Phuoc Le, Gil Rabara, Dilnoza Saidova, Vivian Tran
TCSS 456: Natural Language Processing
April 23, 2023

Dataset

Dataset

The Jigsaw Toxic Comment Classification Challenge dataset available on Kaggle is a collection of comments from Wikipedia's talk page edits. The comments are labeled based on their degree of toxicity, with a focus on six categories: toxic, severe toxic, obscene, threat, insult, and identity hate. The goal of the competition was to develop models that can accurately classify comments based on their level of toxicity

Size and Format

The dataset contains two CSV files: "train.csv" and "test.csv". The "train.csv" file contains over 159,000 comments, along with their corresponding labels. The "test.csv" file contains over 63,000 comments, but without their corresponding labels, as they are reserved for evaluation by the competition organizers.

Each row of the "train.csv" and "test.csv" files contains the following columns:

- "id": unique identifier for each comment
- "comment_text": the text of the comment
- "toxic": binary label indicating whether the comment is toxic or not
- "severe_toxic": binary label indicating whether the comment is severely toxic or not
- "obscene": binary label indicating whether the comment is obscene or not
- "threat": binary label indicating whether the comment contains threats or not
- "insult": binary label indicating whether the comment is insulting or not
- "identity_hate": binary label indicating whether the comment expresses hatred towards a specific group or identity or not

Preprocessing Information

The Kaggle dataset requires preprocessing before it can be used for training an NLP model for toxic text classification. Following are the preprocessing steps that will be taken:

1. Combine the "ham" and "spam" datasets into one dataset.
2. Remove any duplicate text.
4. Remove any HTML tags and non-alphabetic characters.
5. Tokenize the text into individual words.
6. Remove stop words (i.e., commonly-used words, "the", "and", etc.).
7. Convert all words to lowercase.
8. Stem or lemmatize the words (reduce them to their base form) to normalize the text.

Training Data:

	A	B	C	D	E	F	G	H	I
1	id	comment_	toxic	severe_to	obscene	threat	insult	identity_hate	
2	000099793	Explanati	0	0	0	0	0	0	
3	000103f0d	D'aww! He	0	0	0	0	0	0	
4	000113f07	Hey man,	0	0	0	0	0	0	
5	0001b41b:	"	0	0	0	0	0	0	
6	0001d958c	You, sir, a	0	0	0	0	0	0	
7	00025465c	"	0	0	0	0	0	0	
8	0002bcb3c	COCKSUCI	1	1	1	0	1	0	
9	00031b1e9	Your vand	0	0	0	0	0	0	
10	00037261f	Sorry if th	0	0	0	0	0	0	
11	00040093b	alignment	0	0	0	0	0	0	

Testing Data:

	A	B	C	D	E	F	G	H	I	J	K
1	id	comment_text									
2	00001cee3	Yo bitch Ja Rule is more succesful then you'll ever be whats up with you and hating you sad mofuckas...i									
3	000024786	== From									
4	00013b17a	"									
5	00017563c	:If you have a look back at the source, the information I updated was the correct form. I can only guess t									
6	00017695a	I don't anonymously edit articles at all.									
7	0001ea871	Thank you for understanding. I think very highly of you and would not revert without discussion.									
8	00024115c	Please do not add nonsense to Wikipedia. Such edits are considered vandalism and quickly undone. If y									
9	000247e83	:Dear god this site is horrible.									
10	00025358c	"									
11	00026d109	==									

Processed Data:

	A	B	C	D	E	F	G
1	id	toxic	severe_to	obscene	threat	insult	identity_hate
2	00001cee3	0.5	0.5	0.5	0.5	0.5	0.5
3	000024786	0.5	0.5	0.5	0.5	0.5	0.5
4	00013b17a	0.5	0.5	0.5	0.5	0.5	0.5
5	00017563c	0.5	0.5	0.5	0.5	0.5	0.5
6	00017695a	0.5	0.5	0.5	0.5	0.5	0.5
7	0001ea871	0.5	0.5	0.5	0.5	0.5	0.5
8	00024115c	0.5	0.5	0.5	0.5	0.5	0.5
9	000247e83	0.5	0.5	0.5	0.5	0.5	0.5
10	00025358c	0.5	0.5	0.5	0.5	0.5	0.5
11	00026d109	0.5	0.5	0.5	0.5	0.5	0.5

Issues With and Benefits of the Dataset

One issue with the dataset is that the labels were generated using a crowdsourcing platform, which means that they may not always be accurate. Additionally, the dataset contains comments with offensive and discriminatory language, which can be difficult to work with for some people. However, the dataset provides a unique opportunity to develop models that can detect and classify toxic language, which can be valuable for online content moderation and related applications. Overall, the dataset is a good resource for training models to identify toxic

language and provides a starting point for researchers to develop more advanced models for text classification.

Sources

Jigsaw/Conversation AI. (2018). *Toxic Comment Classification Challenge - Dataset Description*. Kaggle. https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data