Name: Faris Chaudhry
Batch: LISUM25

Hate Speech Detection

Week 8

Team Member Details

Name: Faris Chaudhry

Email: faris.chaudhry@outlook.com

Country: United Kingdom

University: Imperial College London

Specialization: NLP

Problem Description

“The term hate speech is understood as any type of verbal, written or behavioural communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, colour, ancestry, sex or another identity factor.

Hate Speech Detection is generally a task of sentiment classification. So, for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. We will use the Twitter tweets to identify tweets containing Hate speech.”

Name: Faris Chaudhry
Batch: LISUM25

Fields and Datatypes

Training Data

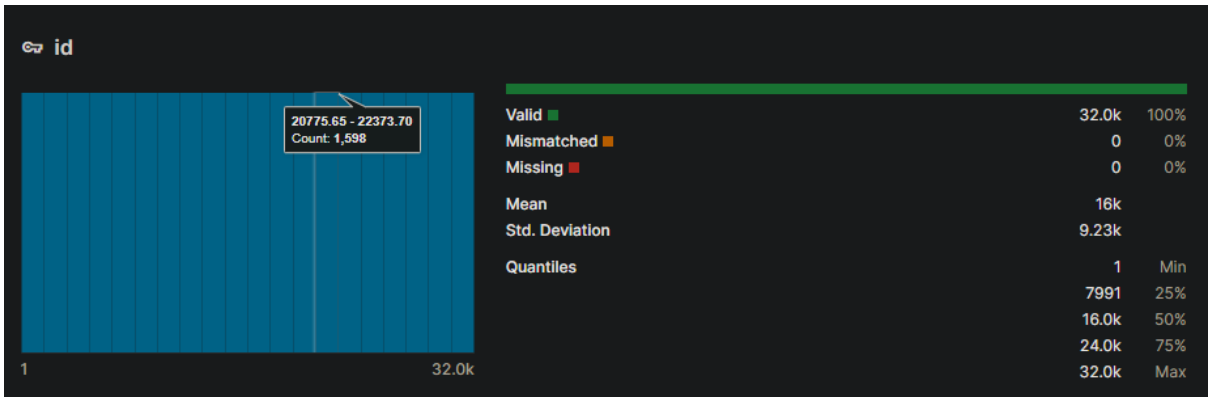
Field	Datatype	Other Notes
id	Id/Integer	Primary key; ascending.
label	Integer	Enumerable: 0 (not hate speech) or 1 (hates speech)
tweet	String	Character limit of 200. Contains special characters and alphanumeric characters.

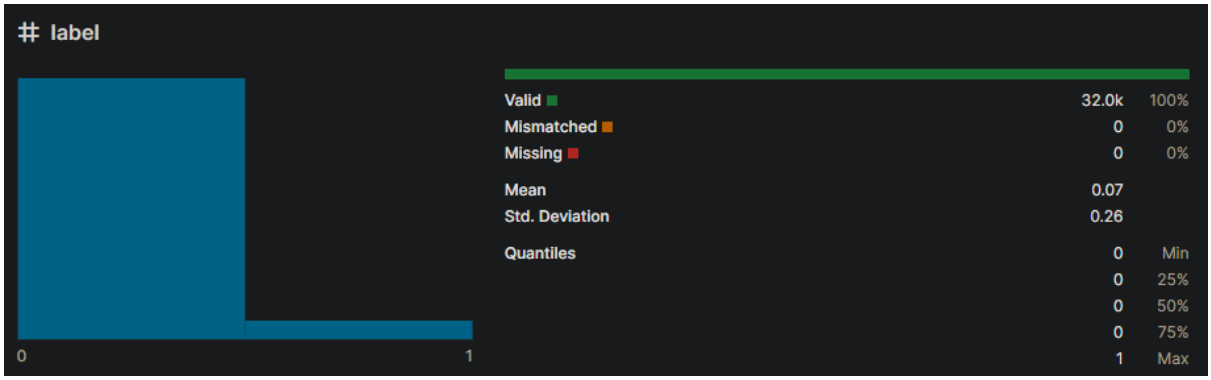
Testing Data

Field	Datatype	Other Notes
id	Id/Integer	Primary key; ascending.
tweet	String	Character limit of 200. Contains special characters and alphanumeric characters.

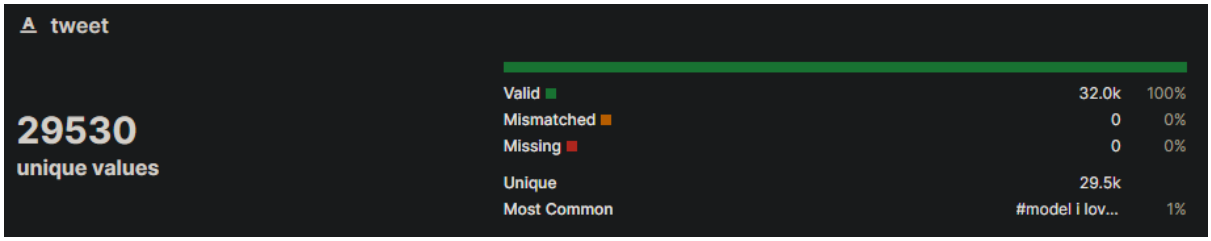
Data Distribution and Quality Analysis

Training Data

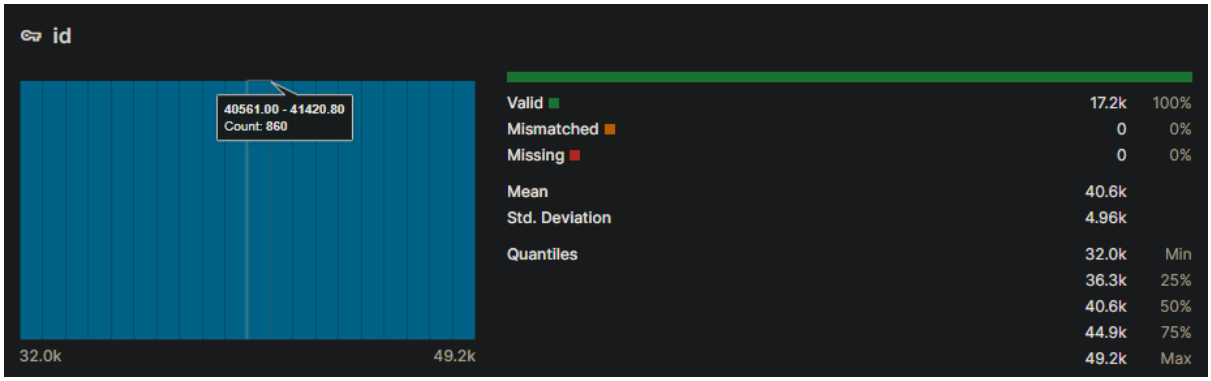


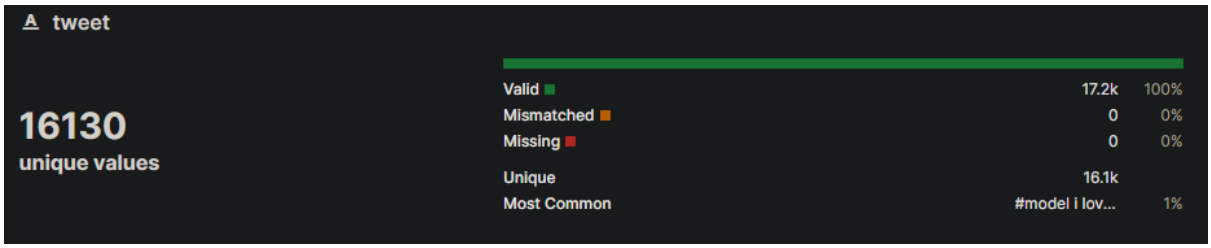


Name: Faris Chaudhry
Batch: LISUM25



Testing Data





From the above it can be seen that there are no missing, invalid, or mismatched values in either data set. Furthermore, it is not possible to tell if there are duplicate values from the visualisations above, so this will have to be done. NA value checks will be done on all columns despite the above graph. Where there is a null value in the tweet column, the data will be deleted; where there is a null value in the id column, a new, unique id will be assigned; where there is a null value in the label column of the training data set, either we can classify it ourselves or delete the data as it is not useful. The approach will depend on how many such examples we find. In the case of duplicate ids, all but one will be changed to a new, unique id. In the case of duplicate tweets, all but one will be deleted. In the case of the same tweet being classified in two different ways (i.e., as hate speech and not hate speech) either both values will be removed, or one will be removed depending on the discretion of the reviewer.

There is no concern for outliers and skewed data since there is no numerical factors explicitly. However, feature engineering will be done to create features such as the length of the tweet, how many blacklisted words are used, the number of capital letters used (this can signify rage), and there may be outliers within these new numerical features. The label data in the training set is quite biased towards non-hate speech at about a 94:6 split; we assume that this is proportional the amount of hate speech on the platform and must account for this when training the model (otherwise it would classify a tweet as ‘not hate speech’ because it is more likely). Therefore, we must choose between minimising type 1 and type 2 errors when designing the model and cleaning the data.

Name: Faris Chaudhry
Batch: LISUM25

Other problems include the need for data cleaning after feature extraction, such as removing special characters, making all tweets lowercase so that word checks are not case-sensitive, removing commonly used words which don't affect the sentiment analysis. This will be done with regex.