TOXIC STATEMENT ANALYSER

Suraj Desai
NLP Info 7610

Prof. Subrata Das

PRESENTED BY SURAJ DESAI 002778614

PROBLEM STATEMENT

In the world where toxic comments on social media are on rise it ss importtant to flag them and give appropriate reason after flagging.

PRESENTED BY SURAJ DESAI 002778614

APPROACH TO TOXIC ANALYZER

EXTRACTIVE QUESTION ANSWERING

Extractive Question Answering involves finding exact answers to questions from a given text

PROCESS OVERVIEW

- Input: Text and a Question.
- Analysis: AI identifies relevant parts of the text.
- Output: Direct answer from the text.

EXAMPLE

TEXT: You are a bad person. I saw this person while I was taking a morning walk.

QUESTION: What toxic statement is used in this text?

ANSWER: You are a bad person.

APPROACH TO TOXIC ANALYZER

TEXT GENERATION

Text generation is the process of automatically creating human-like text using artificial intelligence. Such models are also called as casual learning models as they use autoregressive aprroach for training purpose

PROCESS OVERVIEW

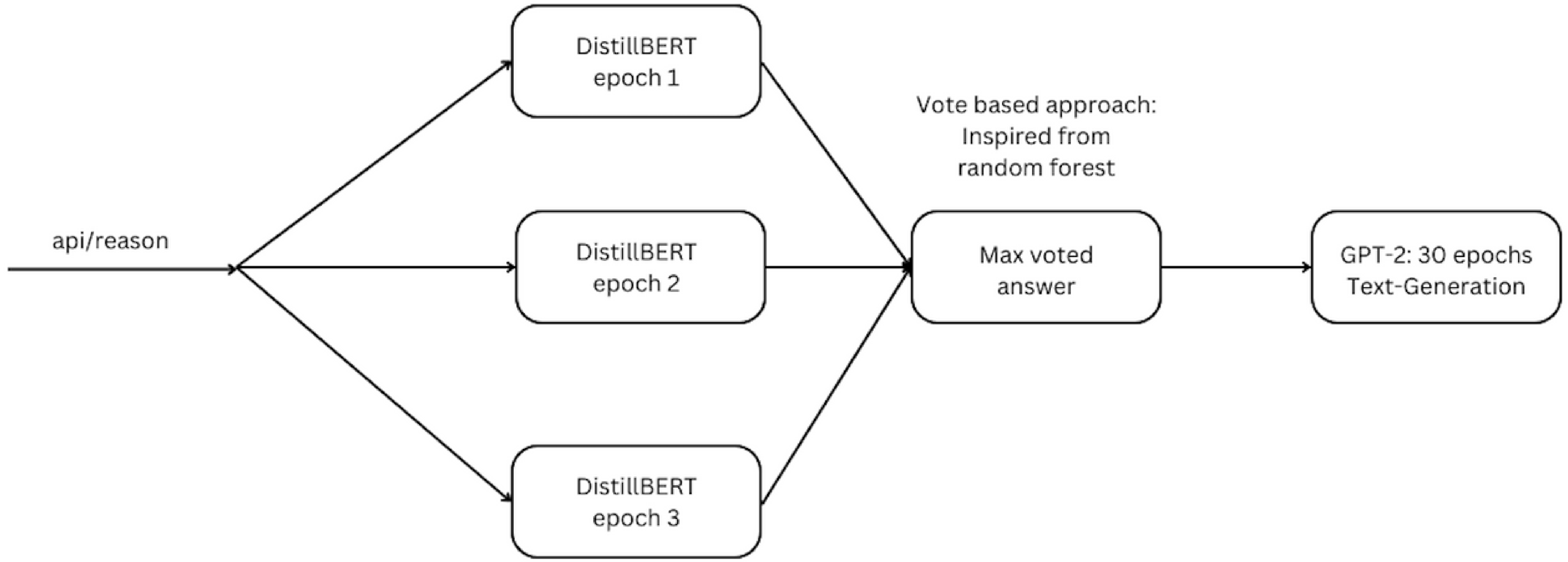
- Input: A prompt or data.
- Processing: AI uses language models to generate text.
- Output: Coherent, contextually relevant text.

EXAMPLE

- Prompt: "Write a short story about a space adventure."
- Generated Text: [Include a brief, generated text snippet]

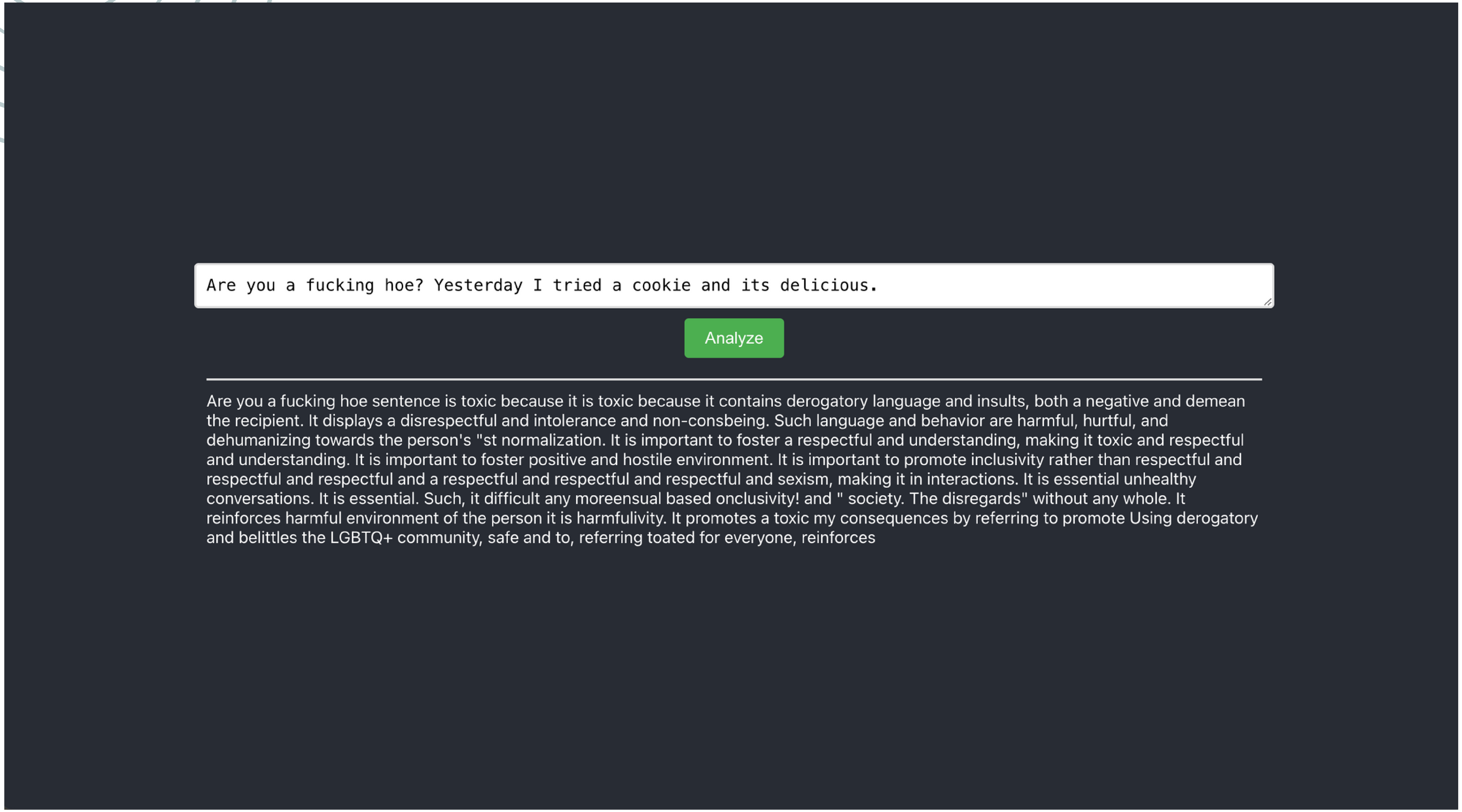
PRESENTED BY SURAJ DESAI 002778614

ARCHITECTURE



PRESENTED BY SURAJ DESAI 002778614

USER EXPERIENCE AND INTERFACE DESIGN

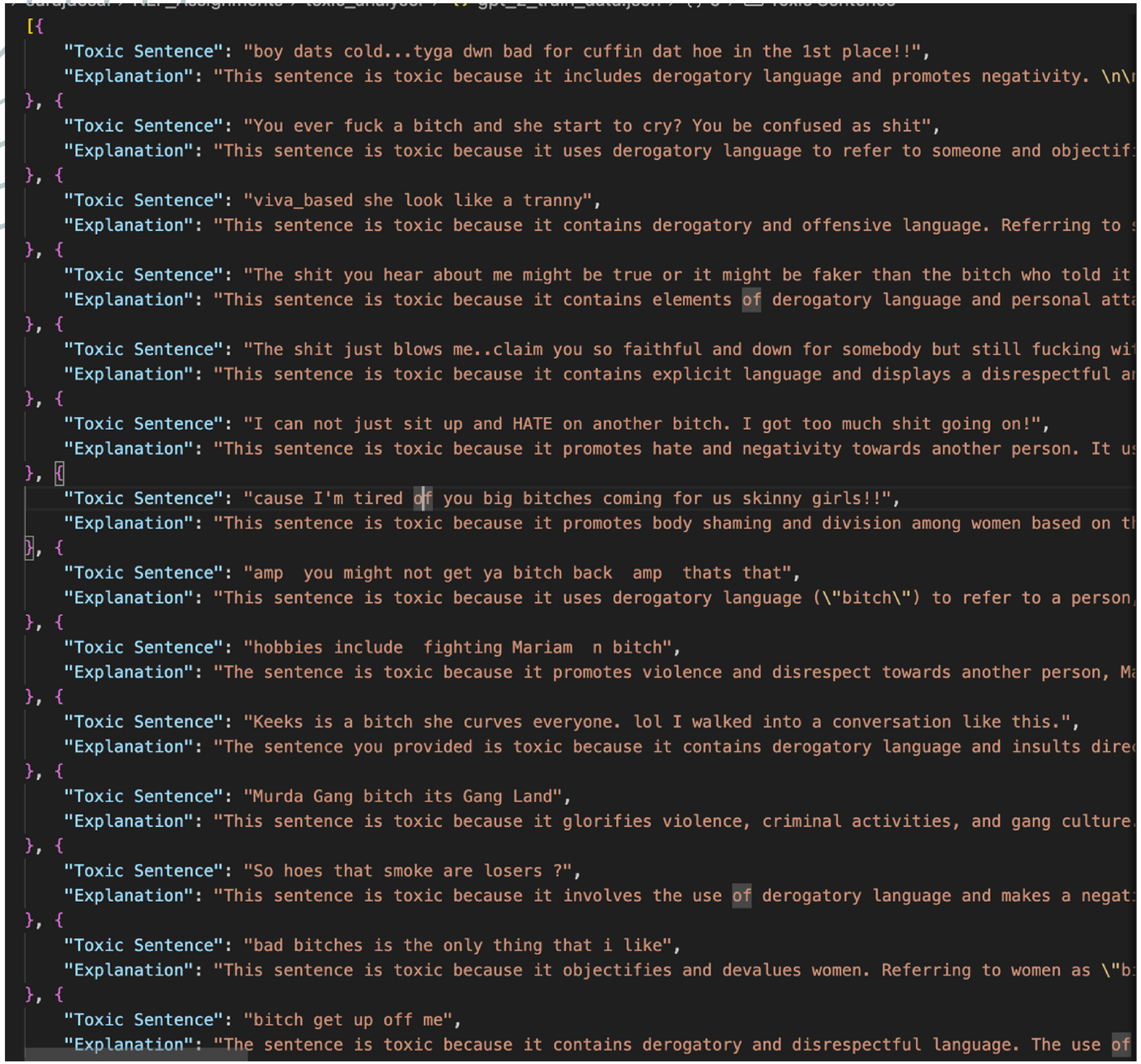


Are you a fucking hoe? Yesterday I tried a cookie and its delicious.

Are you a fucking hoe sentence is toxic because it is toxic because it contains derogatory language and insults, both a negative and demean the recipient. It displays a disrespectful and intolerance and non-consbeing. Such language and behavior are harmful, hurtful, and dehumanizing towards the person's "st normalization. It is important to foster a respectful and understanding, making it toxic and respectful and understanding. It is important to foster positive and hostile environment. It is important to promote inclusivity rather than respectful and respectful and respectful and a respectful and respectful and respectful and sexism, making it in interactions. It is essential unhealthy conversations. It is essential. Such, it difficult any moreensual based onclusivity! and " society. The disregards" without any whole. It reinforces harmful environment of the person it is harmfulivity. It promotes a toxic my consequences by referring to promote Using derogatory and belittles the LGBTQ+ community, safe and to, referring toated for everyone, reinforces

PRESENTED BY SURAJ DESAI 002778614

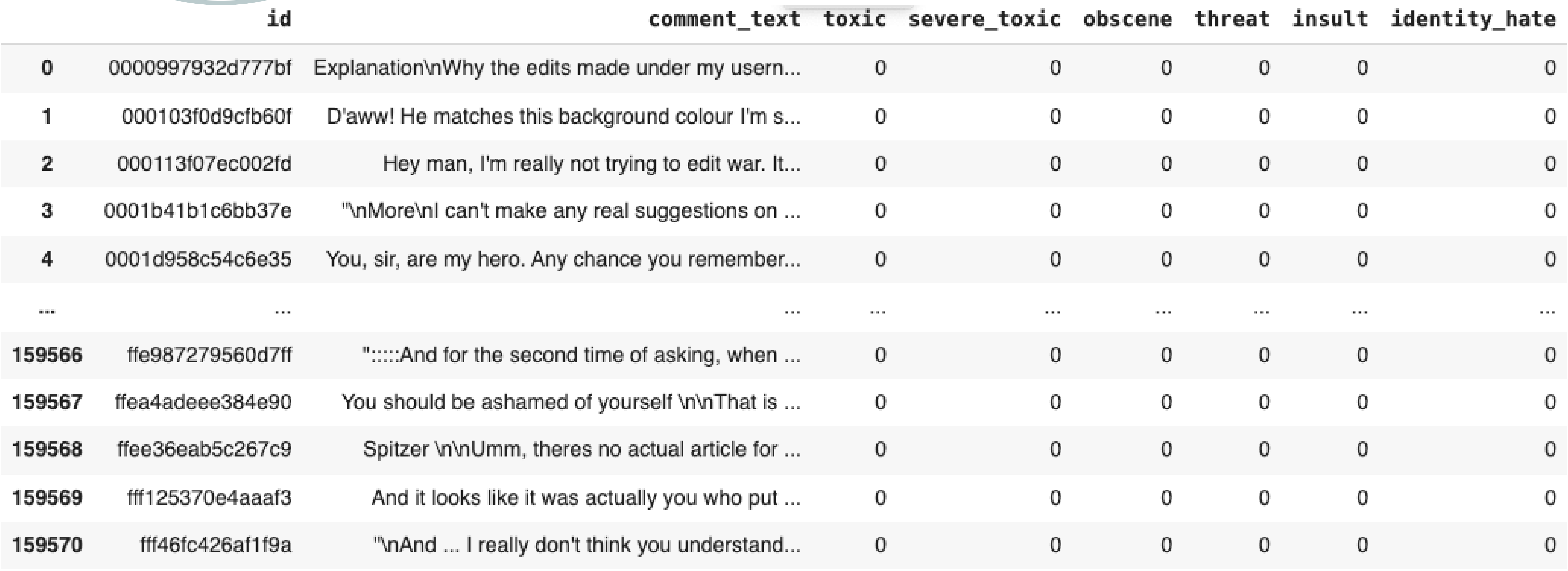
GPT-2 TRAIN DATA



```
[{
    "Toxic Sentence": "boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!",
    "Explanation": "This sentence is toxic because it includes derogatory language and promotes negativity. \n\r
}, {
    "Toxic Sentence": "You ever fuck a bitch and she start to cry? You be confused as shit",
    "Explanation": "This sentence is toxic because it uses derogatory language to refer to someone and objectif
}, {
    "Toxic Sentence": "viva_based she look like a tranny",
    "Explanation": "This sentence is toxic because it contains derogatory and offensive language. Referring to s
}, {
    "Toxic Sentence": "The shit you hear about me might be true or it might be faker than the bitch who told it
    "Explanation": "This sentence is toxic because it contains elements of derogatory language and personal atta
}, {
    "Toxic Sentence": "The shit just blows me..claim you so faithful and down for somebody but still fucking wit
    "Explanation": "This sentence is toxic because it contains explicit language and displays a disrespectful ar
}, {
    "Toxic Sentence": "I can not just sit up and HATE on another bitch. I got too much shit going on!",
    "Explanation": "This sentence is toxic because it promotes hate and negativity towards another person. It us
}, {
    "Toxic Sentence": "cause I'm tired of you big bitches coming for us skinny girls!!",
    "Explanation": "This sentence is toxic because it promotes body shaming and division among women based on th
}, {
    "Toxic Sentence": "amp  you might not get ya bitch back  amp  thats that",
    "Explanation": "This sentence is toxic because it uses derogatory language (\"bitch\") to refer to a person,
}, {
    "Toxic Sentence": "hobbies include  fighting Mariam  n bitch",
    "Explanation": "The sentence is toxic because it promotes violence and disrespect towards another person, Ma
}, {
    "Toxic Sentence": "Keeks is a bitch she curves everyone. lol I walked into a conversation like this.",
    "Explanation": "The sentence you provided is toxic because it contains derogatory language and insults dire
}, {
    "Toxic Sentence": "Murda Gang bitch its Gang Land",
    "Explanation": "This sentence is toxic because it glorifies violence, criminal activities, and gang culture.
}, {
    "Toxic Sentence": "So hoes that smoke are losers ?",
    "Explanation": "This sentence is toxic because it involves the use of derogatory language and makes a negat
}, {
    "Toxic Sentence": "bad bitches is the only thing that i like",
    "Explanation": "This sentence is toxic because it objectifies and devalues women. Referring to women as \"bi
}, {
    "Toxic Sentence": "bitch get up off me",
    "Explanation": "The sentence is toxic because it contains derogatory and disrespectful language. The use of
```

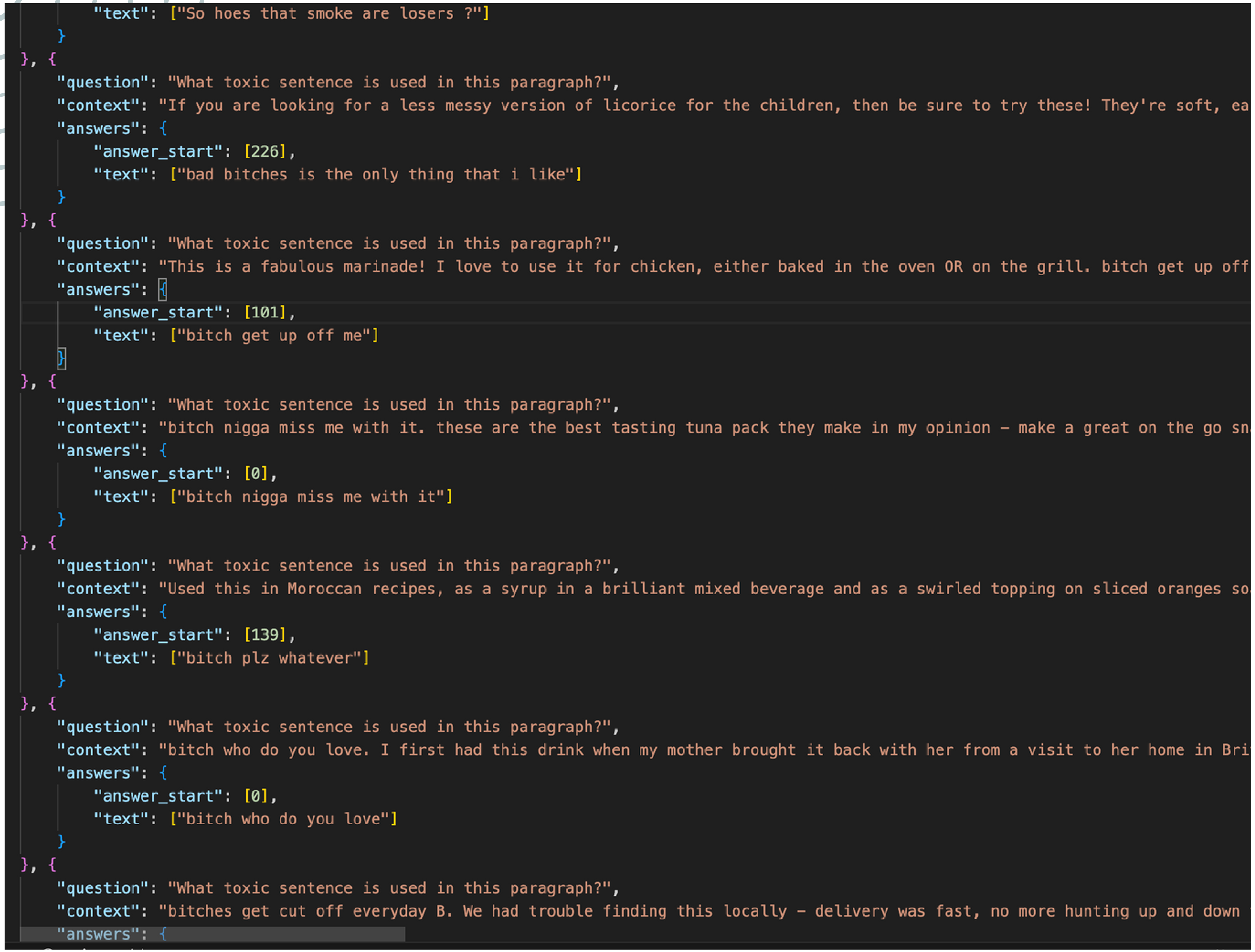
PRESENTED BY SURAJ DESAI 002778614

DISTILLBERT MODEL TRAIN DATA



	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
159566	ffe987279560d7ff	":::::And for the second time of asking, when ...	0	0	0	0	0	0
159567	ffea4adeee384e90	You should be ashamed of yourself \n\nThat is ...	0	0	0	0	0	0
159568	ffee36eab5c267c9	Spitzer \n\nUmm, theres no actual article for ...	0	0	0	0	0	0
159569	fff125370e4aaaf3	And it looks like it was actually you who put ...	0	0	0	0	0	0
159570	fff46fc426af1f9a	"\nAnd ... I really don't think you understand...	0	0	0	0	0	0

DISTILLBERT MODEL TRAIN DATA



```
        "text": ["So hoes that smoke are losers ?"]
    }
}, {
    "question": "What toxic sentence is used in this paragraph?",
    "context": "If you are looking for a less messy version of licorice for the children, then be sure to try these! They're soft, ea
    "answers": {
        "answer_start": [226],
        "text": ["bad bitches is the only thing that i like"]
    }
}, {
    "question": "What toxic sentence is used in this paragraph?",
    "context": "This is a fabulous marinade! I love to use it for chicken, either baked in the oven OR on the grill. bitch get up off
    "answers": {
        "answer_start": [101],
        "text": ["bitch get up off me"]
    }
}, {
    "question": "What toxic sentence is used in this paragraph?",
    "context": "bitch nigga miss me with it. these are the best tasting tuna pack they make in my opinion - make a great on the go sn
    "answers": {
        "answer_start": [0],
        "text": ["bitch nigga miss me with it"]
    }
}, {
    "question": "What toxic sentence is used in this paragraph?",
    "context": "Used this in Moroccan recipes, as a syrup in a brilliant mixed beverage and as a swirled topping on sliced oranges so
    "answers": {
        "answer_start": [139],
        "text": ["bitch plz whatever"]
    }
}, {
    "question": "What toxic sentence is used in this paragraph?",
    "context": "bitch who do you love. I first had this drink when my mother brought it back with her from a visit to her home in Bri
    "answers": {
        "answer_start": [0],
        "text": ["bitch who do you love"]
    }
}, {
    "question": "What toxic sentence is used in this paragraph?",
    "context": "bitches get cut off everyday B. We had trouble finding this locally - delivery was fast, no more hunting up and down
    "answers": {
```

CHALLENGES AND LIMITATIONS

- Training gpt-2 using hugging face model for reasoning purpose was very difficult. I used only 200 datapoints for training and still it took 12 hrs despite having 900 gb ram and 60 GPU’s on work. If training would have been easy the results could have been much better.

- I had to generate the data for text generation on my own as its not available on internet and when I asked chatgpt to generate it was flagging me as it was violating their poilcy. I had more data it would also would have helped in better results

CONCLUSION

The tool harnesses the power of GPT-2 and DIstillBert and has produced promising results in identifying and explaining toxic content, paving the way for healthier online interactions and informed content moderation.