



QUICKSCAN - CANVAS Learning Model for Classifying Tweet's Content


NAME: Machine Learning Model for Classifying Tweet's Content 

DATE: September 5, 2024 4:26 AM

DESCRIPTION OF TECHNOLOGY
 A model which classifies the content of text (tweets specifically) in 3 classes: Neutral Speech, Offensive Speech and Hate Speech; using a dataset with already annotated tweets for training and testing purposes.

HUMAN VALUES 


The hate speech detection technology impacts user identity by potentially altering how individuals express themselves online. For friends using social media, it can enhance relationships by creating a safer communication space free from hate speech, which may encourage more open interaction. However, if the technology over-moderates, it could detract from authentic expression, leading to frustrations or misunderstanding about why certain content is censored.

TRANSPARENCY 


Yes, the operation and goals of the hate speech detection technology are broadly explained to users and stakeholders. We detail the types of data used, the purpose of the technology, and the methods employed in detecting hate speech and offensive language on our website and in user documentation. However, while we strive for transparency, the specific inner workings of the machine learning algorithms like how they process data and arrive at specific determinations are not fully disclosed for two reasons: first, t...

IMPACT ON SOCIETY 


The problem this technology addresses is the prevalence of hate speech and offensive language on social media, which undermines safety and inclusivity online. This "pain" is felt by social media users exposed to harmful content, causing emotional distress and discouraging engagement. Platform administrators also struggle to moderate content effectively without infringing on free speech. This technology aims to ease these challenges by enhancing the ability to detect an...

STAKEHOLDERS 


- Social Media Platforms
- Blogs, Forums

SUSTAINABILITY 


The hate speech detection technology primarily operates in cloud environments, which do consume significant amounts of energy. However, we partner with cloud service providers committed to sustainability, utilizing energy-efficient data centers often powered by renewable energy sources. While the model is computationally intensive, efforts are made to optimize algorithms for better efficiency, reducing unnecessary computations and data transfers. We continuously explore advancements in technology that could...

HATEFUL AND CRIMINAL ACTORS 


Hate speech detection technology, while designed to improve social media safety, could potentially be misused in ways that infringe on privacy or freedom of speech. For instance, governments or other entities might exploit the technology to surveil and suppress dissenting voices under the guise of moderating hate speech. This could lead to censorship of political opposition or marginalized groups, impacting their right to express dissent. Additionally, the technology could be manipulated to wrongfully flag content, causing harassment...

DATA 


Yes, the technology is designed with an awareness of data limitations such as subjectivity, incompleteness, bias, and the complexity of language. We recognize that the dataset may not fully represent all nuances of language and cultural expressions, which can lead to inaccuracies in hate speech detection. To mitigate these issues, the model incorporates diverse data sources to balance representation, employs techniques to adjust for class imbalance, and includes regular updates and feedback mechanisms to refine its...

FUTURE 

With widespread adoption, this hate speech detection technology could significantly alter online communication norms, encouraging more respectful interactions and possibly reducing the prevalence of harmful content on social media. As communities adapt to these moderated environments, there might be a shift towards more positive online behavior. However, there's also potential for overreliance on automated moderation, possibly leading to suppression of legitimate speech or failure to capture nuanced expressions...

PRIVACY 

The hate speech detection technology inherently processes personal data as it analyzes content generated by individuals on social media platforms. This includes text from tweets, which may contain identifiable information about the user, potentially revealing their identity, location, or personal beliefs. Although the primary intent is to assess the content for hate speech or offensive language, the nature of social media data means that personal and sometimes sensitive information...

INCLUSIVITY 


Yes, the technology could potentially have built-in biases due to the nature of its data sources and the design of its algorithms. The dataset, primarily composed of Twitter data, may inherently reflect the biases of the users who generated the tweets, including geographical, linguistic, or cultural biases. Additionally, the annotators who labeled the dataset might have introduced subjective interpretations of what constitutes hate speech or offensive language. Recognizing these limitations, the design includes regular reviews and...

FIND US ON WWW.TICT.IO

THIS CANVAS IS PART OF THE TECHNOLOGY IMPACT CYCLE TOOL. THIS CANVAS IS THE RESULT OF A QUICKSCAN. YOU CAN FILL OUT THE FULL TICT ON WWW.TICT.IO


  

QUICKSCAN - CANVAS HEAPS Big Model for Classifying Tweet's Content

NAME: Machine Learning Model for Classifying Tweet's Content 

DATE: September 5, 2024 4:26 AM


DESCRIPTION OF TECHNOLOGY
 A model which classifies the content of text (tweets specifically) in 3 classes: Neutral Speech, Offensive Speech and Hate Speech; using a dataset with already annotated tweets for training and testing purposes.

HUMAN VALUES 

How is the identity of the (intended) users affected by the technology?


To help you answer this question think about sub questions like:

- If two friends use your product, how could it enhance or detract from their relationship?
- Does your product create new ways for people to interact?...

TRANSPARENCY 


Is it explained to the users/stakeholders how the technology works and how the business model works?

- Is it easy for users to find out how the technology works?
- Can a user understand or find out why your technology behaves in a certain way?
- Are the goals explained?
- Is the idea of the technology explained?
- Is the technology company transparent about the way their...

IMPACT ON SOCIETY 

What is exactly the problem? Is it really a problem? Are you sure?


Can you exactly define what the challenge is? What problem (what 'pain') does this technology want to solve? Can you make a clear definition of the problem? What 'pain' does this technology want to ease? Whose pain? Is it really a problem? For who? Will solving the problem make the world better? Are you sure? The problem definition will help you to determine...

STAKEHOLDERS 

Who are the main users/targetgroups/stakeholders for this technology? Think about the intended context by...


When thinking about the stakeholders, the most obvious one are of course the intended users, so start there. Next, list the stakeholders that are directly affected. Listing the users and directly affected stakeholders also gives an impression of the intended context of the technology.

...

SUSTAINABILITY 


In what way is the direct and indirect energy use of this technology taken into account?

One of the most prominent impacts on sustainability is energy efficiency. Consider what service you want this technology to provide and how this could be achieved with a minimal use of energy. Are improvements possible?

HATEFUL AND CRIMINAL ACTORS 

In which way can the technology be used to break the law or avoid the consequences of breaking the law?


Can you imagine ways that the technology can or will be used to break the law? Think about invading someone's privacy. Spying. Hurting people. Harassment. Steal things. Fraud/identity theft and so on. Or will people use the technology to avoid facing the consequences of breaking the law (using trackers to evade speed radars or using bitcoins to launder...)

DATA 

Are you familiar with the fundamental shortcomings and pitfalls of data and do you take this sufficiently into...


There are fundamental issues with data. For example:

- Data is always subjective;
- Data collections are never complete;
- Correlation and causation are tricky concepts;
- Data collections are often biased;...

FUTURE 


What could possibly happen with this technology in the future?

Discuss this quickly and note your first thoughts here. Think about what happens when 100 million people use your product. How could communities, habits and norms change?

PRIVACY 

Does the technology register personal data? If yes, what personal data?

If this technology registers personal data you have to be aware of privacy legislation and the concept of privacy. Think hard about this question. Remember: personal data can be interpreted in a broad way. Maybe this technology does not collect personal data, but can be used to assemble personal data. If the technology collects special personal data (like...

INCLUSIVITY 

Does this technology have a built-in bias?

Do a brainstorm. Can you find a built-in bias in this technology? Maybe because of the way the data was collected, either by personal bias, historical bias, political bias or a lack of diversity in the people responsible for the design of the technology? How do you know this is not the case? Be critical. Be aware of your own biases....

FIND US ON WWW.TICT.IO

THIS CANVAS IS PART OF THE TECHNOLOGY IMPACT CYCLE TOOL. THIS CANVAS IS THE RESULT OF A QUICKSCAN. YOU CAN FILL OUT THE FULL TICT ON WWW.TICT.IO