HATE SPAN DETECTION

Aryan Dua (2020CS50475)
Parmar Hirenkumar Hareshbhai (2020CS50435)
Viraj Agashe (2020CS10567)

PROBLEM STATEMENT

This task aims to detect the various hateful spans within a sentence already considered hateful.

MODELS USED

01 - BERT

02 - CRF

BERT

STEP 1

Loading of text data from CSV files.

STEP 2

Breakdown of text into smaller tokens using a tokenizer.

STEP 3

Conversion of data to dataloaders for batchwise efficient training

STEP 4

Transformation of textual data into numerical embeddings for model input.

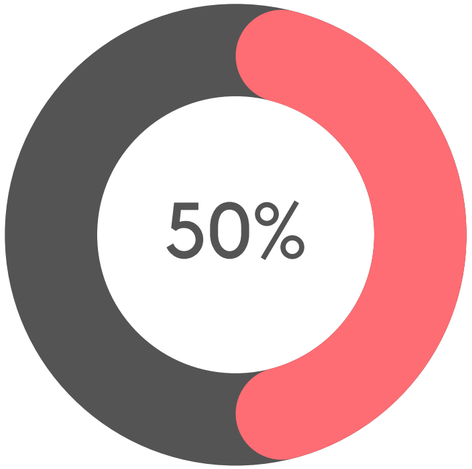
STEP 5

Selection of pre-trained BERT model for hate span detection.

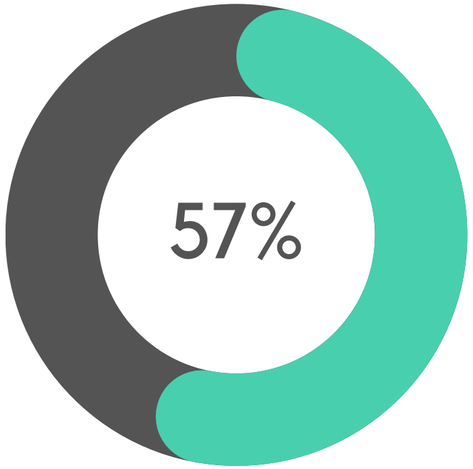
STEP 6

Optimization of model performance over multiple epochs.

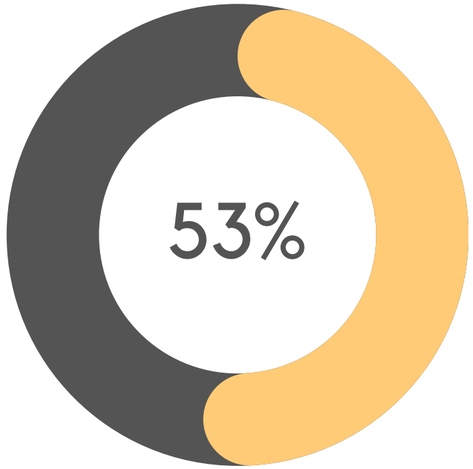
RESULTS



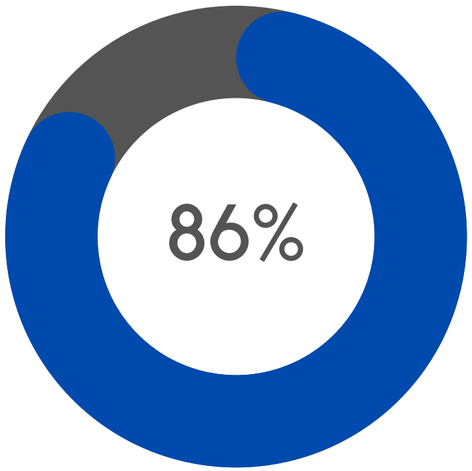
01 - PRECISION



02 - RECALL



03 - F1 SCORE



04 - ACCURACY

We achieved an accuracy of **23.6%** on the test.csv dataset using CRF model.

CONDITIONAL RANDOM FIELDS

Conditional Random Fields (CRFs) are probabilistic models used for sequence labeling tasks, such as named entity recognition and part-of-speech tagging.

CRFs model the conditional probability of a sequence of labels given an input sequence, taking into account the dependencies between neighboring labels.

CRFs can effectively capture the sequential nature of language in hate speech detection tasks, making them suitable for identifying patterns and contexts indicative of hate speech.

01 - FEATURE EXTRACTION

The word2features function extracts features from each word in the sentence, such as its lowercase form, suffixes, capitalization, and numerical properties. These features capture relevant information for hate speech detection.

02 - SEQUENCE REPRESENTATION

The sentence2features function processes the entire sentence to convert it into a sequence of feature vectors, which serve as input (X_train) for the CRF model.

03 - LABEL PREPARATION

The sentence2labels function extracts the corresponding labels for each sentence, creating the target labels (Y_train) for the CRF model.

CONDITIONAL RANDOM FIELDS

During training, the CRF model learns the dependencies between input features and output labels, leveraging the sequential information encoded in the data.

Standard metrics such as precision, recall, and F1-score are computed to assess the model's accuracy and effectiveness in hate speech detection.

The classification report provides insights into the model's performance across different hate speech categories and its ability to generalize to unseen data.

RESULTS

	precision	recall	f1-score	support
B	1.00	0.89	0.94	485
I	0.78	0.67	0.72	246
O	0.99	1.00	0.99	479
micro avg	0.96	0.89	0.92	1210
macro avg	0.92	0.85	0.88	1210
weighted avg	0.95	0.89	0.92	1210
samples avg	0.96	0.89	0.91	1210

We achieved an accuracy of **31.8%** on the test.csv dataset using CRF model.