Cyberbullism Detection System

Course Final Project aa:2024/2025

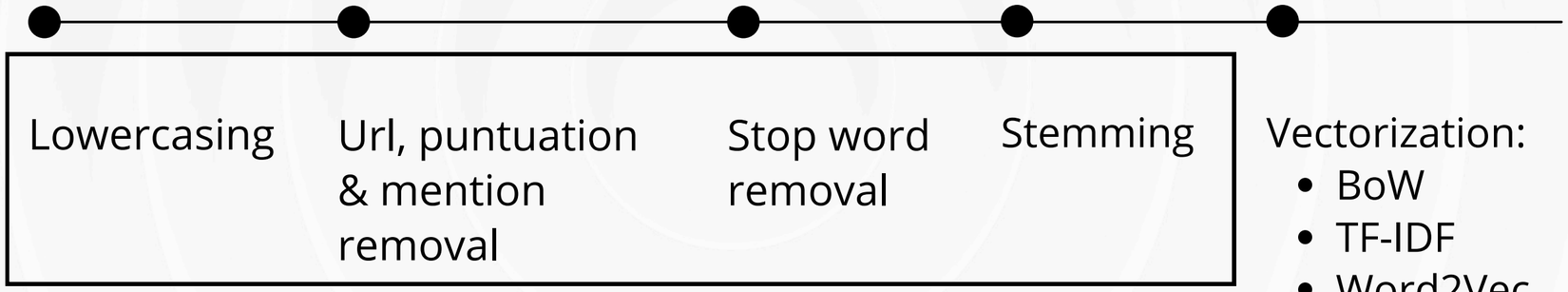
Segreto Mattia

Index

Dataset

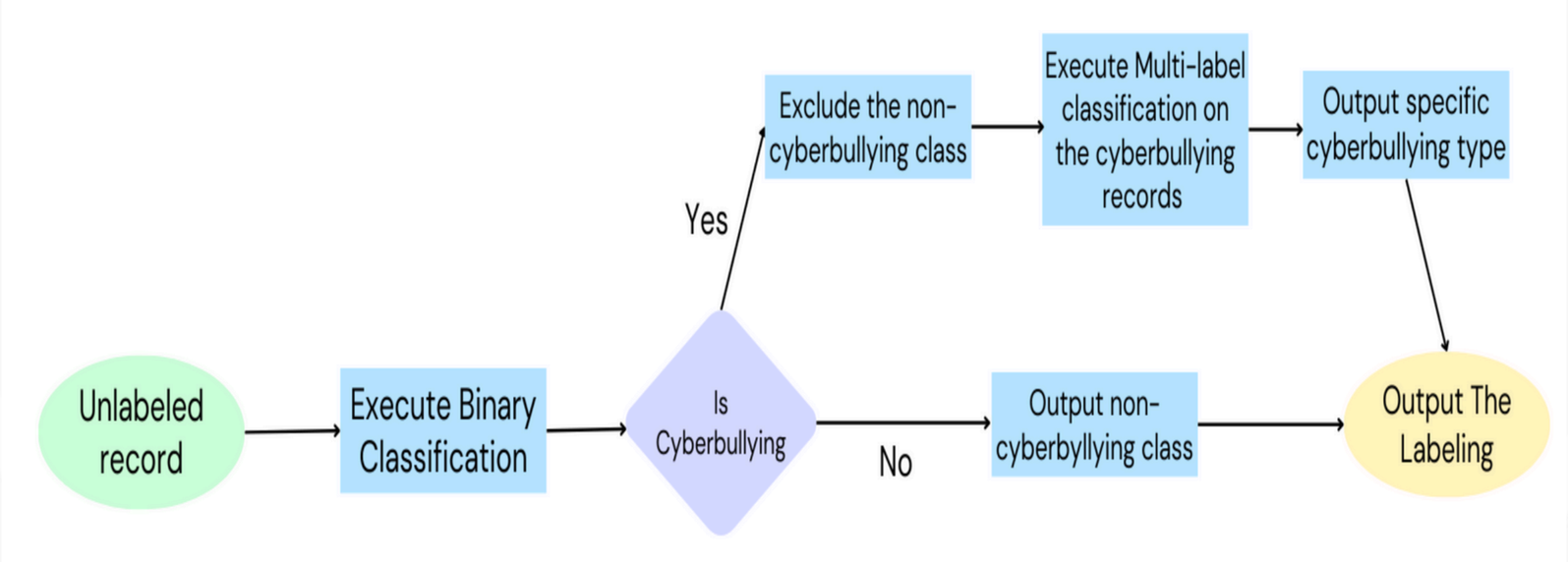
- Labaled Dataset
- Almost **50000** tweets
- Balanced Among Multiclass Label
- Original Feature:
  - **tweet_text**
  - **cyberbullying_type:** it’s the label used for **multiclass classification step**
- Created Feature:
  - **cyberbullying**: it’s a binary label created for **binary classification step**

Data Preprocessing & Vectorization



Preprocessing Phase

Classification Pipeline



Binary Classification

- **LogisticRegression** is the **best** model in this classification step
- The **vectorization** methods that overperform is **TF-IDF.**

Model Evaluation & Model Comparison Tables

Vectorizer	Classifier	Mean f1_macro	Mean Acc.	Mean Prec.	Mean Rec.	Mean F1
TF-IDF	LogisticRegression	0.720981	0.794931	0.961023	0.785829	0.864601
BoW	LogisticRegression	0.719513	0.789034	0.968343	0.772117	0.859136
TF-IDF	LinearSVM	0.718948	0.791078	0.963769	0.778595	0.861309
TF-IDF	RandomForest	0.717019	0.794695	0.956629	0.790359	0.864864
BoW	RandomForest	0.716926	0.777737	0.984917	0.744724	0.848107
BoW	LinearSVM	0.716482	0.783792			
W2V-1	RandomForest	0.695382	0.790842			
W2V-1	LogisticRegression	0.684053	0.768564			
W2V-1	LinearSVM	0.683352	0.765916			

Model	Score
LogisticRegression + TF-IDF	7
LogisticRegression + BoW	6
LinearSVM + TF-IDF	4
RandomForest + TF-IDF	0
RandomForest + BoW	1
LinearSVM + BoW	0
RandomForest + W2V-1	-4
LogisticRegression + W2V-1	-7
LinearSVM + W2V-1	-7

Multiclass Classification

- All classifiers achieve good results on the evaluated metrics
- **RandomForest + TF-IDF** over perform w.r.t all the other classifiers

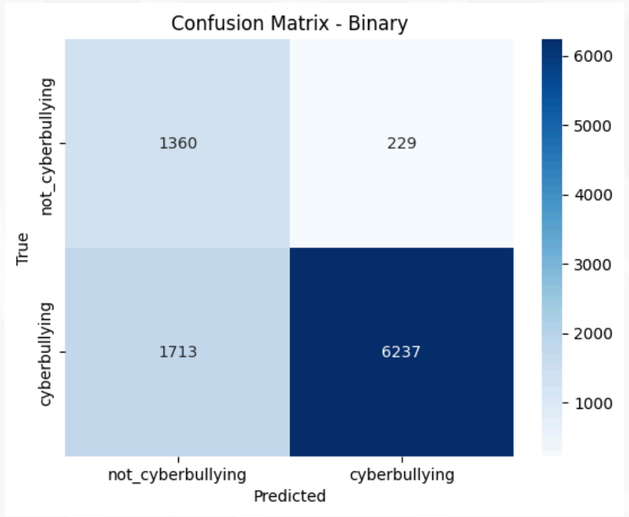
Model Evaluation & Model Comparison Tables

Vectorizer	Classifier	Mean F1	Mean Acc.	Mean Prec.	Mean Rec.
TF-IDF	RandomForest	0.932600	0.932195	0.934327	0.932014
BoW	LogisticRegression	0.928662	0.928075	0.931726	0.927980
BoW	RandomForest	0.928478	0.928075	0.929942	0.927852
BoW	LinearSVM	0.928478	0.927855	0.932123	0.927780
TF-IDF	LogisticRegression	0.927957	0.927383	0.930868	0.927272
TF-IDF	LinearSVM	0.927945			
W2V-1	RandomForest	0.891557			
W2V-1	LinearSVM	0.878282			
W2V-1	LogisticRegression	0.878051			

Model	Score
RandomForest + TF-IDF	8
LogisticRegression + BoW	4
RandomForest + BoW	2
LinearSVM + BoW	2
LinearSVM + TF-IDF	1
LogisticRegression + TF-IDF	1
RandomForest + W2V-1	-4
LinearSVM + W2V-1	-7
LogisticRegression + W2V-1	-7

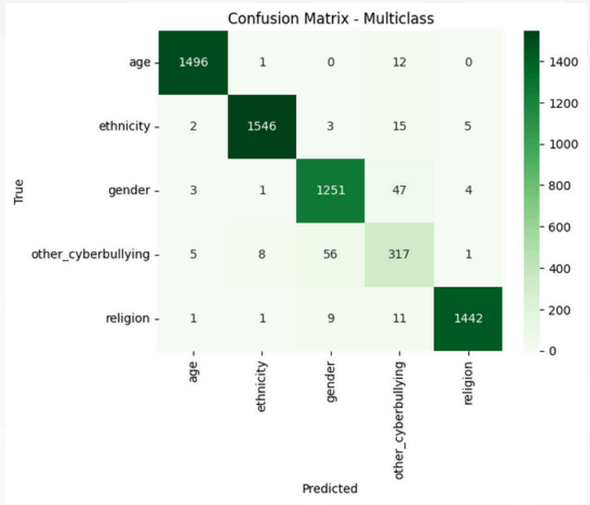
Pipeline Evaluation

Binary Stage:



LogisticRegression+BoW

Multiclass Stage:



Randomforest + TF-IDF

Explanation Module Overview

- Global Explanation
  - Global Classifier Explanation
  - Pattern Mining
- Local Explanation

Pattern Mining

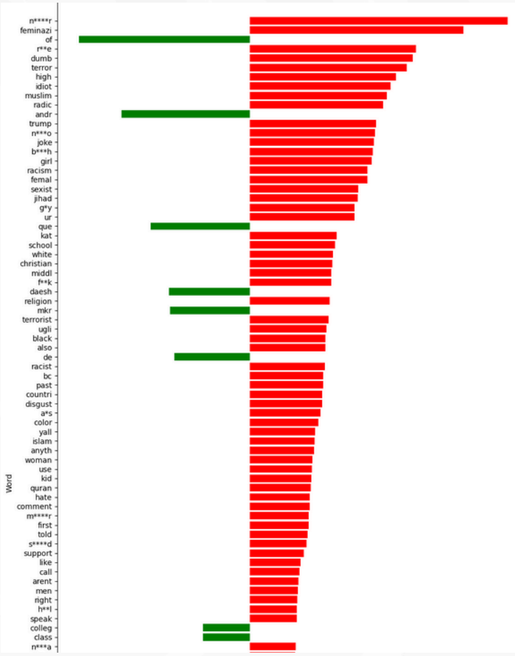
Closed and maximal itemset were analyzed in order to discover if the cyberbullyng phenomenon has **frequent pattern** that shold be useful to discover **interesting informations**

Class	Number of closed itemsets	Number of maximal itemsets
Multiclass Target		
age	147	25
ethnicity	131	25
religion	63	30
gender	62	20
other_cyberbullying	8	8
Binary Target		
cyberbullying	51	23
not_cyberbullying	5	5

Binary Model Global Explanation

LogisticRegression + BoW:

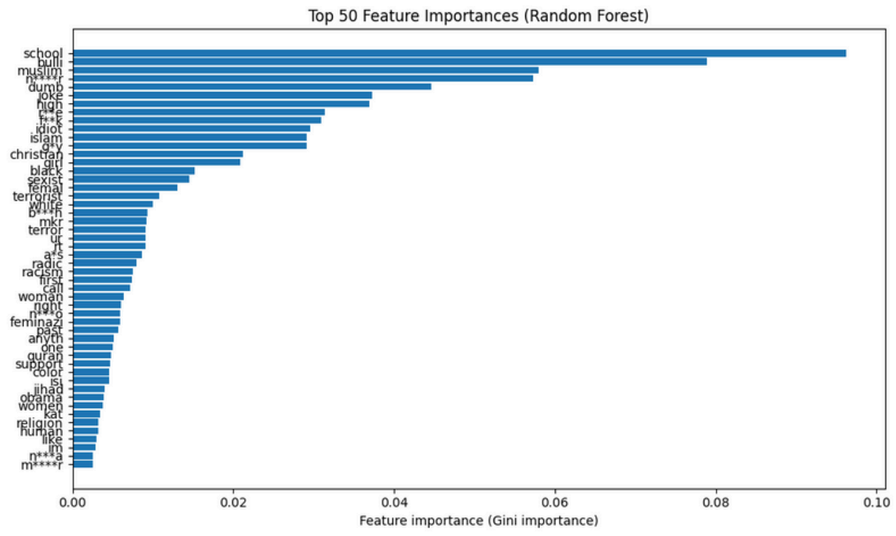
- Coefficient of Logistic Regression has been used(**.coef_**)
- The overall system is recognizing the **cyberbullying** class isted of distinguish the two classes.
- Module can mislead, global trend no.



Multiclass Model Global Explanation

RandomForest + TF-IDF:

- built in method has been used (**.feature_importances_**)
- most frequent terms can be inuitively associated to **harmful content** and specific categories

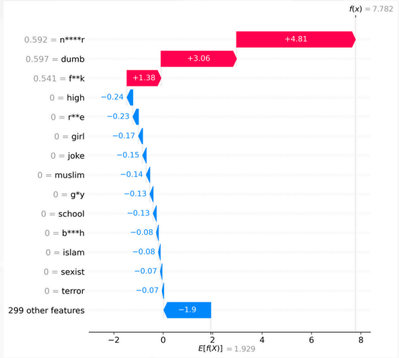


Binary Model Local Explanation

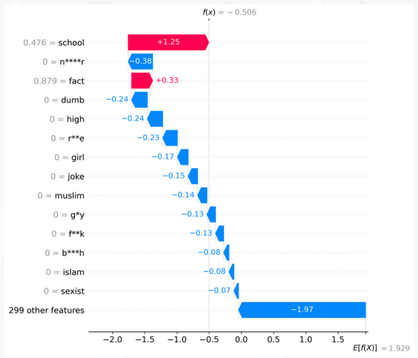
LogisticRegression + BoW:

- **Shap** has been used as post hoc methods
- **positive** coefficients lead towards **cyberbullying**
- **negative** coefficients lead toward **not_cyberbullying**

Cyberbullying:
0.9996 as SHAP prediction



Not_cyberbullying:
0.3761 as SHAP prediction



Multiclass Model Local Explanation

RandomForest + TF-IDF:

- **TreeInterpreter** has been used as post hoc methods for **tree-based** mode
- **positive** coefficients lead towards **predicted label**
- **negative** coefficients lead toward **other label**

feature	contribution	tfidf_value
n****r	0.311113	0.588248
dumb	0.264423	0.596447
f**k	0.181319	0.546091
a*s	-0.014693	0.000000
bulli	0.014067	0.000000
school	0.012138	0.000000
muslim	0.006880	0.000000
black	-0.006662	0.000000
joke	0.005607	0.000000
high	0.005505	0.000000
obama	-0.005396	0.000000
r**e	0.005181	0.000000
white	-0.004923	0.000000
g*y	0.004485	0.000000
idiot	0.004453	0.000000

feature	contribution	tfidf_value
sexist	0.457824	0.434186
rt	0.112435	0.324170
im	0.071577	0.316331
woman	0.062021	0.405737
joke	-0.027306	0.000000
r**e	-0.025821	0.000000
g*y	-0.022894	0.000000
christian	0.021899	0.000000
bulli	0.019601	0.000000
right	0.015546	0.000000
hate	-0.013856	0.400914
school	0.013801	0.000000
n****r	0.012050	0.000000
femal	-0.010885	0.000000
f**k	0.010327	0.000000

References

- Wang, J., Fu, K., & Lu, C. T. (2020). ***"SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection.* 2020 IEEE International Conference on Big Data".** https://ieeexplore.ieee.org/document/9378065
- Ahmadinejad, M., Shahriar, N., & Fan, L(2023). **"Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset".** https://www.proquest.com/openview/2e6b484d78e3a1fe0486ec1217dd574c/1?cbl=18750&diss=y&pq-origsite=gscholar
- Sharma, P., Mirzan, S. R., Bhandari, A., Pimpley, A., Eswaran, A., Srinivasan, S., & Shao, L. (2020). ***"Evaluating Tree Explanation Methods for Anomaly Reasoning: A Case Study of SHAP TreeExplainer and TreeInterpreter"***. https://arxiv.org/abs/2010.06734

Thanks for the attention