Artificial Intelligence

Detection Of Toxic Comments

Murat Tinal
23MD0442

Introduction

In recent years, with the development of social networks and online platforms, the problem of spreading toxic comments has been growing. Toxic comments can include insults, threats, racism, sexism, and other types of negative behavior, which negatively affects the atmosphere of communication on the Internet. To automate the process of content moderation and prevent the spread of harmful messages, it is becoming increasingly important to use machine learning methods to classify toxic comments.

The classification of texts such as comments is one of the key tasks in the field of natural language processing . It involves analyzing texts to determine their category or label (for example, toxic or non-toxic). To solve this problem, various machine learning algorithms are used, which can be both simple and quite complex, depending on the required accuracy and complexity of the data.

The purpose of the study

The purpose of the study is to develop and compare various ML methods for classifying toxic comments. The main task is to create a model that can effectively distinguish between toxic and non-toxic comments, which can be useful for filtering unwanted messages on social networks and other online platforms.

Three models were selected for this purpose:

- Logistic Regression;
- Random Forest;
- Voting Classifier ;

I evaluated their performance against several metrics and determined which model performs better.

Models

Logistic Regression:

Logistic regression is a linear classifier that is used to predict the probability that a comment will be toxic. The model evaluates the impact of each word in a comment on the likelihood that it is a toxic statement. For example, the comment "You're stupid!" It can be classified as toxic because the word "stupid" has a high probability of being associated with an insult. Logistic regression will look at the frequency of occurrence of such words in the data and, based on these characteristics, decide whether a comment is toxic.

Random Forest:

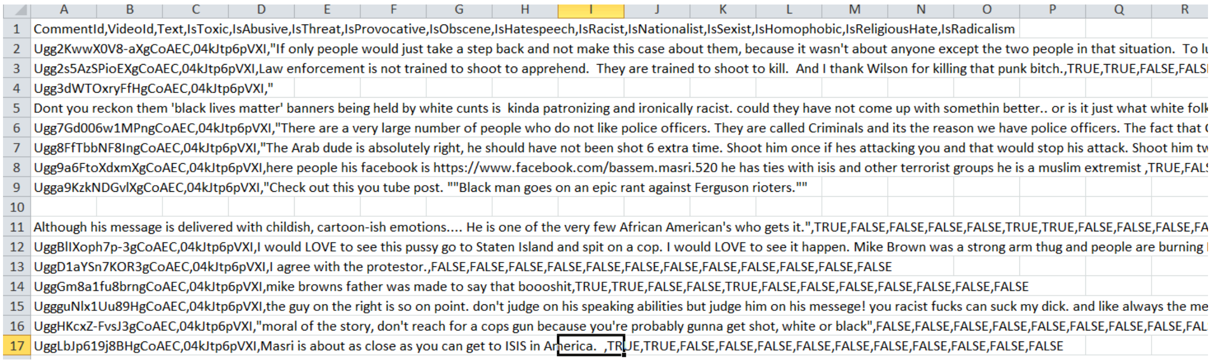
A random forest is an ensemble method consisting of many decision trees. Each tree makes a decision based on a subset of features (for example, individual words or phrases in the text), and the result of all trees is averaged to make a final decision. This model can effectively recognize more complex dependencies between words and context, which is important when analyzing toxic comments. For example, the comment "I hope you die" can be classified as toxic because the words "hope" and "die" in combination often indicate a threat.

Voting Classifier:

The Voting Classifier combines several basic classifiers, such as Logistic Regression and Random Forest, and makes a decision based on their joint findings. The "soft voting" approach is used, in which models give probabilities for each class (toxic or non-toxic), and the final decision is made based on averaging these probabilities. This method improves overall accuracy by taking into account the strengths of several models. For example, if one model cannot accurately classify a comment as toxic, another model can help correct the error.

Database

About the database, I took it from Kaggle in ***Fig.1***.The source dataset consists of comment texts with several types of toxicity indicating their toxicity.



```
   A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R
1  CommentId,VideoId,Text,IsToxic,IsAbusive,IsThreat,IsProvocative,IsObscene,IsHatespeech,IsRacist,IsNationalist,IsSexist,IsHomophobic,IsReligiousHate,IsRadicalism
2  Ugg2KwwX0V8-aXgCoAEC,04kJtp6pVXI,"If only people would just take a step back and not make this case about them, because it wasn't about anyone except the two people in that situation.  To l
3  Ugg2s5AzSPioEXgCoAEC,04kJtp6pVXI,Law enforcement is not trained to shoot to apprehend.  They are trained to shoot to kill.  And I thank Wilson for killing that punk bitch.,TRUE,TRUE,FALSE,FALS
4  Ugg3dWTOxryFfHgCoAEC,04kJtp6pVXI,"
5  Dont you reckon them 'black lives matter' banners being held by white cunts is  kinda patronizing and ironically racist. could they have not come up with somethin better.. or is it just what white folk
6  Ugg7Gd006w1MPngCoAEC,04kJtp6pVXI,"There are a very large number of people who do not like police officers. They are called Criminals and its the reason we have police officers. The fact that
7  Ugg8FfTbbNF8IngCoAEC,04kJtp6pVXI,"The Arab dude is absolutely right, he should have not been shot 6 extra time. Shoot him once if hes attacking you and that would stop his attack. Shoot him tw
8  Ugg9a6FtoXdxmXgCoAEC,04kJtp6pVXI,here people his facebook is https://www.facebook.com/bassem.masri.520 he has ties with isis and other terrorist groups he is a muslim extremist ,TRUE,FAL
9  Ugga9KzkNDGvlXgCoAEC,04kJtp6pVXI,"Check out this you tube post. ""Black man goes on an epic rant against Ferguson rioters.""
10
11 Although his message is delivered with childish, cartoon-ish emotions.... He is one of the very few African American's who gets it.",TRUE,FALSE,FALSE,FALSE,FALSE,TRUE,TRUE,FALSE,FALSE,FALSE,FA
12 UggBlIXoph7p-3gCoAEC,04kJtp6pVXI,I would LOVE to see this pussy go to Staten Island and spit on a cop. I would LOVE to see it happen. Mike Brown was a strong arm thug and people are burning
13 UggD1aYSn7KOR3gCoAEC,04kJtp6pVXI,I agree with the protestor.,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE
14 UggGm8a1fu8brngCoAEC,04kJtp6pVXI,mike browns father was made to say that boooshit,TRUE,TRUE,FALSE,FALSE,TRUE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE
15 UggguNlx1Uu89HgCoAEC,04kJtp6pVXI,the guy on the right is so on point. don't judge on his speaking abilities but judge him on his messege! you racist fucks can suck my dick. and like always the me
16 UggHKcxZ-FvsJ3gCoAEC,04kJtp6pVXI,"moral of the story, don't reach for a cops gun because you're probably gunna get shot, white or black",FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FAL
17 UggLbJp619j8BHgCoAEC,04kJtp6pVXI,Masri is about as close as you can get to ISIS in America.  ,TRUE,TRUE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE
```

Fig.1

Methodology

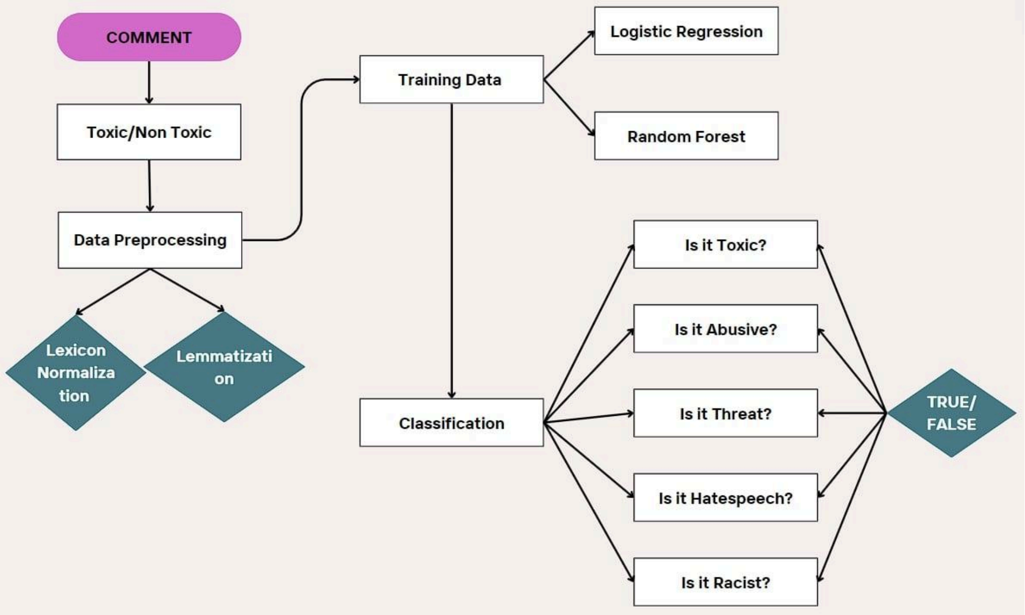
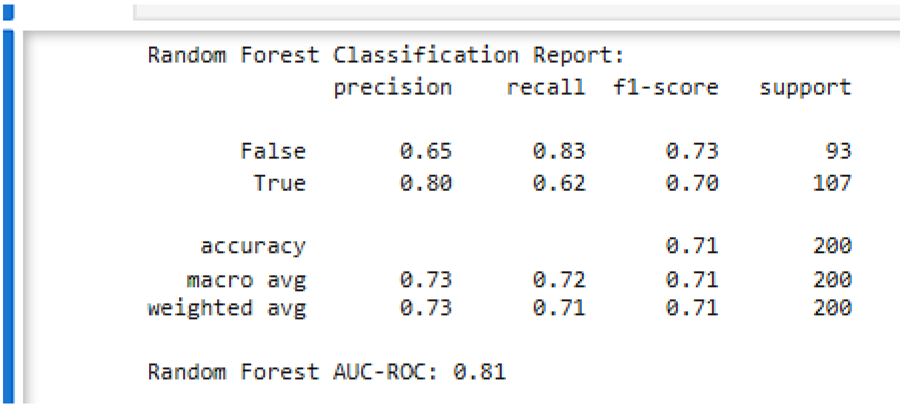


Fig.2

-Data Separation: All the data I used was divided into two parts. 80% of the data was used to train the model (this is called a training sample), and the remaining 20% was used to check how well the model works;
-TF-IDF vectorization: This allows the model to "understand" the text. TF-IDF helps figure out which words are important and which ones aren't so important;
-Model Training: I trained two types of models: Logistic Regression and Random Forest. Both models learned from the training data to figure out how to make predictions. Logistic Regression helps to classify data, and Random Forest uses several decision trees to make better predictions;
-Combining models in Voting Classifier: To make our predictions more accurate, I combined the two models into a Voting Classifier. This means that the two models "vote" on the result. If one model says "yes" and the other says "no", the Voting Classifier picks the answer that gets the most votes;

Evaluation metrics

The Random Forest model gave good results with 71% accuracy in ***Fig.3,*** meaning it correctly predicted 71% of the time. It was especially good at identifying non-toxic texts, with 83% recall for the negative class. However, it wasn't as good at spotting toxic texts, with a 62% recall for the positive class. The model's AUC-ROC score of 0.81 shows it does a solid job at separating the two classes and generally performs well.

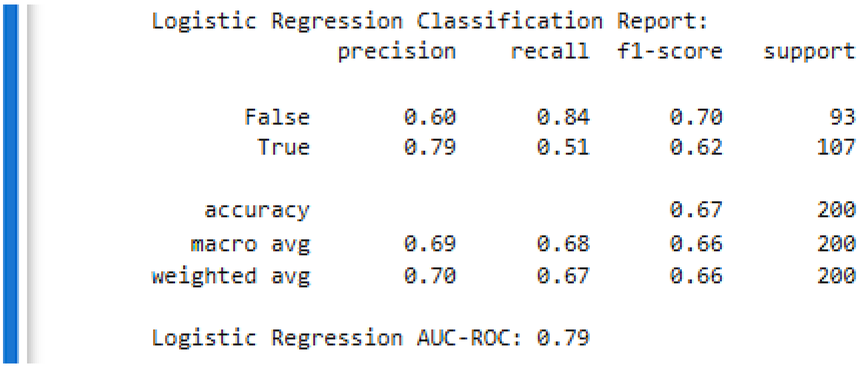


```
Random Forest Classification Report:
              precision    recall  f1-score   support

       False       0.65      0.83      0.73        93
        True       0.80      0.62      0.70       107

    accuracy                           0.71       200
   macro avg       0.73      0.72      0.71       200
weighted avg       0.73      0.71      0.71       200

Random Forest AUC-ROC: 0.81
```

Fig.3

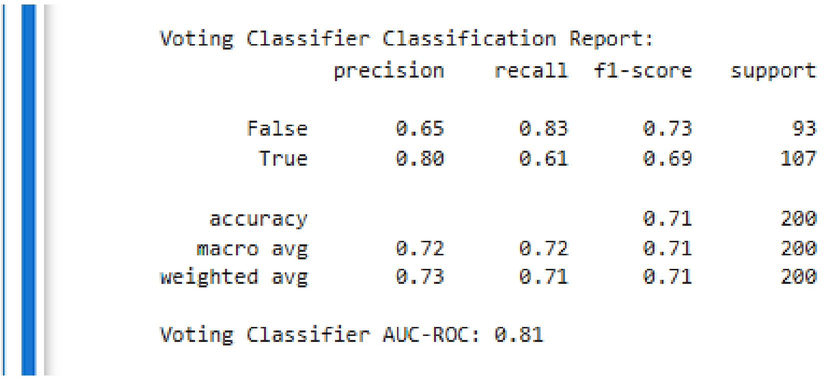
Logistic Regression, on the other hand, had a lower accuracy of 67%, which means it made more mistakes in ***Fig.4***. While it performed well at detecting non-toxic texts (84% recall for the negative class), it struggled with identifying toxic texts, only getting 51% recall for the positive class. Its AUC-ROC score was 0.79, slightly lower than Random Forest, indicating it wasn't as good at distinguishing between the two classes.

The Voting Classifier in ***Fig.5,*** which combines multiple models, showed similar results to Random Forest in Fig.5. It also had 71% accuracy, and its recall for the negative class (83%) and positive class (61%) were very close to the Random Forest model's results. The AUC-ROC score was also 0.81, confirming that combining models in the Voting Classifier helps improve accuracy while maintaining good separation between the classes.



```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

       False       0.60      0.84      0.70        93
        True       0.79      0.51      0.62       107

    accuracy                           0.67       200
   macro avg       0.69      0.68      0.66       200
weighted avg       0.70      0.67      0.66       200

Logistic Regression AUC-ROC: 0.79
```

Fig.4



```
Voting Classifier Classification Report:
              precision    recall  f1-score   support

       False       0.65      0.83      0.73        93
        True       0.80      0.61      0.69       107

    accuracy                           0.71       200
   macro avg       0.72      0.72      0.71       200
weighted avg       0.73      0.71      0.71       200

Voting Classifier AUC-ROC: 0.81
```

Fig.5

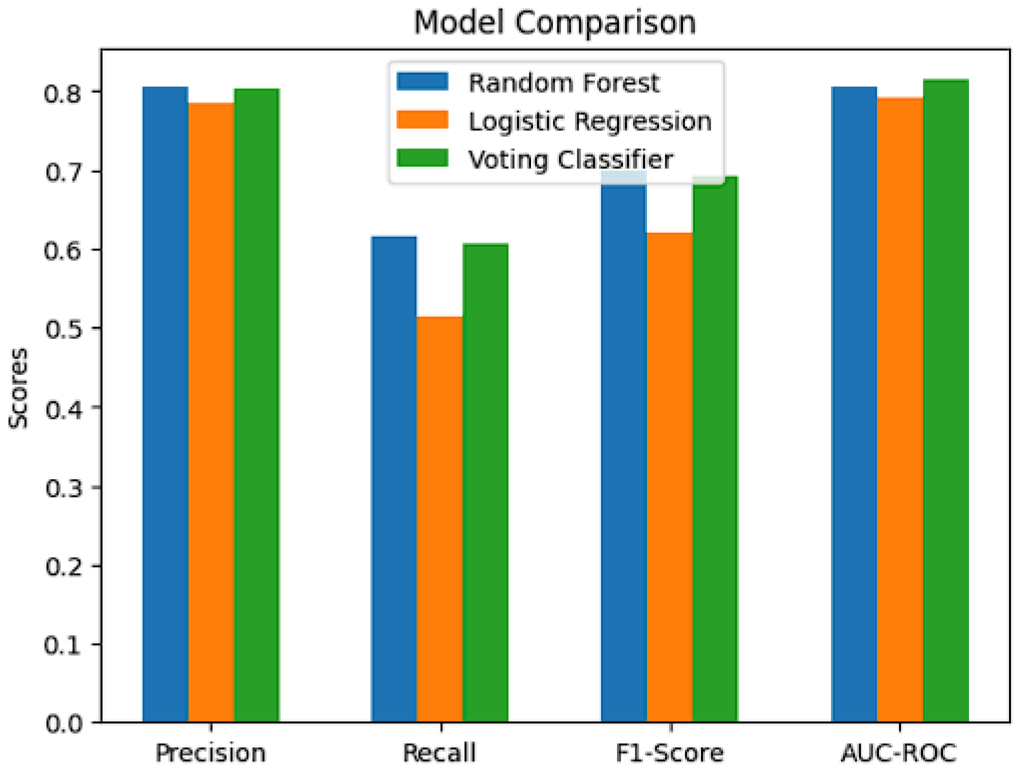


Fig.6

Fig.6 measured using four metrics: Precision, Recall, F1-Score, and AUC-ROC. All models have similar high precision and AUC-ROC scores. However, Random Forest has a lower recall compared to the others. The Voting Classifier performs slightly better in terms of the F1-Score and maintains a good balance across all metrics.

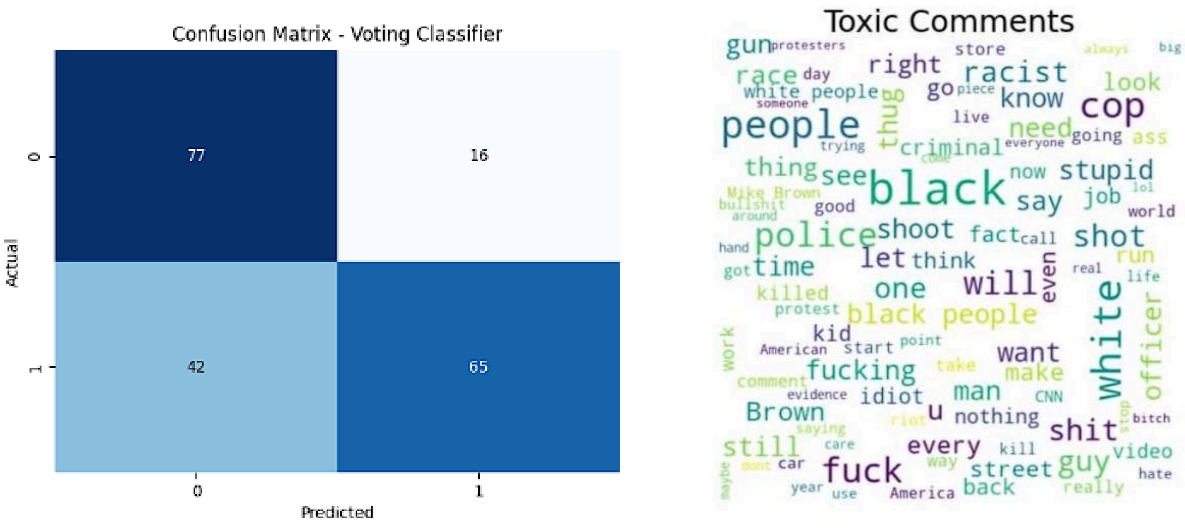


Fig.7

Fig.7 shows how well the Voting Classifier works. The model correctly predicted 77 cases as 0, meaning it got these right. It made 16 mistakes by predicting 1 instead of 0. The model also correctly predicted 65 cases as 1. However, it made 42 mistakes by predicting 0 instead of 1. This table helps us understand where the model is accurate and where it makes errors.

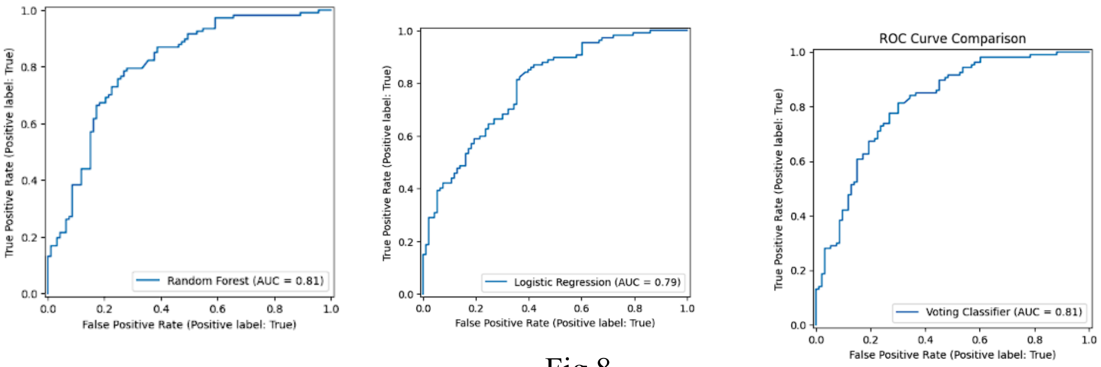


Fig.8

The Random Forest model and Voting Classifier have the same AUC value of 0.81, which indicates better performance compared to Logistic Regression, which has an AUC of 0.79 shown in ***Fig.8.***

The higher the AUC, the better the model copes with the classification. Random Forest and Voting Classifier show equally high accuracy, which indicates their ability to distinguish well between positive and negative classes. Logistic Regression is slightly inferior to them but still shows stable results. Such comparisons are important for choosing the model that will be most effective for a particular task.

Conclusion

In conclusion, I investigated the task of classifying texts into toxic and non-toxic using machine learning methods. The best accuracy (71%) was achieved by the Random Forest and Voting Classifier models, which is also confirmed by the high AUC-ROC value (0.81). These models showed a good balance between recognizing toxic and non-toxic texts.