Group Name: Sedats
Name: Sedat CAN (one-member group)
Email: sedatcancan1@gmail.com
Country: Germany Specialization: Data Science, NLP, Data Analyst

https://github.com/sedat-can/Final-project-

Exploratory Data Analysis

<Twitter Hate Speach Detection>

16/12/2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

The hate speech detection model would evaluate and classify whether tweets include hate speech and filter these tweets to offer a better user experience.

The data does not have null value. I can be easily cleaned. emmatizer processes the texts well. labels are dis-proportionational

The project would train a model that evaluates tweets better and gives a more persuasive prediction.

Data Glacier
Your Deep Learning Partner

Problem Statement

Hate speech is one of the biggest problems that social media users face on daily basis. Especially, Twitter users are harassed and discriminated against because of their religion, ethnicity, nationality, race, color, ancestry, sex, etc. The hate speech detection model would evaluate and classifies whether tweets include hate speech or not and filters these tweets to offer a better user experience.

Approach

As the project offers a model that selects tweets containing hate speech, It evaluates data and after the cleaning process, it trains a model that evaluates tweets better and gives a more persuasive prediction.

EDA

The data:

10197*3 values.

The three columns:

id, tweet and label.

‘tweet’:

the user, text and hashtag, special characters ‘(‘ ‘&’,

EDA

Does the data has null value?

```
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      17197 non-null  int64
 1   tweet   17197 non-null  object
 2   label   17197 non-null  int64
dtypes: int64(2), object(1)
memory usage: 403.2+ KB
None
NULLS?
id       0
tweet    0
label    0
dtype: int64
```

EDA

Data Glacier
Your Deep Learning Partner

Clear data

```
df['cleartweet'] = df['tweet'].apply(lambda  tweet: re.sub("(@[A-Za-z0-
9]+)|(#)|(RT[\s]+)|(https?:\/\/\S+)|([^a-zA-Z0-9 -])", "", tweet))
```

Lemmitizer

Tweet before processed

tweet
#studiolife #aislife #requires #passion #dedic...
@user #white #supremacists want everyone to s...
safe ways to heal your #acne!! #altwaystohe...
is the hp and the cursed child book up for res...
3rd #bihday to my amazing, hilarious #nephew...

Processed tweet data

lematizer
studiolife aislife requires passion dedication...
white supremacists want everyone to see the ...
safe ways to heal your acne altwaystoheal h...
is the hp and the cursed child book up for res...
rd bihday to my amazing hilarious nephew el...

EDA Summary

Data Glacier
Your Deep Learning Partner

Executive Summary

The data does not have null value. I can be easily cleaned. emmatizer processes the texts well. labels are dis-proportionational.

Recommendations

Two models:

Random Forest

Artificial Neural Networks

Sedat Can

Data Glacier

Your Deep Learning Partner

Thank You