HATE SPEECH DETECTION ON TURKISH TWEETS

Nil Akay

150210737

twitter

People have a free platform to openly express their emotions thanks to social media. Twitter is one of the most popular of them.

Go to the page

Whats happening?

The enormous size of the data influence of Turkish social media contains a lot of harmful content that causes major content filtering issues, such as hate speech, cyberbullying, and insulting material.
The rise of these online insults directed towards other countries, ethnicities, faiths, and other groups has an effect and disturbs social peace. Due to its irony and sarcasm, Turkish hate speech is hard to detect.

Nil

Tweets	Following	Followers
38	38	38

Trends For You

#DataMining

#HateSpeech

#IstanbulTechnicalUniversity

#AI

There are some hate speech to detect!

Take a look

Who to follow

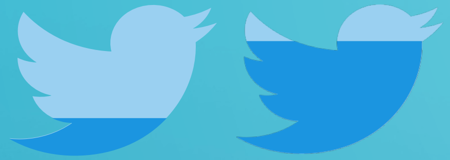
Refresh

Deniz

Alin

Issues About Related Work

Imbalanced



- Generalizability problems
- Tend to favor the majority class for accuracy

Biased

- Towards some entities and religions
- Simple mention of the entities in question, model can label that instance offensive.

Mislabeled

- Can train the model wrongly
- Subjective

Datasets

“The OffensEval-20"

Constructed and often used for offensive language categorization,

Manually classified

“HATC”

Homophobic-Abusive Turkish Comments

Was collected from Instagram

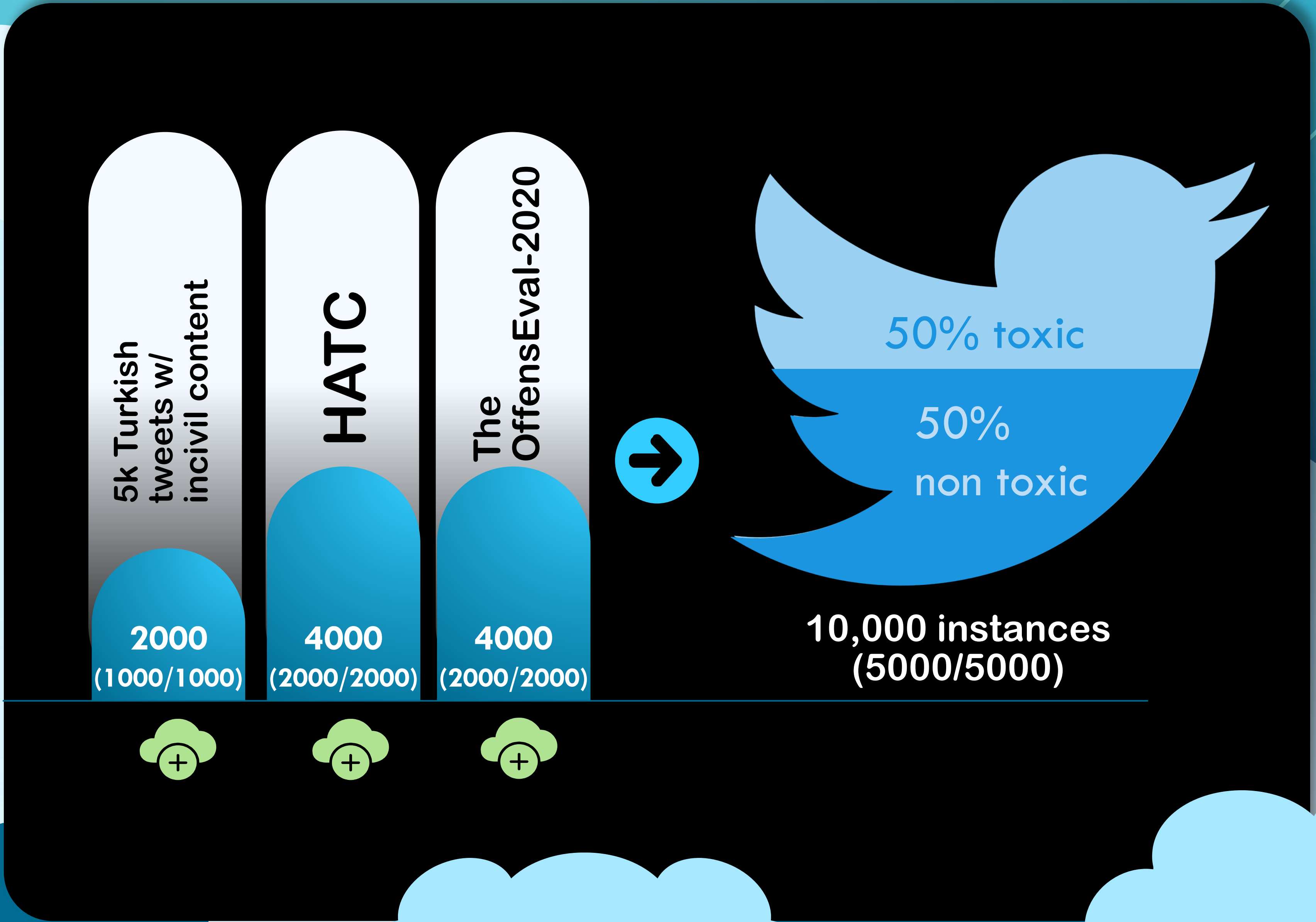
List of 201 words that would cause high morphological uncertainty was created and removed from this dataset

“5k Turkish tweets with incivil content”

A collection of Turkish tweets from twitter.

2,073 of the 5,054 total samples are offensive

Dataset


5k Turkish tweets w/ incivil content
2000
(1000/1000)
HATC
4000
(2000/2000)
The OffensEval-2020
4000
(2000/2000)
50% toxic
50% non toxic
10,000 instances
(5000/5000)

The Model Refresh

Combination of

BERT – transformed based ML technique

CNN-BiLSTM – a deep learning pipeline

outperforms most of the other models

The Model includes:

#BERTurk

A Turkish BERT model with 128k uncased vocabulary

Extracts Turkish language features

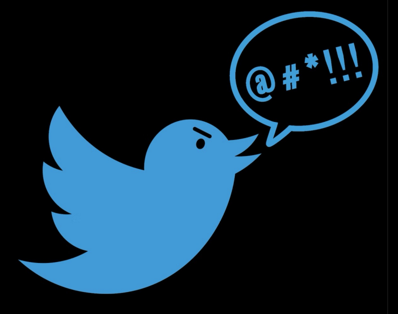
#1D CNN

Since they can extract as many features from the text as possible

#BiLSTM

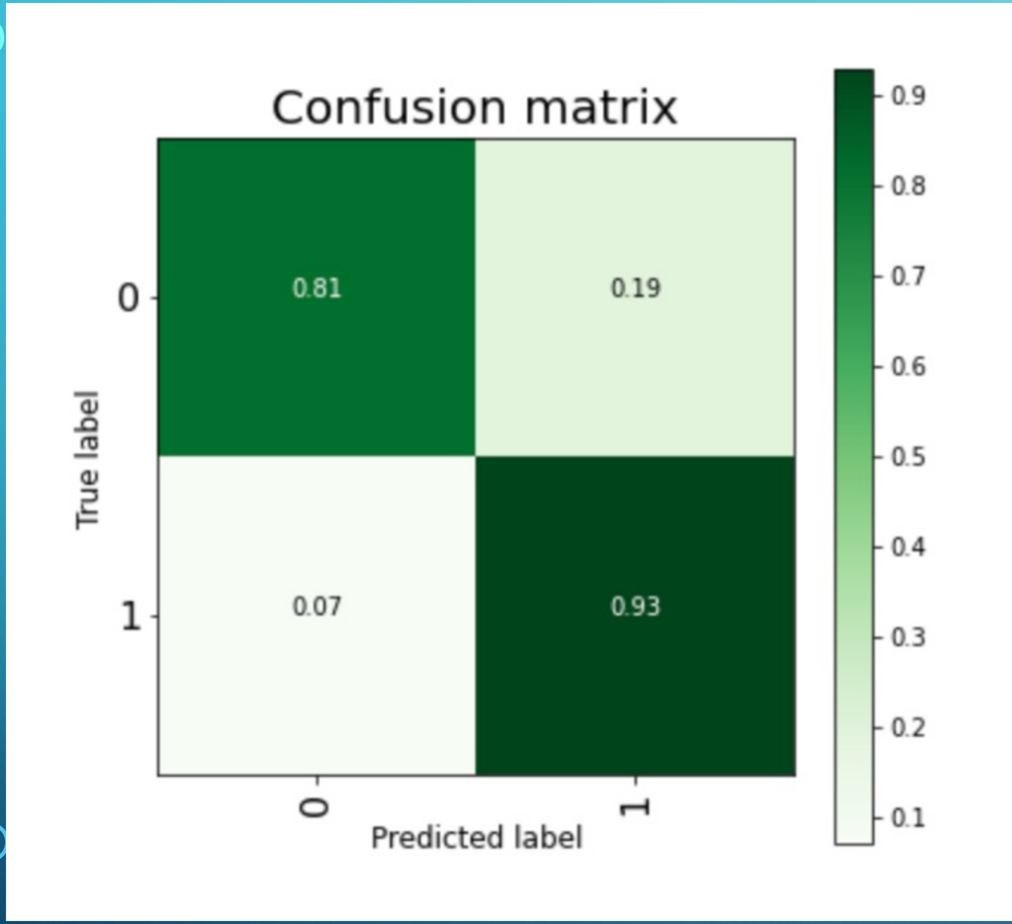
It efficiently expands the network's information pool, enhancing the context that the algorithm has access to.

Uses the extracted features to learn bidirectional long-term dependencies between words



BERT-CNN-BiLSTM

Results



Precision: 0.8387
Recall: 0.9293
Accuracy: 0.8730
F1 Score: 0.8816

	precision	recall	f1-score	support
0	0.92	0.81	0.86	491
1	0.84	0.93	0.88	509
accuracy			0.87	1000
macro avg	0.88	0.87	0.87	1000
weighted avg	0.88	0.87	0.87	1000

Conclusion & Future Work

Different instances were gathered from different datasets to have more various samples in the set. The samples were arranged to be the same number per class to get rid of the imbalanced data problem.

The dataset’s size can be increased by adding more positive instances about specific races, religions, entities, etc. In this way, the bias toward those entities can be reduced, and the generalizability can be improved. Also, we can reduce the harm that mislabeled data can give to the model by customizing the loss function.