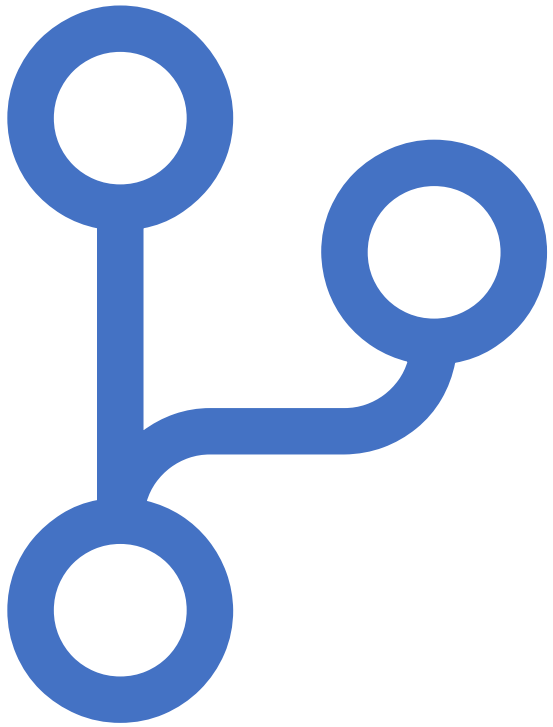
Hate speech detection using transformers (deep learningh

Submitted by : Noura Alsahli
Internship domain : NLP
Batch : LISUM30
Date: 28-April-2024

Content



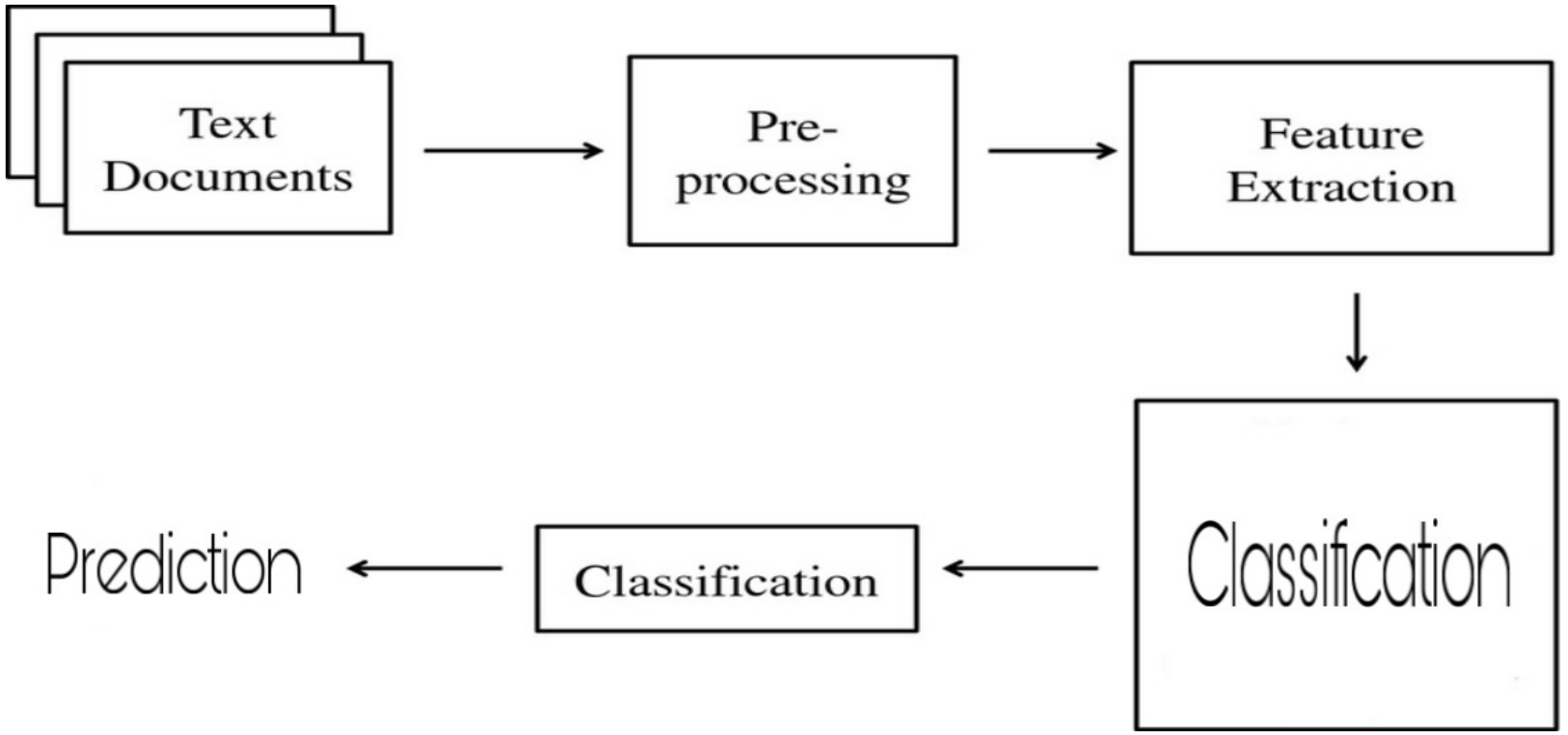
- Problem Statement.
- architecture.
- Data collection.
- Data processing.
- Feature extraction.
- Deep learning model.
- Evaluation.
- Design.
- Conclusion.

Data Glacier

Problem statement :

- The term hates speech is understood as any type of verbal written or behavioural communication that attacks or use derogatory or discriminatory language against a person or a group based on what they are. For example ,based on their religion, ethnicity ,nationality ,race ,colour ,ancestry ,sex or another identity factor in this problem I will take you through a hate speech detection model made with machine learning and python.
- Hate speech detection is a task of sentiment classification so for training a model that can classify hate speech from free speech based on certain piece of text, it can be achieved by training it on data that is used to classify sentiments so for the task of the speech inspection model I will use Twitter to identify tweets containing hate speech.

System architecture:



Data collection:

- The data taken from Kaggle is a Twitter hate speech data that contains three futures and 31962 of observations. The dataset using the Twitter hates speech. Data was used to research hate speech detection. The text is classified as hate speech, offensive language and neither. The data set contains text that can be considered racist sexist, homophobic or offensive.

Tabular data details:

Total number of observations	31962
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	2.95 MB

Data processing:

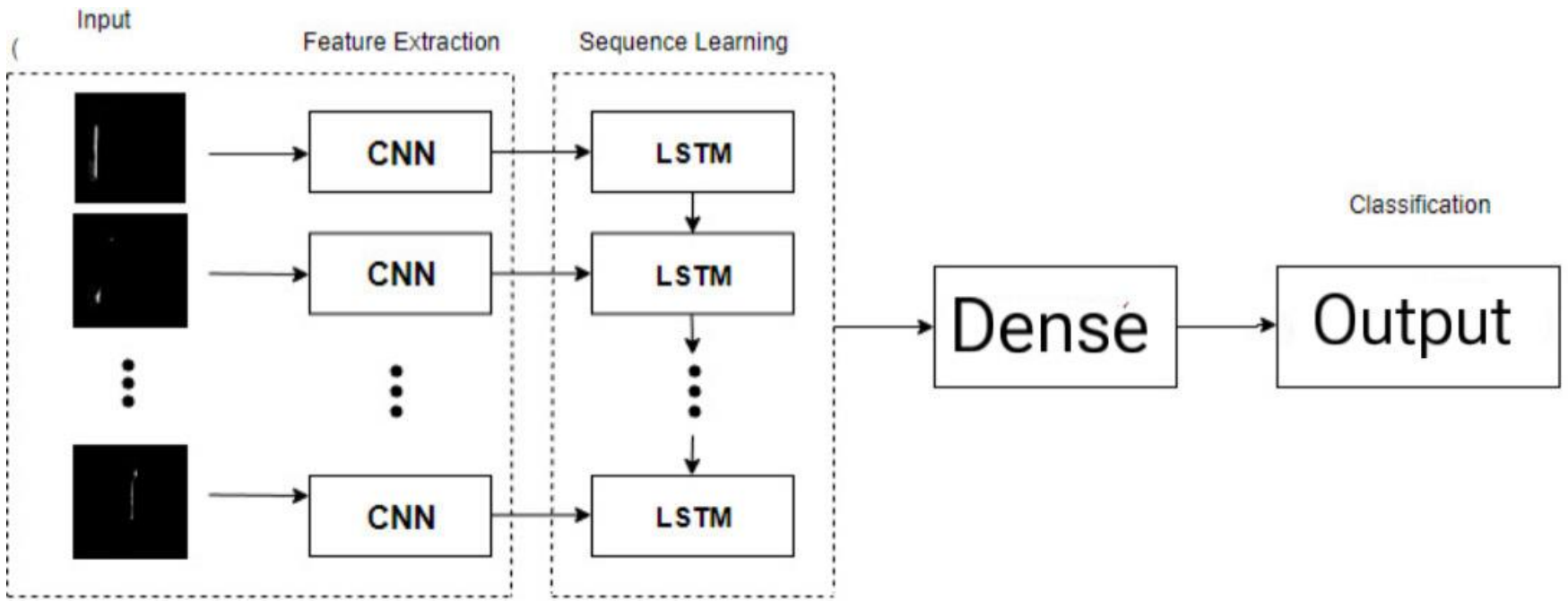
- Text cleaning
  - Lowercase.
  - Remove punctuations.
  - Remove URL.
  - Remove tags.
  - Remove special characters.

- Processing operations
  - Tokenization.
  - Removing stop words.
  - Lemmatization.

Feature extraction:

- TF-IDF Model
  - creating a histogram.
  - Frequent words from dictionaries.
  - TF matrix.
  - IDF matrix.
  - TF – IDF calculation.

Deep learning model:



Result evaluation:

- The confession matrix visualisation contains the following components
  - True negative :there are 5952 instances accounting for 93.10%.
  - False positive: there are 33 instance representing 0.52%.
  - False negative: there are 237 instance making up 3.71%.
  - True Positive: there are 171 instance which correspondence to 2.67% .

- The matrix has two axis labelled as zero and one representing predicted classes a colour bar on the right side indicate the scale of instance counts ranging from 0 to over 5000.

- The different performance of the matrix : The accuracy was close to one, The precision was approximately 0.8, the TPR were around 0.4 ,the FPR slightly above zero ,the F-score close to one and specificity was approximately 0.6.

Application design:

- The application design contains:
  - User interface HTML.
  - Python flask application.
  - Process data TF – IDF vectorizations .
  - Send processed results to application.
  - Result displayed in HTML.
  - Send the results to HTML page.

Conclusion:

- The objective of this project was to identify effective methods and configurations for detecting hate speech and free speech on Twitter. While the model is not error free some misclassifications may still occur. Looking ahead I plan to enhance the work by incorporating some techniques including temporal conversional network And random multi model deep learning. The ongoing effort will contribute more robust and accurate speech detection in online platform.

Thank You