **Name and description : Git and Github** is a web-based platform for hosting and collaboration on software development project using Git version control system.
- **Version :** 2.44.0(latest version)
- **Operating system** : Compatible with major operating system such as windows , mac and linux .
- **Tools :** Git bash , Git GUI
- **Source** : Git official documentation https://docs.github.com/en
- **Repository** : https://github.com/chethanv-20/toxicity_detection

5.2 FUNCTIONAL REQUIREMENTS

5.2.1 Validity test on input

- **Validity check** : Ensure the format can be only text or audio.
- **Error handling** : provide information message for incorrect input .
- **User guidance** : offer suggestion to correct input issues , enhancing user experience .

5.2.2 Sequence of operation

- Processing data from frontend to backend to process.
- **Large Language Model processing** : implement a stepwise approach for a large language model to analyse toxicity and capture the context .
- **Classification algorithm** : define a sequence of steps for the classification algorithm to classify text or audio accordingly.

5.2.3 Error Handling and Recovery

- **Detection** : implementation mechanism to to detect and handle errors related to incorrect or malformed genomic substring inputs .
- **Information Message**: Provide clear and informative error messages guiding users on how to rectify input issues .
- **Detection**: establish error detection mechanism in case the large language the model encounters difficulties in processing specific genomic substring .
- **Recovery Strategies** : Define recovery strategies , such as reprocessing or alternative model usage , to mitigate the impact of interference errors.

5.3 EXTERNAL INTERFACE REQUIREMENTS

5.3.1 User Interface

- **Input submission** : provide users an interface to enter audio and video options .
- **Result** : Show results of toxicity in detail and show resultant flags for further validation .
- **Error Message** : Clear and descriptive error messages for invalid inputs or system errors . Suggest for users to correct input mistakes .

5.3.2 Hardware Requirements

The hardware of the project is important to support the processing needs of large genomic data and machine learning studies . This section provides an overview of the logical and physical characteristics of the connections between software and hardware.

- **Device Type** :We thrive to build applications that can be accessed from various devices , including desktop , laptop and smartphone when deployed to the internet and optimal display across different screen sizes and resolutions.
- **Cloud service provider** : open-source cloud provider by AWS/Google Cloud and native cloud given by streamlit for deployment will be used on the scalability , flexibility and wide range of computing resources .
- **Hardware Dependencies** : The processing and analysis tasks are done on cloud or server , so that the hardware demands are not very high . This enables you to just interact with the browser .

5.4 NON FUNCTIONAL REQUIREMENTS

5.4.1 Performance Requirement (function specific performance):

- Large language model interface .
- Large language model interaction with BERT Model.
- with a good average response time per query .
- **Quality attribute** : speed and efficiency.

5.4.2 Scalability :

- Data Volume Scalability : The system must scale to accommodate growing data and ensure optimal performance as the dataset expands .
- **Quality Attribute** : Scalability , performance

5.4.3 Reliability , Availability and Fault tolerance :

- The system should gracefully handle faults , ensuring continued functionality in the presence of unforeseen errors or disruption .
- **Quality Attribute** : Robustness , reliability

5.4.4 Resource Utilization :

- **Objective** : The system should maintain optimal CPU and memory utilization to prevent resource saturation and ensure stability .
- **Quality Attribute** : Resource efficiency , Stability , database query response time , database queries should be executed within a time limit to maintain efficient data retrieval and processing , Database performance , efficiency .

5.4.5 Safety requirement

- Regulatory compliance
- the system should comply with relevant industry and regulatory standards.
- **Quality attribute** : compliance , security

5.5 SAFETY REQUIREMENTS

5.5.1 Data security , privacy and confidentiality

- all data submitted by the user must be treated confidentiality .
- **Security Measure** : Implement t encryption protocols for data and to prevent unauthorised access .

5.5.2 Privacy compliance

- use relevant data protection regulations and privacy standards.
- **Compliance Measure** : regularly check t data handling practices to ensure alignment with applicable privacy law.

5.5.3 User interaction , communication , and clear user guidance

- provide clear instruction and guidance to users regarding data submission and interpretation of results .
- **User interface design** : implement tooltips , help section and user-friendly documentations to assist users.

5.5.4 Natural Language interaction safety

- safeguard against potentially sensitive or inappropriate language in natural language interactions .
- **Filtering mechanism** : integrate filters to identity and handle inappropriate language , maintaining a respectful and secure environment .

CHAPTER 6

SYSTEM DESIGN

6.1 Architecture Diagrams

6.1.1 High Level System Design

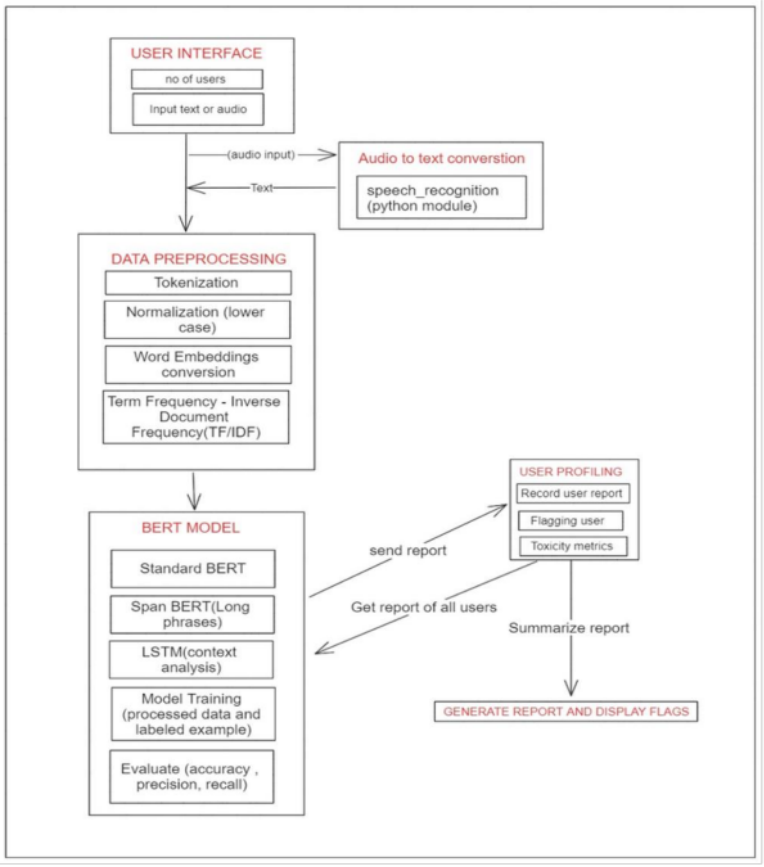


Fig 6.1.1 High Level System Design

6.1.2 Master Class Diagram

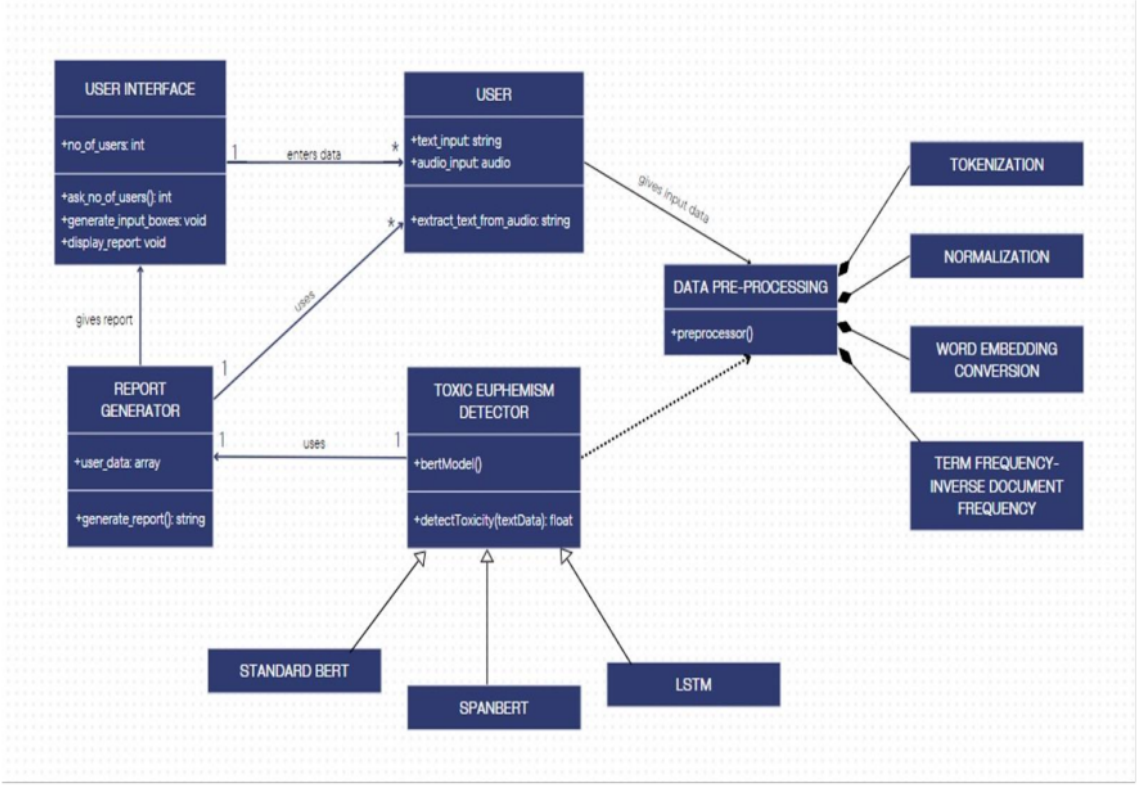


Fig 6.1.2 Master Class Diagram

6.1.3 Architecture Diagram

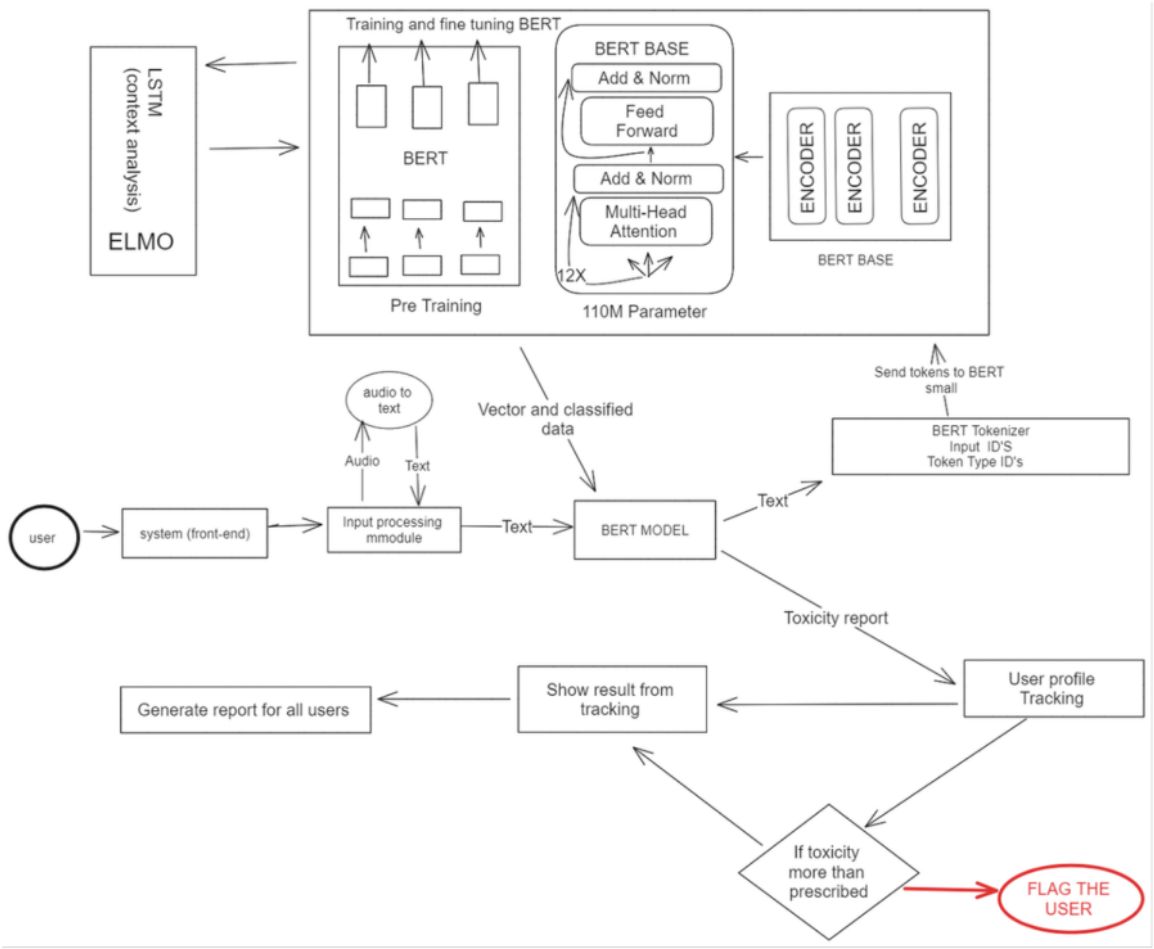


Fig 6.1.3 Architecture Diagram

6.1.4 External Interfaces

Main Page:

Fig 6.1.4.1

Selecting number of users:

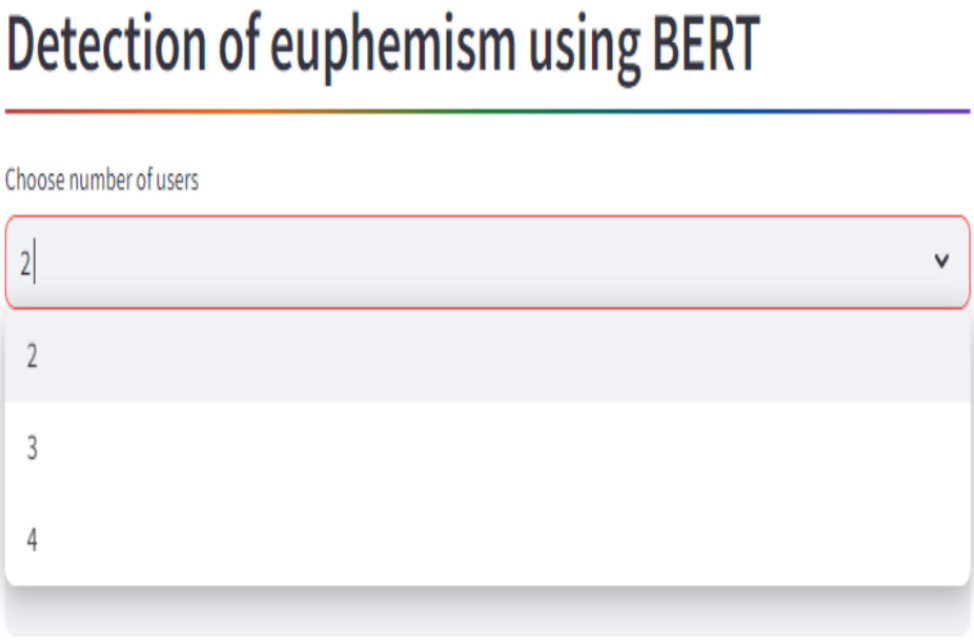


Fig 6.1.4.2

User Page:

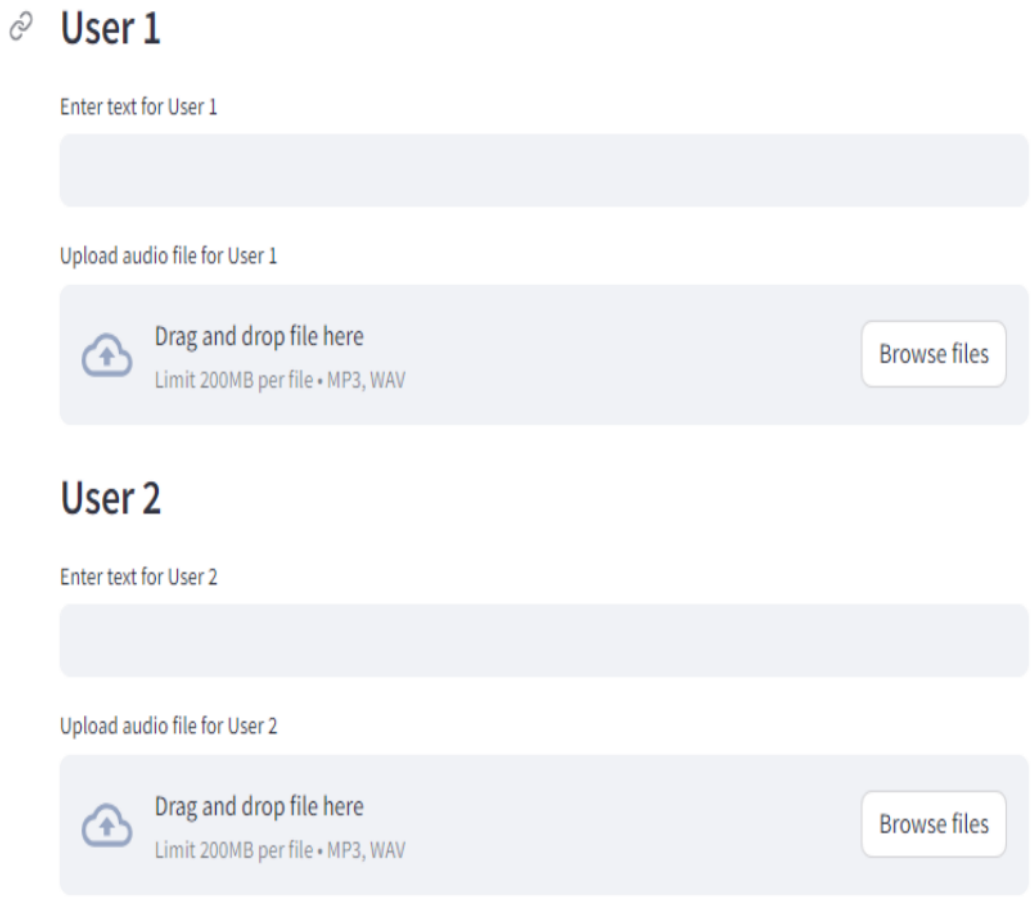


Fig 6.1.4.3

Test Report:

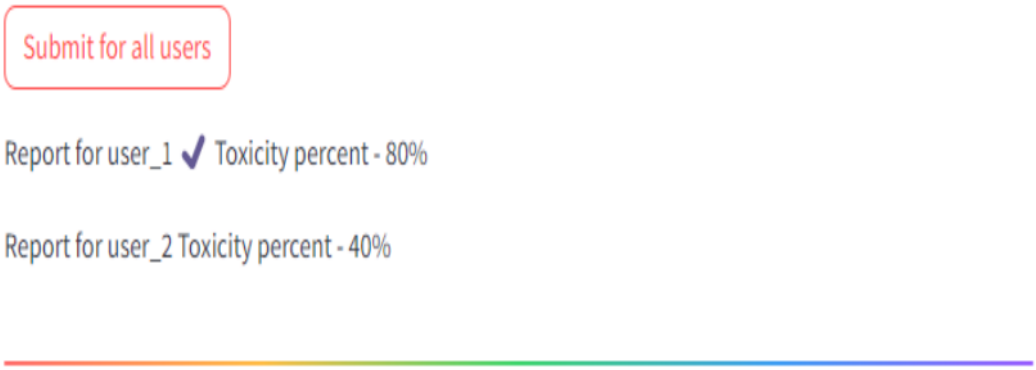


Fig 6.1.4.4

6.2 Design Considerations

6.2.1 Design Goals

- **Speed -** Try to generate the output of all phases in a reasonable amount of time.
- **Scalability -** should be scalable for any further additional updates of features in future.
- **Accuracy -** models which we are using should reach high accuracy.
- **Portability -** the system which we are trying to build will be platform-independent, that will work on different systems.

6.2.2 Architecture Choice

Our project architecture choice is a client-server architecture. In which the client will be a web browser or mobile app, that allows users to give input messages for toxicity analysis.
Then the user input messages are sent to the server, where the analysis for the toxicity detection is done or performed using the models which we have considered such as BERT and LSTM models. After the analysis is done the server will send back the toxicity analysis results to the client, which will be checked by the client or user.

Pros -

- The architecture that is chosen will allow the server to handle multiple client requests, so it will be easier to scale.
- It also provides centralized processing, which can be beneficial for complex hate language detection algorithms.
- Gives security where server can implement security measures such as encryption and authentication, so that the sensitive data is protected.
- Maintenance is easy, as updates and maintenance tasks can be performed centrally on the server, making it easier to manage the system.

Cons -

- Latency may be involved as the request has to be sent to server and processed and returned, so there may be latency in response time.
- Dependency on Server, client depends or relies on the server, so if the server is down or has any issues then the analysis might not take place.
- Complexity may get added as we need to develop both on client side as well as server side.

6.2.3 General Constraints, Assumptions and Dependencies

Computational Resources

- Data Availability
- External Libraries
- Evaluation Metrics

6.3 Design Details

6.3.1 Novelty :-

Input format - We have included audio data in addition to textual data.

Handling Euphemism - Our model focuses on detecting euphemistic language,which is a new and challenging aspect of toxicity detection. This includes symbols *, #, !, alternate words, and other similar expressions that users may employ to convey toxicity indirectly.

Context of the conversation - we will try to understanding the context of the conversation so that we can anlayze language and detecting toxicity properly.

6.3.2 Interoperability :-

Making sure that the system can integrate with the content management systems which are used in different online platforms, like social media networks.

6.3.3 Performance :-

Response time – trying to generate the system output within a time frame or less time.
Scalability - Scaling properly to satisfy the increasing user demand without sacrificing performance.

6.3.4 Maintainability :-

Modularity - Designing a model with such components that can be easily modified and replaced without affecting the other parts of the system.
Documentation- Providing a detailed documentation that explains the project architecture, functionality and usage instructions so it would be easy to understand the project.
Code quality– keeping the code in proper format and documenting it so it will be easier for us to understand and modify the code when needed.

6.3.5 Security :-

Giving data security and privacy high priority, by taking strong methods to protect user inputs, result analysis and any sensitive information.

6.3.6 Portability :-

Our project is a platform-independent application that is able to work together with different operating systems and web browsers.

6.3.7 Reusability :-

Components and protocols in the architecture are designed to be reused, making it flexible for deployment in different network environments and scenarios.

The architecture can be adapted and scaled for use in various organizational settings.

CHAPTER 7

CONCLUSION OF CAPSTONE PHASE-1

In this phase of Capstone, we have explored various models and techniques available for detection of euphemism and toxic content. Our literature survey focused on models like BERT, SpanBERT, LSTM and others.

- We have only considered the models which we think might be best suited to our research. This helps us focus on enhancing the model to increase the efficiency and accuracy.
- We have defined all the diagrams required for the project in review 3 of phase 1. We have also described architecture that includes a detailed view of bert architecture and system architecture .
- We came up with the prototype for user interface using streamlit , it includes selecting numbers , adding text and audio and a sample output with flagged users with a threshold percentage .
- We conclude our phase 1 here and our challenge is to curate a suitable dataset(that has euphemistic language in it) that is most suitable for our projects from already existing dataset .

CHAPTER 8

PLAN OF WORK FOR CAPSTONE PHASE-2

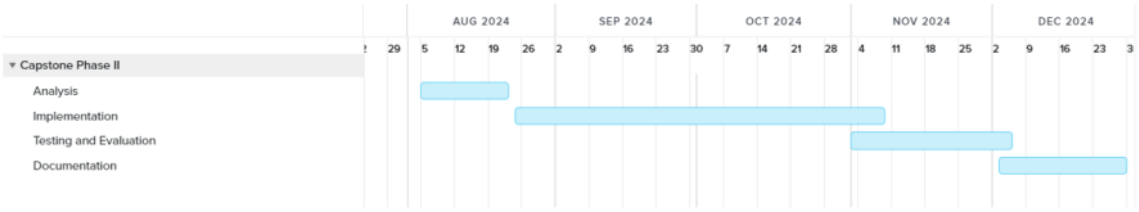


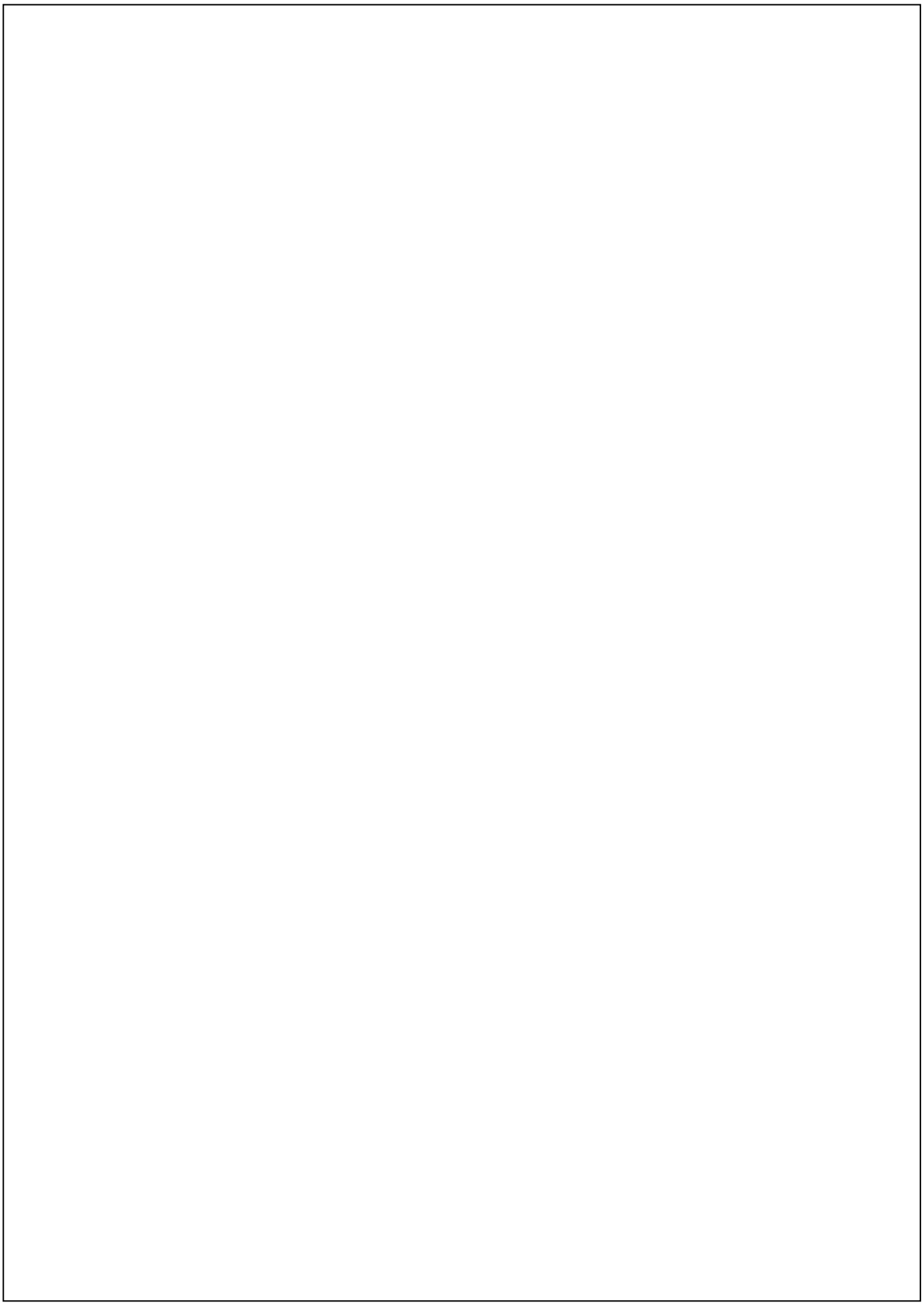
Fig 8 Plan of Work for Capstone Phase-2

- Above, is a visual overview of the Project Phase-2 schedule has been presented using a Gantt Chart. A Gantt chart is a bar chart that shows the planned timeline for the project's phases and key deliverables.

- Vertical Axis: The vertical axis lists the major planned milestones of phase 2 of the project. They include Technology Research, Implementation, Testing, Validation and Documentation.

- Horizontal Axis: The horizontal axis represents the project timeline with the time frame ranging from August'24 to December'24.

- Task Durations: The shaded sections within the cells represent the planned duration for each project phase.



Final_report_team66[2] (1)

ORIGINALITY REPORT

8%	3%	0%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to PES University Student Paper	4%
2	github.com Internet Source	1%
3	www.aclweb.org Internet Source	<1%
4	elar.urfu.ru Internet Source	<1%
5	www.coursehero.com Internet Source	<1%
6	Submitted to University of Greenwich Student Paper	<1%
7	upcommons.upc.edu Internet Source	<1%
8	ijarcce.com Internet Source	<1%
9	hasifnoorprojects.weebly.com Internet Source	<1%

10	Submitted to Kingston University Student Paper	<1%
11	cran.rstudio.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		