Cyberbullism Detection System

Course Final Project aa:2024/2025

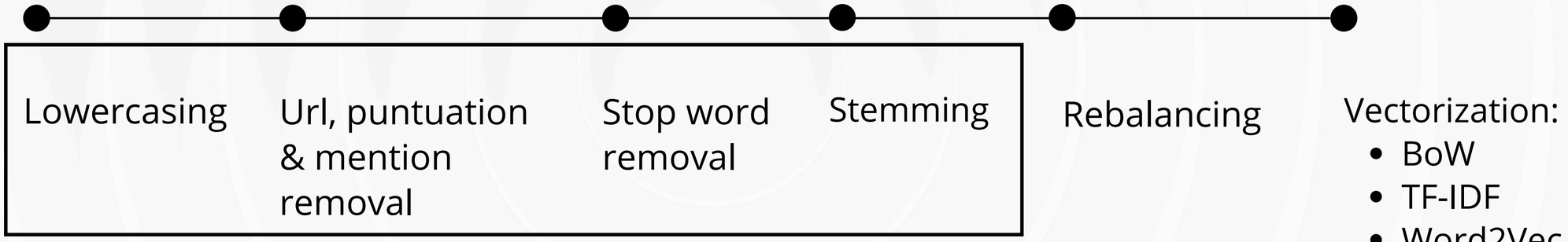
Segreto Mattia

Index

Dataset

- Labaled Dataset
- Almost **45000** 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

Rebalancing Technique

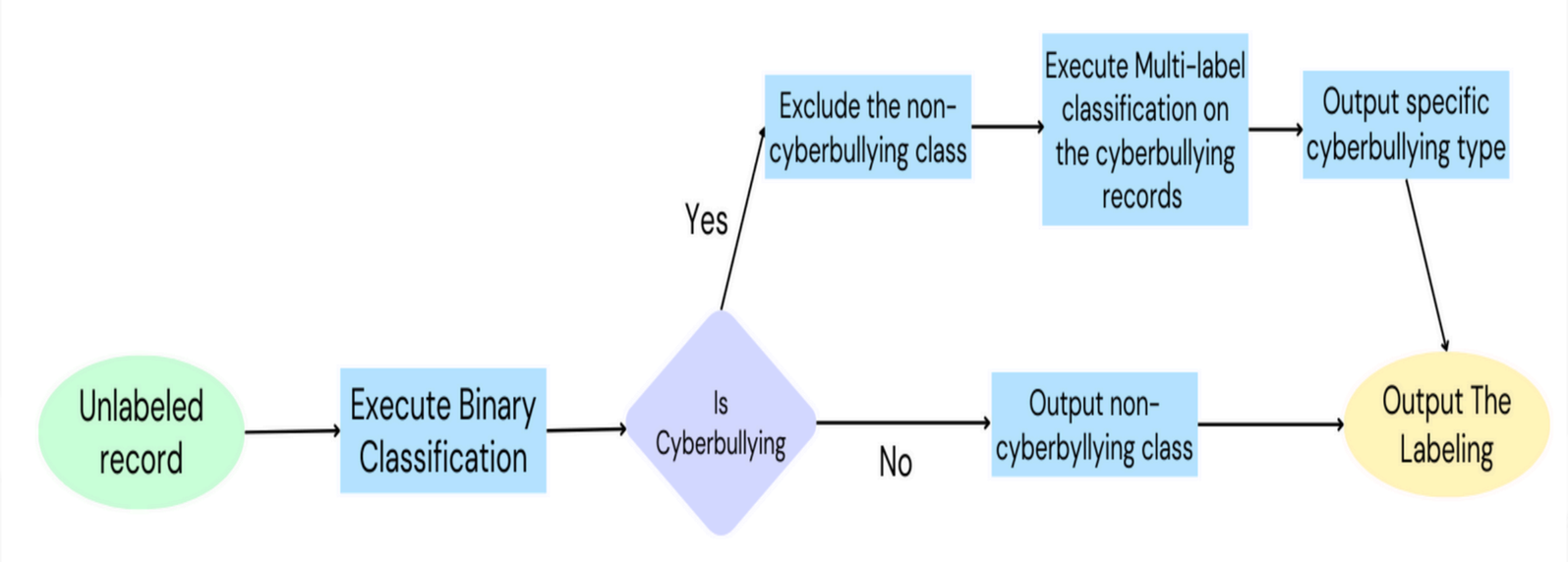
Considered Techniques:

- Domain Agnostic
  - SMOTE
  - SMOTE + round
  - OverSampling
  - UnderSampling
- Domain Specific
  - Synonym Replacement
  - Back Translation

Chosen One:

- Synonym Replacement

Classification Pipeline



Binary Classification

- **RandomForest** 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	Mean Acc.	Mean Prec.	Mean Rec.	Mean F1_macro
TF-IDF	RandomForest	0.905981	0.848335	0.964091	0.854501	0.756927
BoW	RandomForest	0.903343	0.844219	0.962246	0.851271	0.751002
W2V-1	RandomForest	0.902507	0.842162	0.956441	0.854370	0.744036
TF-IDF	LogisticRegression	0.886757	0.821357	0.968221	0.817979	0.731874
TF-IDF	LinearSVM	0.883932	0.817833	0.971030	0.811221	0.730305
BoW	LogisticRegression	0.881775	0.815494			
BoW	LinearSVM	0.876786	0.809010			
W2V-1	LogisticRegression	0.868929	0.793731			
W2V-1	LinearSVM	0.866450	0.790770			

Model	Score
RandomForest + TF-IDF	8
RandomForest + BoW	6
RandomForest + W2V-1	4
LogisticRegression + TF-IDF	2
LinearSVM + TF-IDF	0
LogisticRegression + BoW	-2
LinearSVM + BoW	-4
LogisticRegression + W2V-1	-6
LinearSVM + W2V-1	-8

Multiclass Classification

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

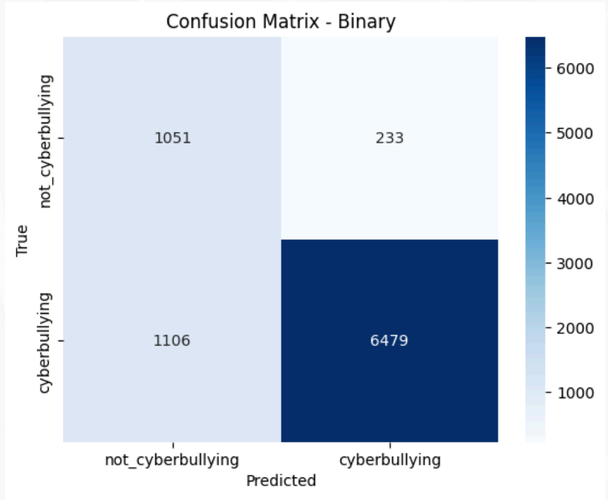
Model Evaluation & Model Comparison Tables

Vectorizer	Classifier	Mean F1	Mean Acc.	Mean Prec.	Mean Rec.
TF-IDF	RandomForest	0.932961	0.936677	0.933369	0.934142
BoW	RandomForest	0.927850	0.931865	0.928208	0.928836
BoW	LogisticRegression	0.927084	0.930480	0.928832	0.928977
TF-IDF	LinearSVM	0.926727	0.930249	0.928194	0.928430
TF-IDF	LogisticRegression	0.926651	0.930085	0.928354	0.928411
BoW	LinearSVM	0.925820			
W2V-1	RandomForest	0.892870			
W2V-1	LinearSVM	0.878508			
W2V-1	LogisticRegression	0.878328			

Model	Score
RandomForest + TF-IDF	8
RandomForest + BoW	4
LogisticRegression + BoW	4
LinearSVM + TF-IDF	2
LogisticRegression + TF-IDF	0
LinearSVM + BoW	0
RandomForest + W2V-1	-4
LinearSVM + W2V-1	-7
LogisticRegression + W2V-1	-7

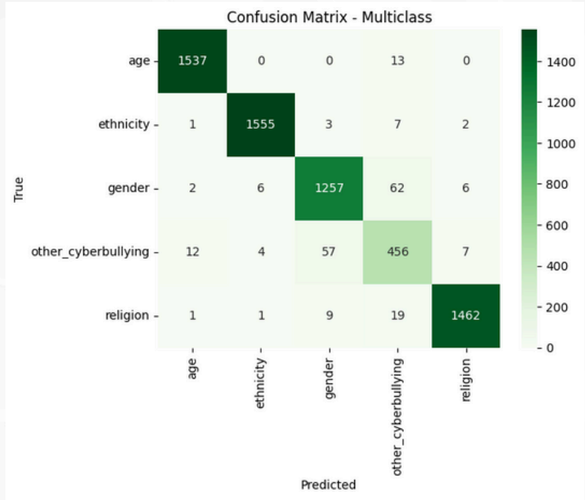
Pipeline Evaluation

Binary Stage:



LogisticRegression+BoW

Multiclass Stage:



Randomforest + TF-IDF

Explanation Module Overview

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

Pattern Mining

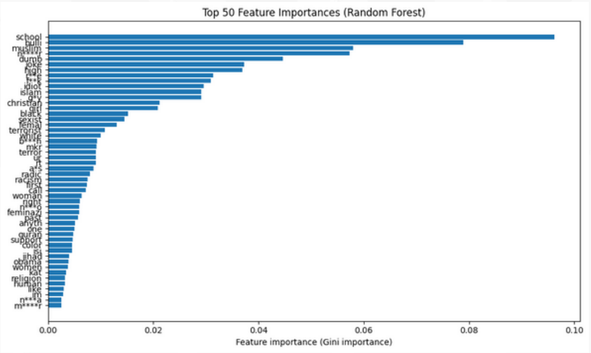
Closed and maximal itemset were analyzed in order to discover if the cyberbullyng phenomenon has **frequent pattern** that shold be useful to discover **interesting informations**

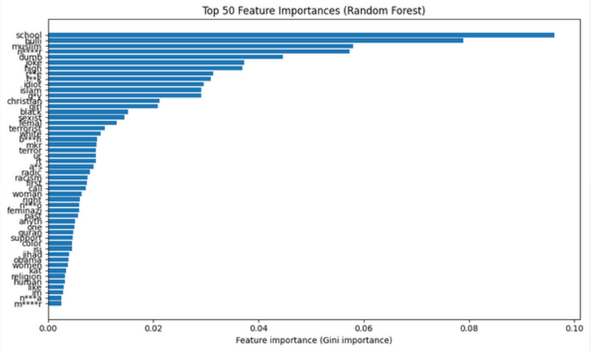
Class	Number of closed itemsets	Number of maximal itemsets
Multiclass Target		
age	144	24
ethnicity	127	26
religion	68	31
gender	62	17
other_cyberbullying	8	8
Binary Target		
cyberbullying	54	23
not_cyberbullying	4	4

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 in both cases





Local Model 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