THE “Family Matters” NON-PROFIT ORGANIZATION

FAMILY-CENTRIC APPROACH TO HEALTHCARE

Prepared and presented by
Cristina Sahoo
on behalf of
FAMILY MATTERS

Problem statement:

Toddler and teenage years are some of the most difficult years to navigate through for parents and children. A non-profit group of doctors and psychologists, called "Family Matters", wants to better serve their community of parents by:

1. *staying relevant by addressing current issues through articles, events, etc.*

2. *directing new calls and messages to the correct department for expert help*

Data Source:

(1) The r/Parenting subreddit: https://www.reddit.com/r/Parenting/

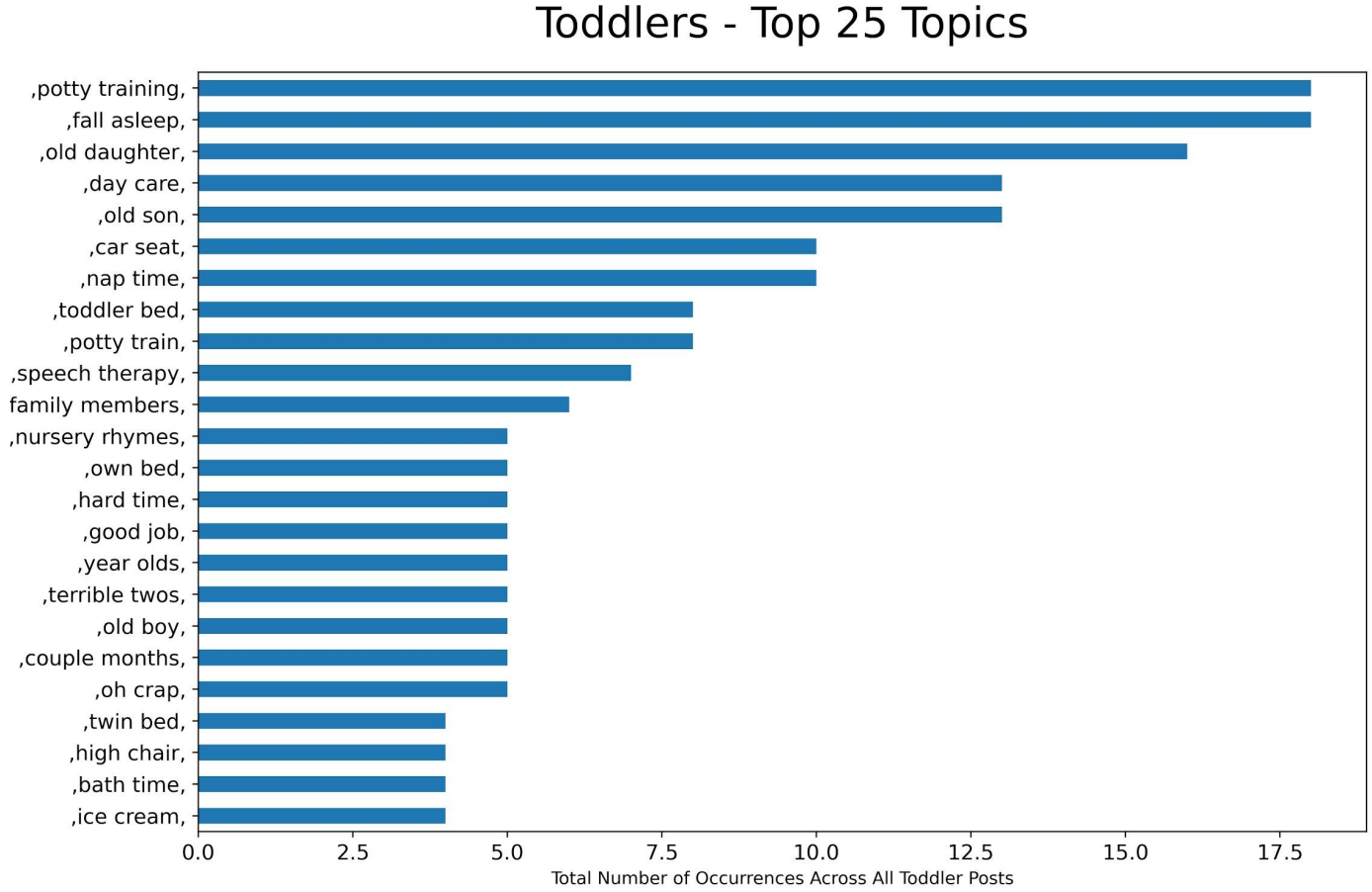
(2) The r/toddlers subreddit: https://www.reddit.com/r/toddlers/

Why is this important?

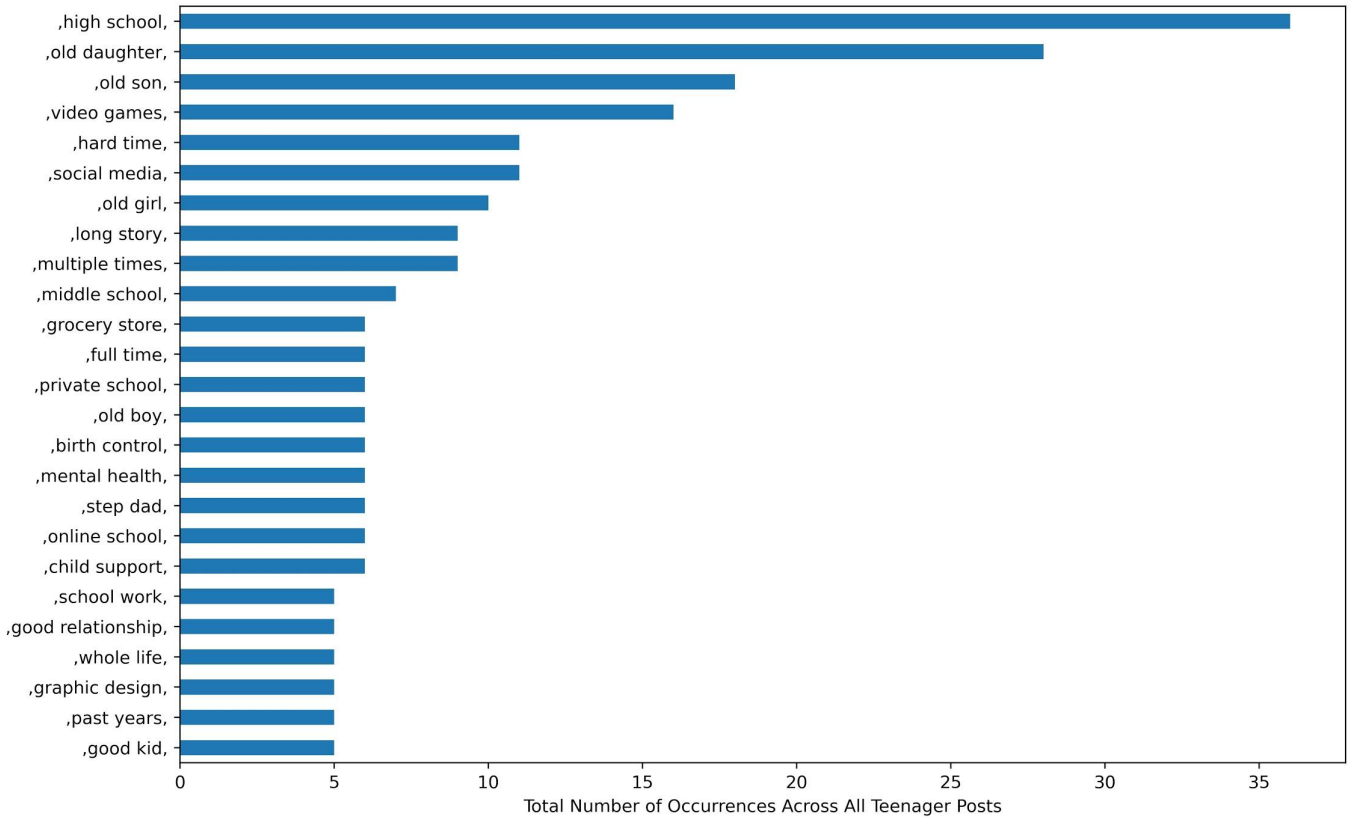
- Due to limited resources, it can be difficult to discuss all topics and address all incoming requests from the community
- By focusing on the top major current issues, the organization can stay useful and relevant
- By routing incoming requests to the correct departments, the organization can provide better advice and help more people

Tools used:

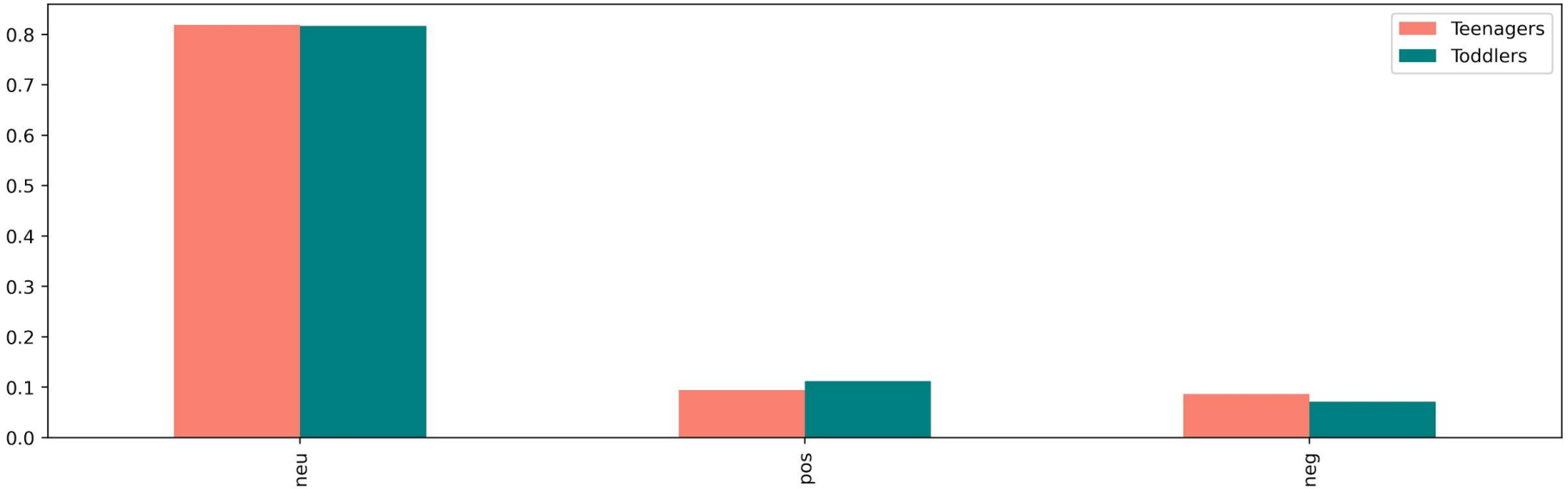
- Natural Language Processing Methods and Techniques
- Exploratory Data Analysis of Text Data
- Classification Models
- Sentiment Analysis


Toddlers - Top 25 Topics
,potty training,
,fall asleep,
,old daughter,
,day care,
,old son,
,car seat,
,nap time,
,toddler bed,
,potty train,
,speech therapy,
family members,
,nursery rhymes,
,own bed,
,hard time,
,good job,
,year olds,
,terrible twos,
,old boy,
,couple months,
,oh crap,
,twin bed,
,high chair,
,bath time,
,ice cream,
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Total Number of Occurrences Across All Toddler Posts

Teenagers - Top 25 Topics



Sentiment for each Class



neg - Negative, neu - Neutral, pos - Positive

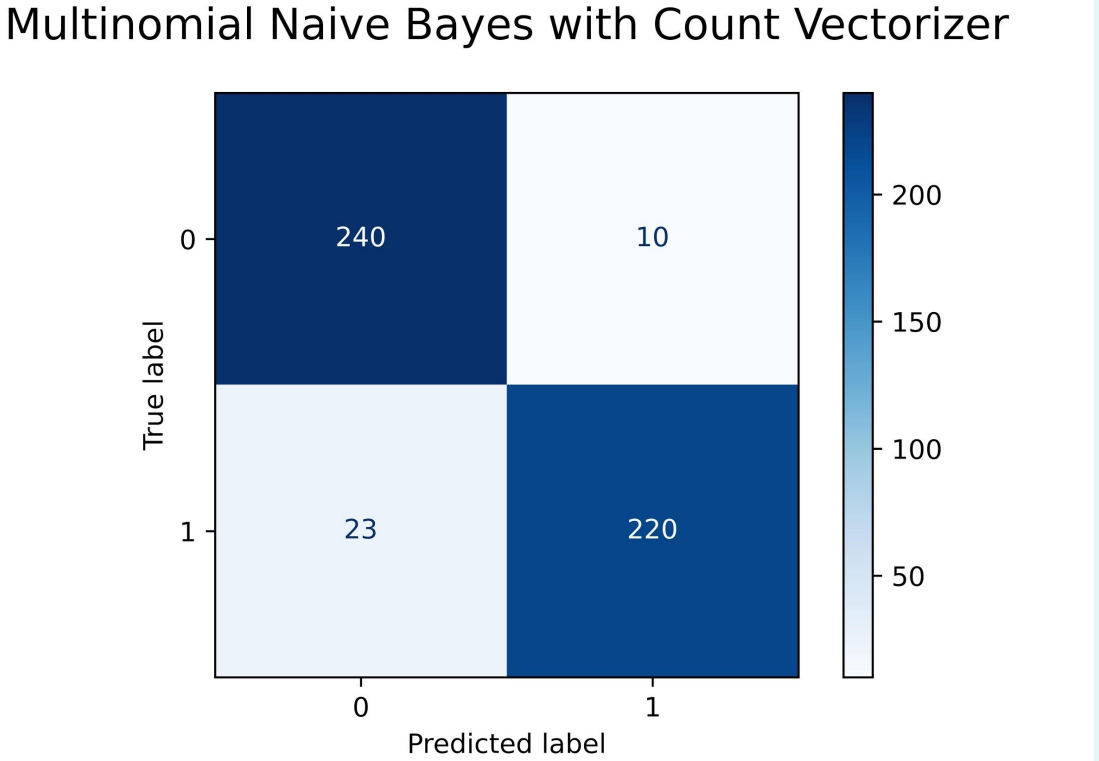
Interpretation:

Out of all posts predicted as class 0 (r/Parenting), we were right 238 times, and we were wrong 12 times.

Out of all posts predicted as class 1 (r/toddler), we were right 218 times, and we were wrong 25 times.

We chose to optimize for accuracy, and the model with the highest accuracy, of 93%, was Multi-binomial NB with CountVectorizer.

This means that for every 100 posts, we classified 93 of them correctly.



Recommendations:

1. Focus resources on developing materials (i.e. books, website content, events) that discuss current issues in the community as previously highlighted.

2. Train new volunteers on addressing these topics when reaching out to the community through education or consulting.

3. Implement our classification model into a new system for routing incoming requests from the community.

Next Steps:

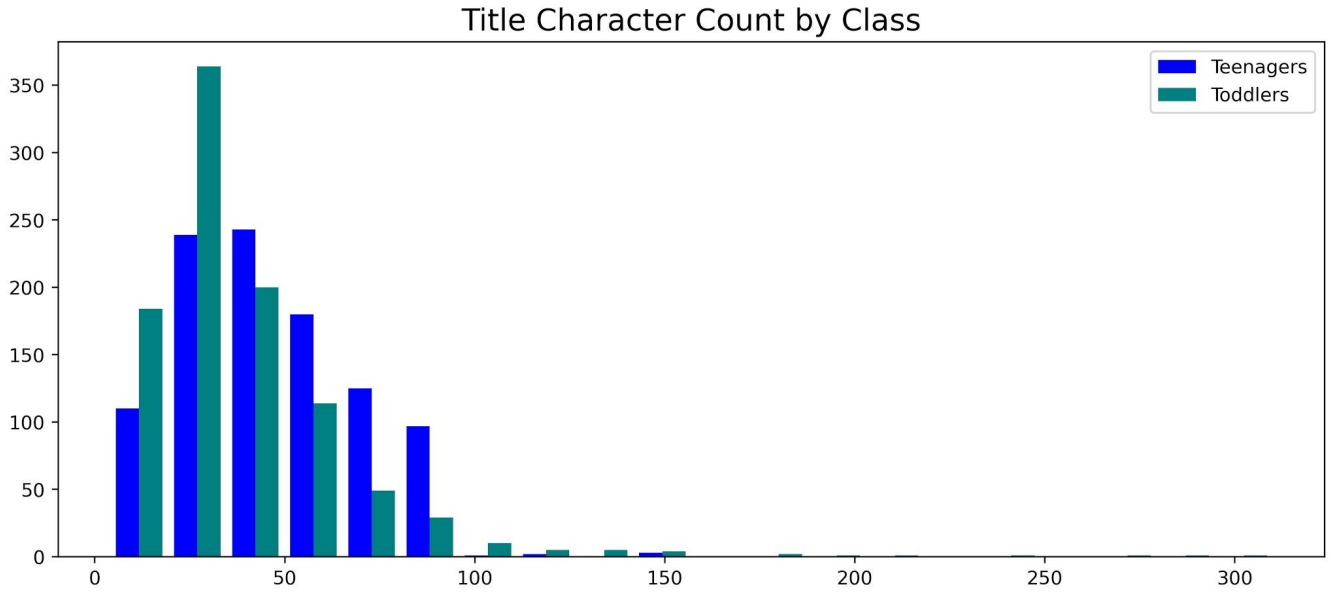
- gather more data
- remove more words that add no value
- consider additional detail, such as post comments
- consider working with other departments of the organization to get a better understanding of the domain
- look into acquiring more funds and computational resources to support the build, run, and evaluation of more models

Resources:

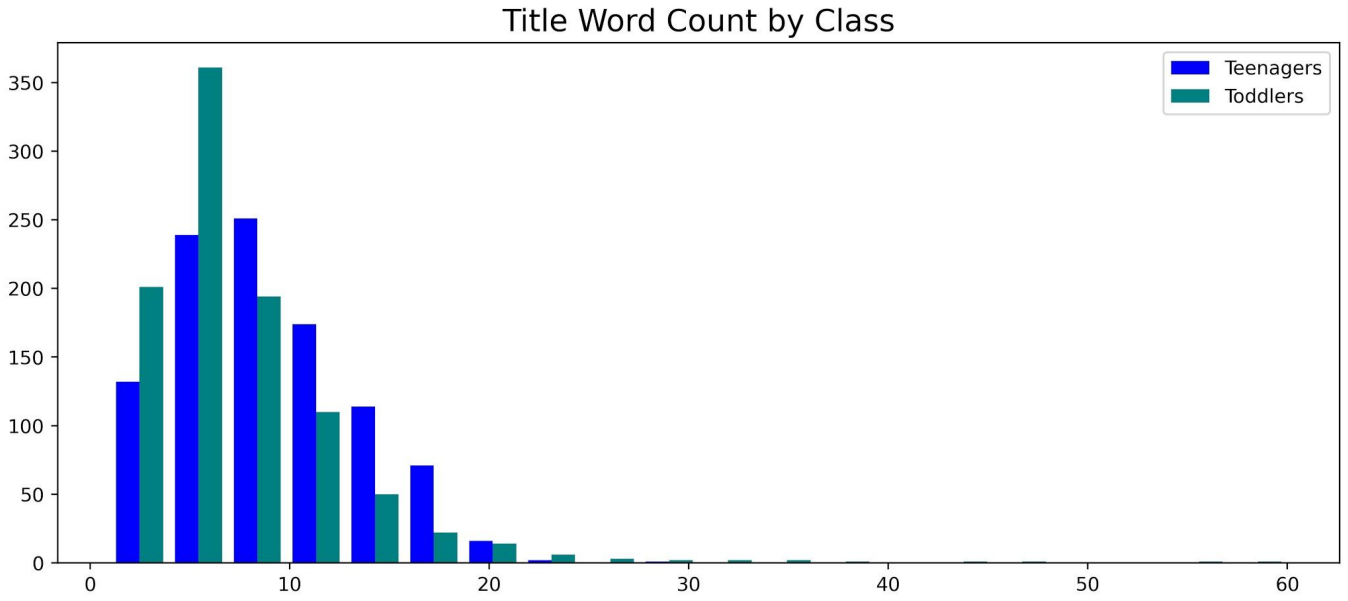
- *Sensitivity and Specificity*
  https://en.wikipedia.org/wiki/Sensitivity_and_specificity
- *Accuracy*
  https://towardsdatascience.com/how-to-evaluate-machine-learning-model-performance-in-python-135b4ae27f7e
- *Precision*
  https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c
- *Executive Summary*
  https://unilearning.uow.edu.au/report/4bi1.html
- *Permission to scrape* https://www.reddit.com/robots.txt
- *Technical Report and Project*
  https://git.generalassemb.ly/cristinasahoo/project_3.git

Questions?

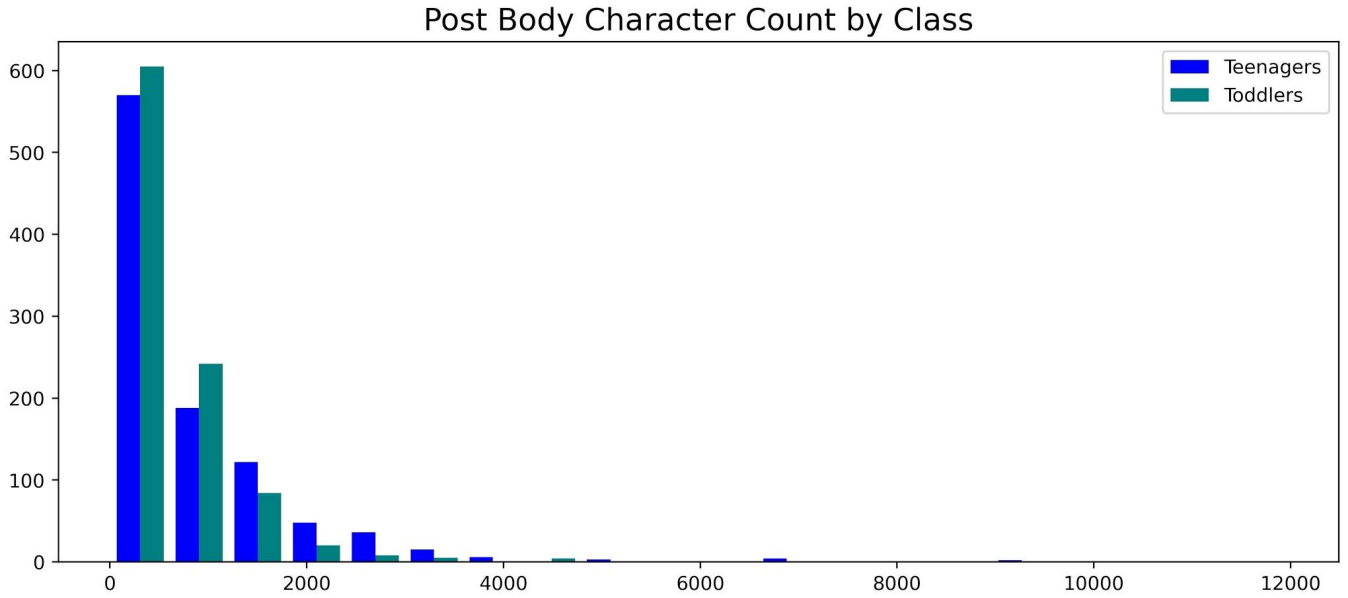
APPENDIX A1.


Title Character Count by Class
Teenagers
Toddlers
350
300
250
200
150
100
50
0
0
50
100
150
200
250
300

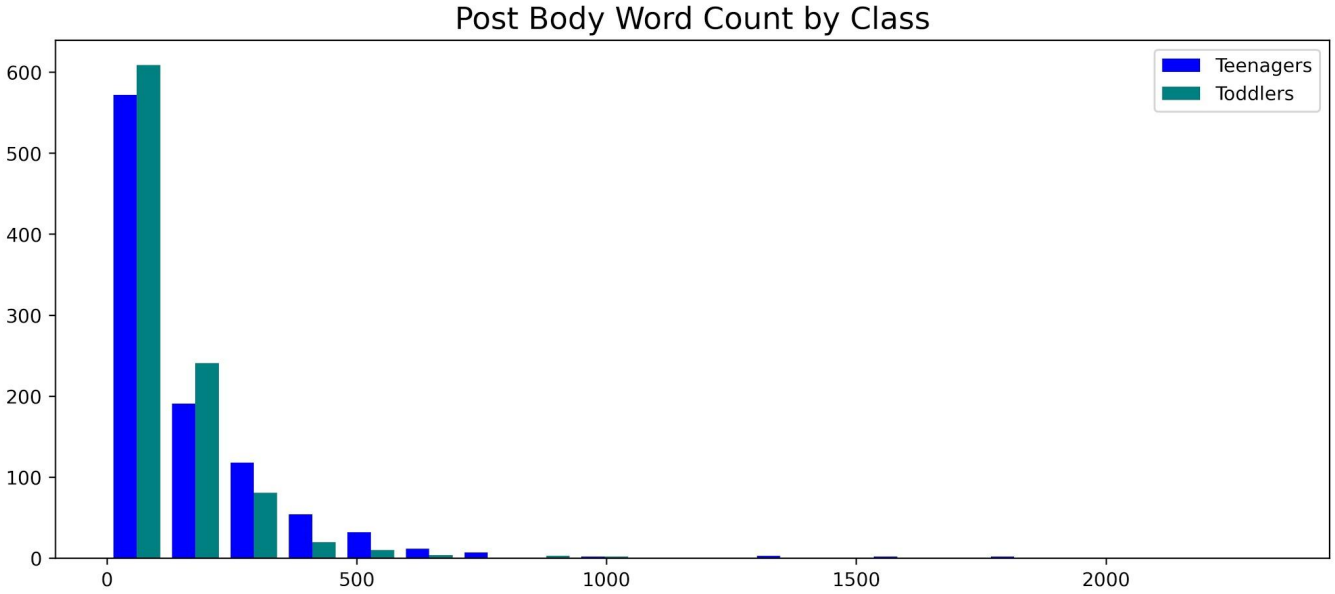
APPENDIX A2.


Title Word Count by Class
Teenagers
Toddlers
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60

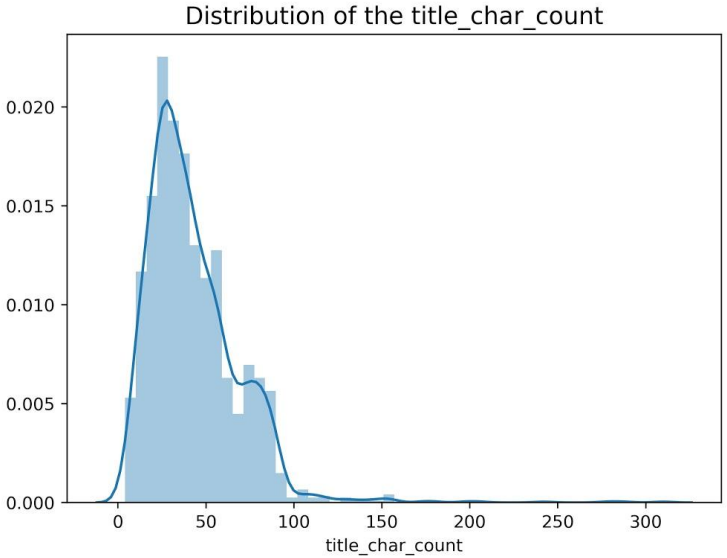
APPENDIX A3.

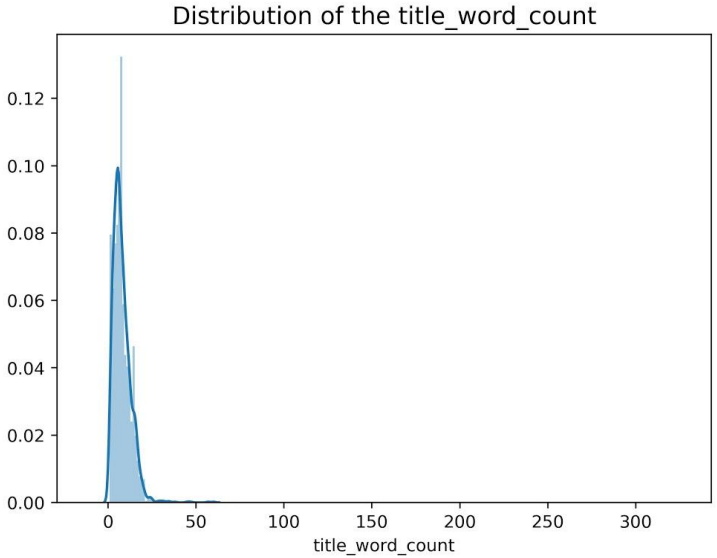

Post Body Character Count by Class
Teenagers
Toddlers
600
500
400
300
200
100
0
0
2000
4000
6000
8000
10000
12000

APPENDIX A4.

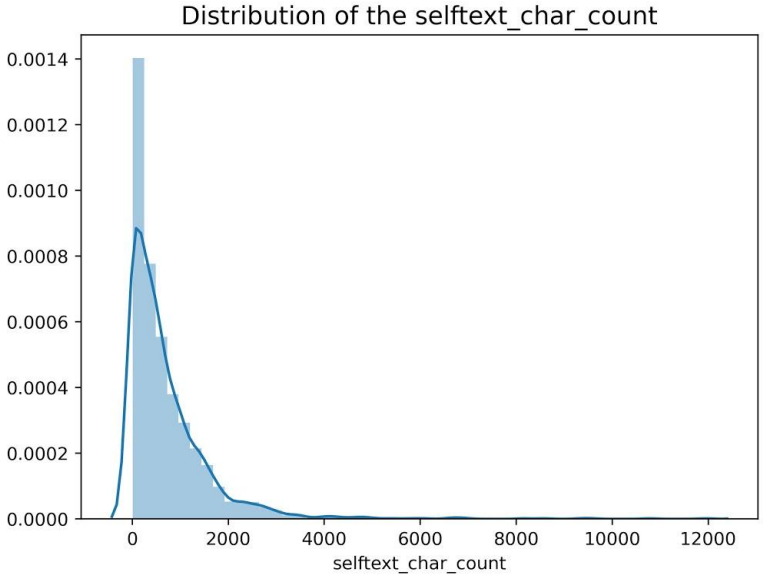

Post Body Word Count by Class
Teenagers
Toddlers
600
500
400
300
200
100
0
0
500
1000
1500
2000

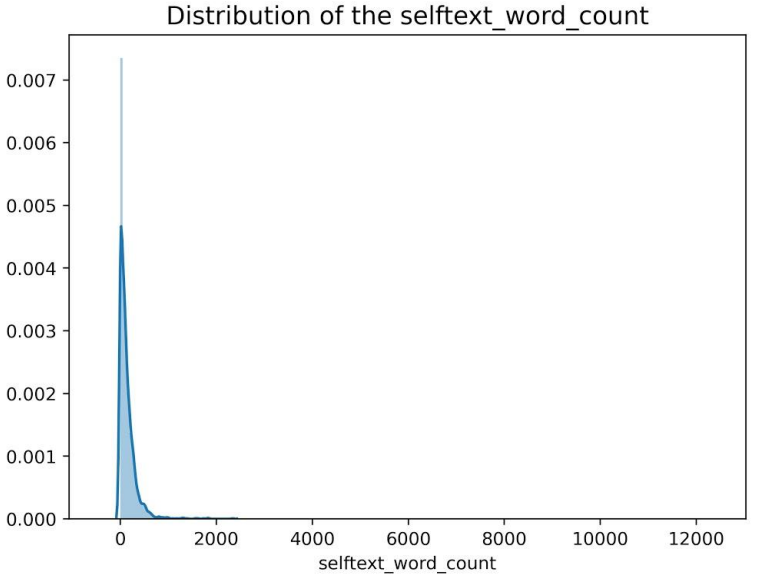
APPENDIX A5.


Distribution of the title_char_count
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
title_char_count

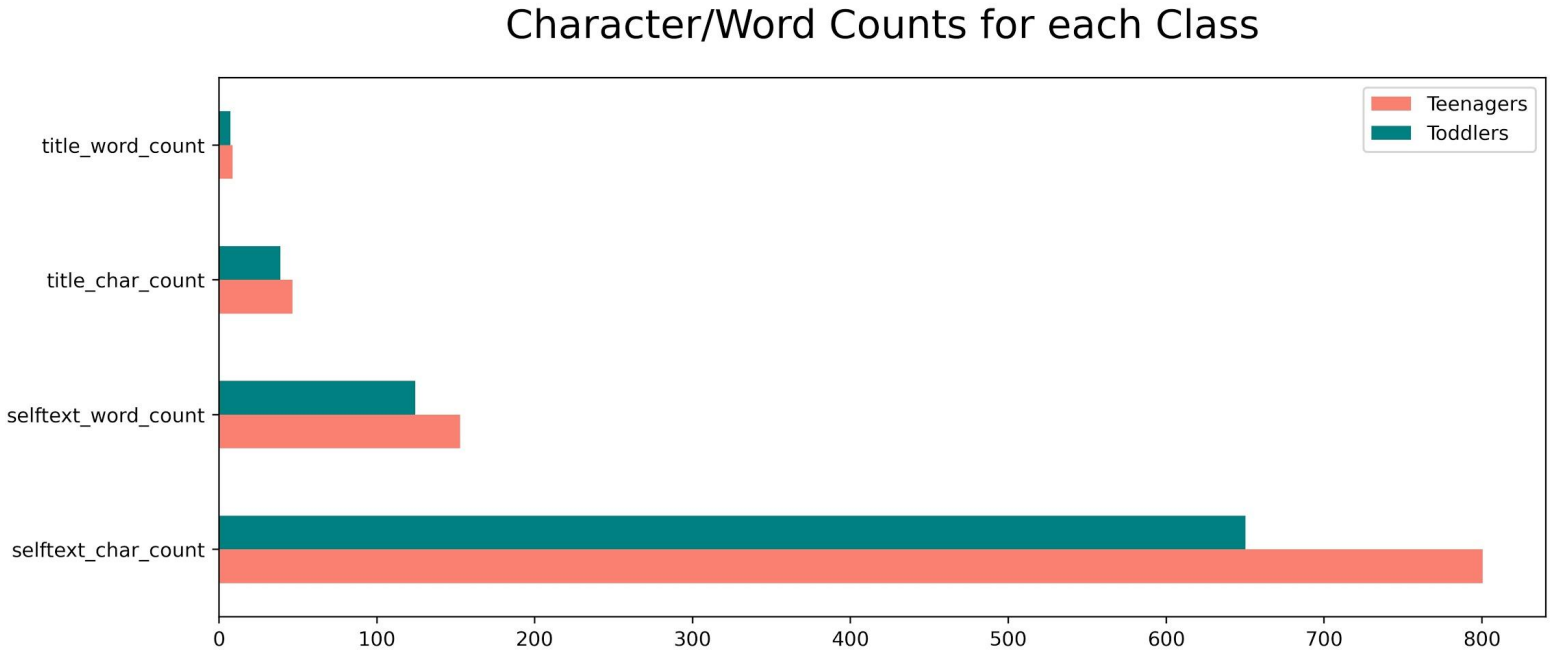

Distribution of the title_word_count
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
250
300
title_word_count

APPENDIX A6.

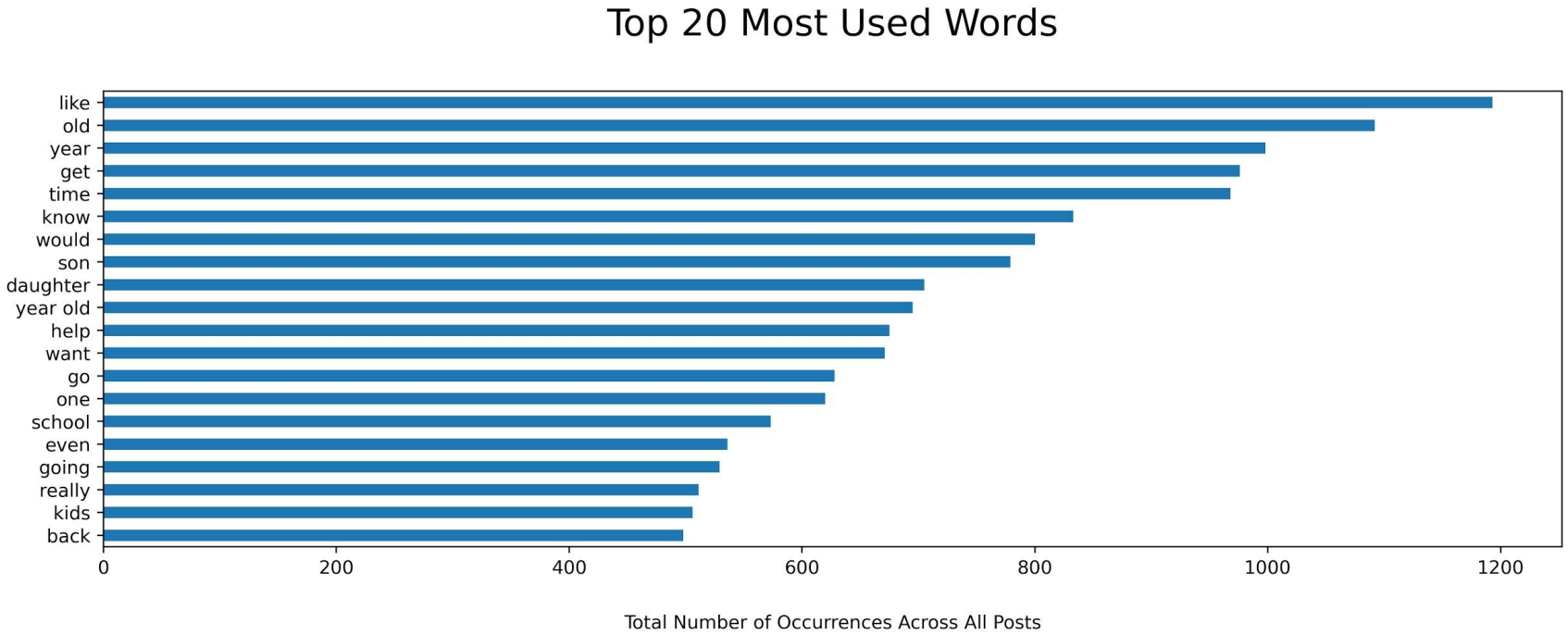

Distribution of the selftext_char_count
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
2000
4000
6000
8000
10000
12000
selftext_char_count


Distribution of the selftext_word_count
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
2000
4000
6000
8000
10000
12000
selftext_word_count

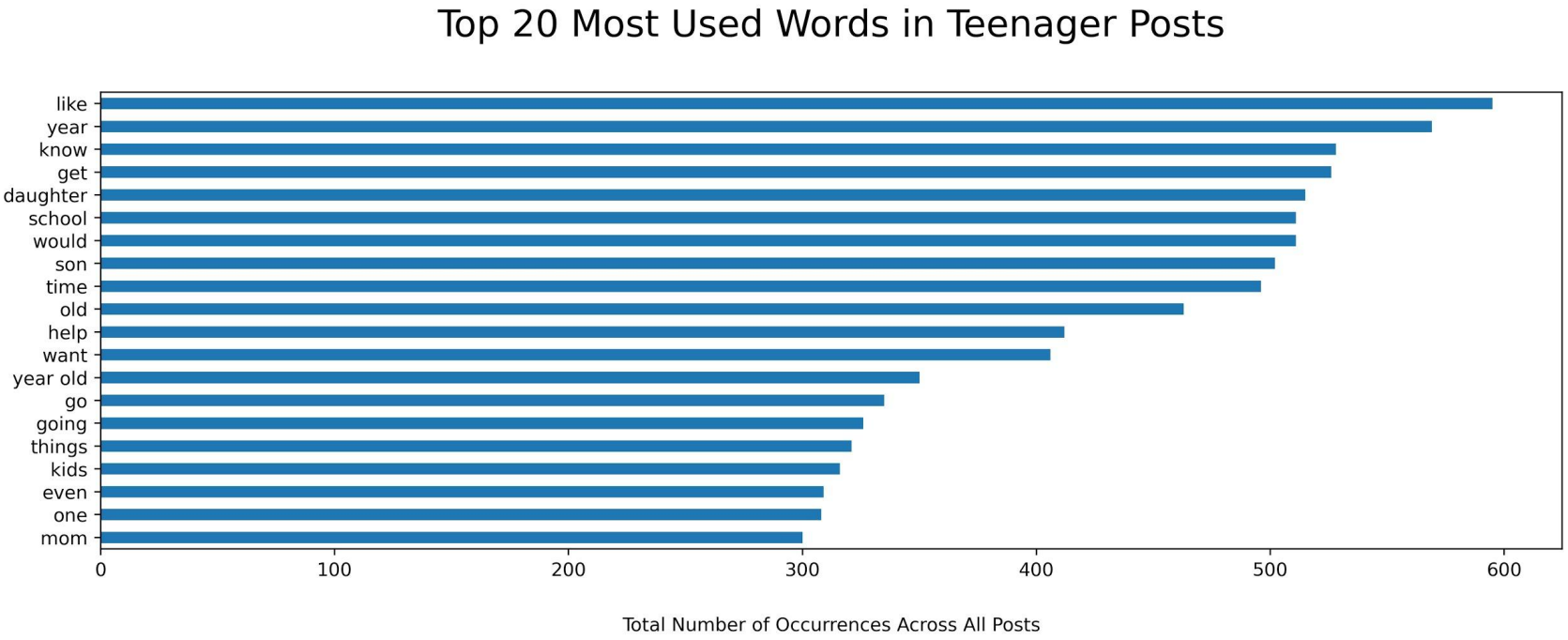
APPENDIX A7.


Character/Word Counts for each Class
Teenagers
Toddlers
title_word_count
title_char_count
selftext_word_count
selftext_char_count
0
100
200
300
400
500
600
700
800

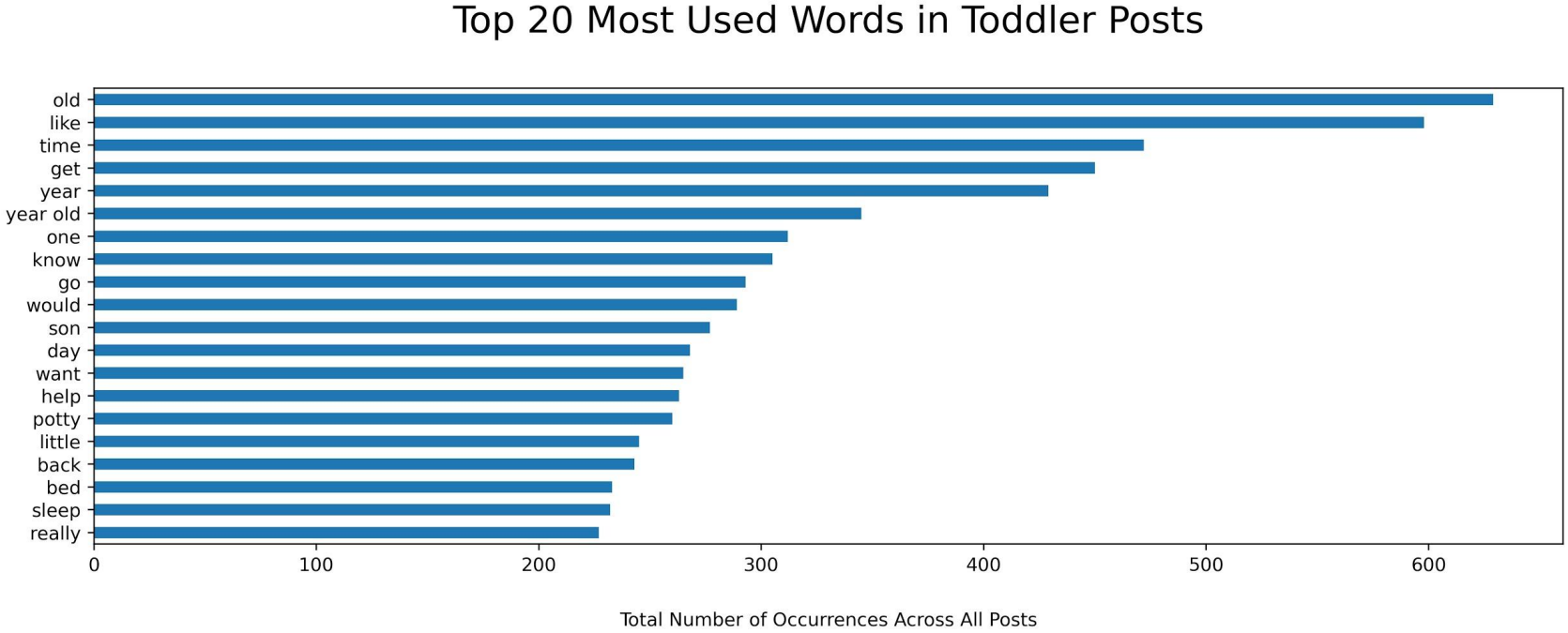
APPENDIX A8.


Top 20 Most Used Words
like
old
year
get
time
know
would
son
daughter
year old
help
want
go
one
school
even
going
really
kids
back
0
200
400
600
800
1000
1200
Total Number of Occurrences Across All Posts

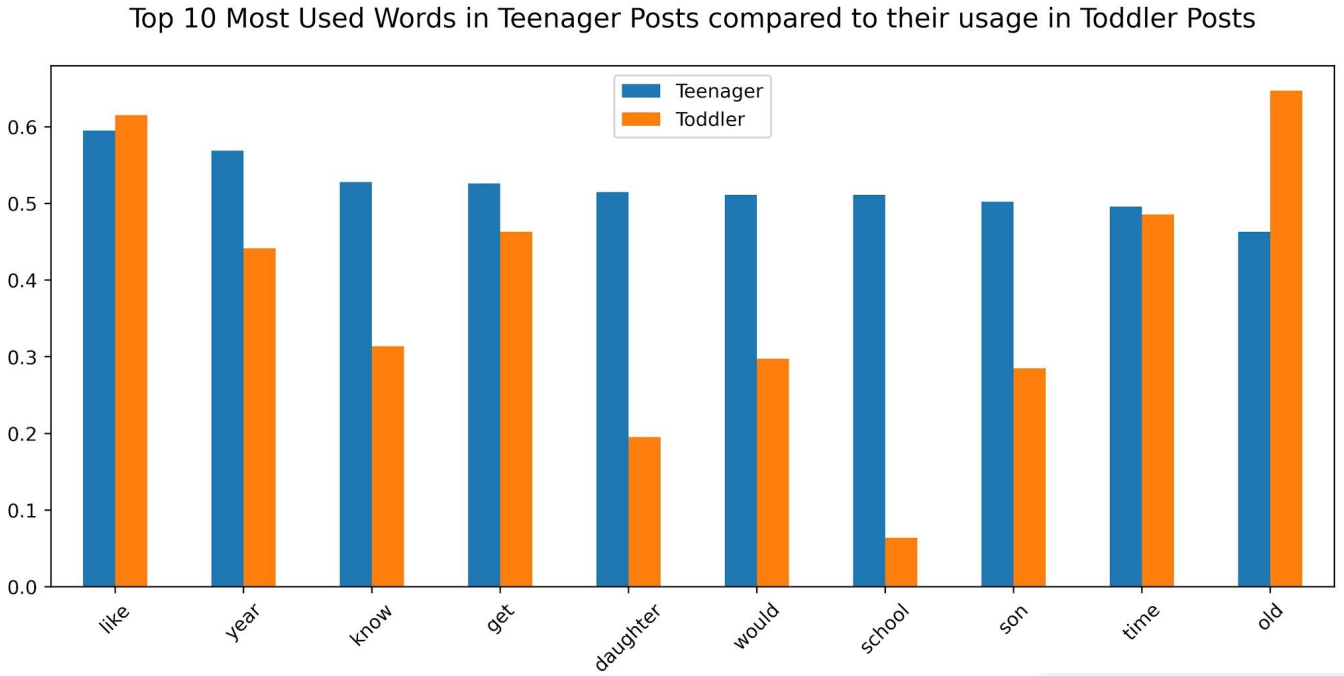
APPENDIX A9.


Top 20 Most Used Words in Teenager Posts
like
year
know
get
daughter
school
would
son
time
old
help
want
year old
go
going
things
kids
even
one
mom
0
100
200
300
400
500
600
Total Number of Occurrences Across All Posts

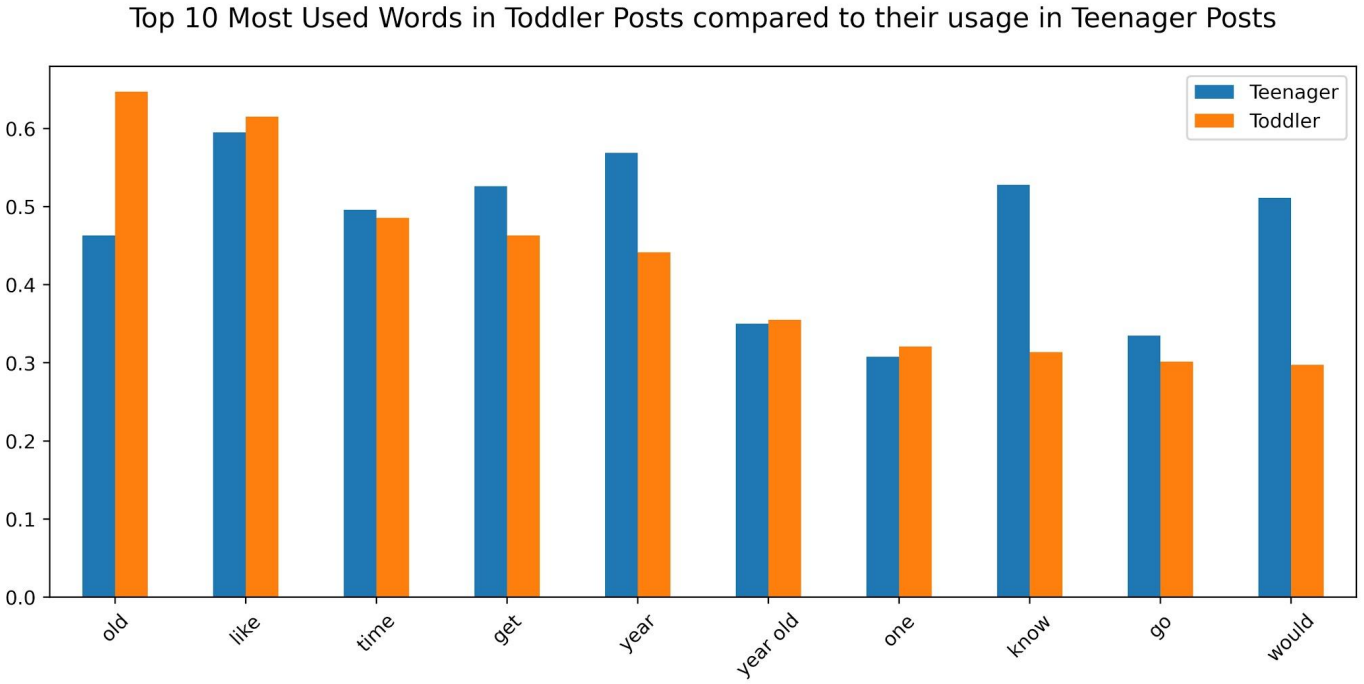
APPENDIX A10.



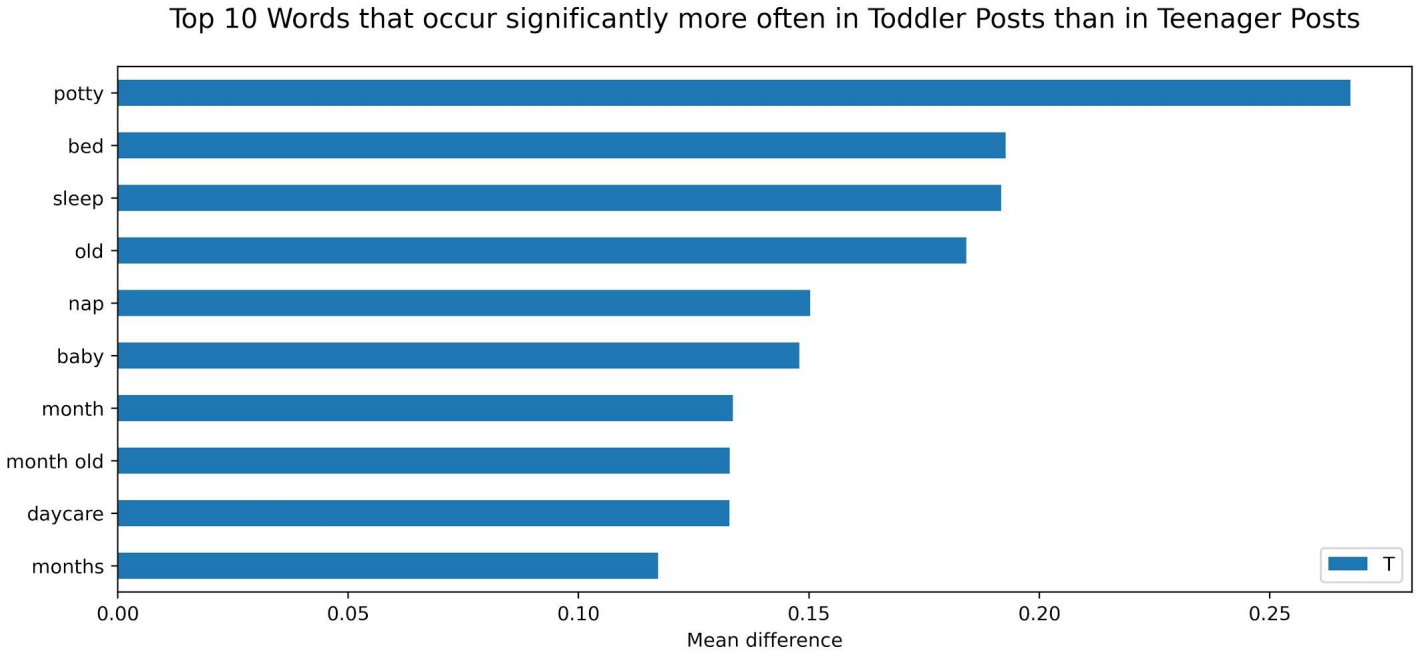
APPENDIX A11.



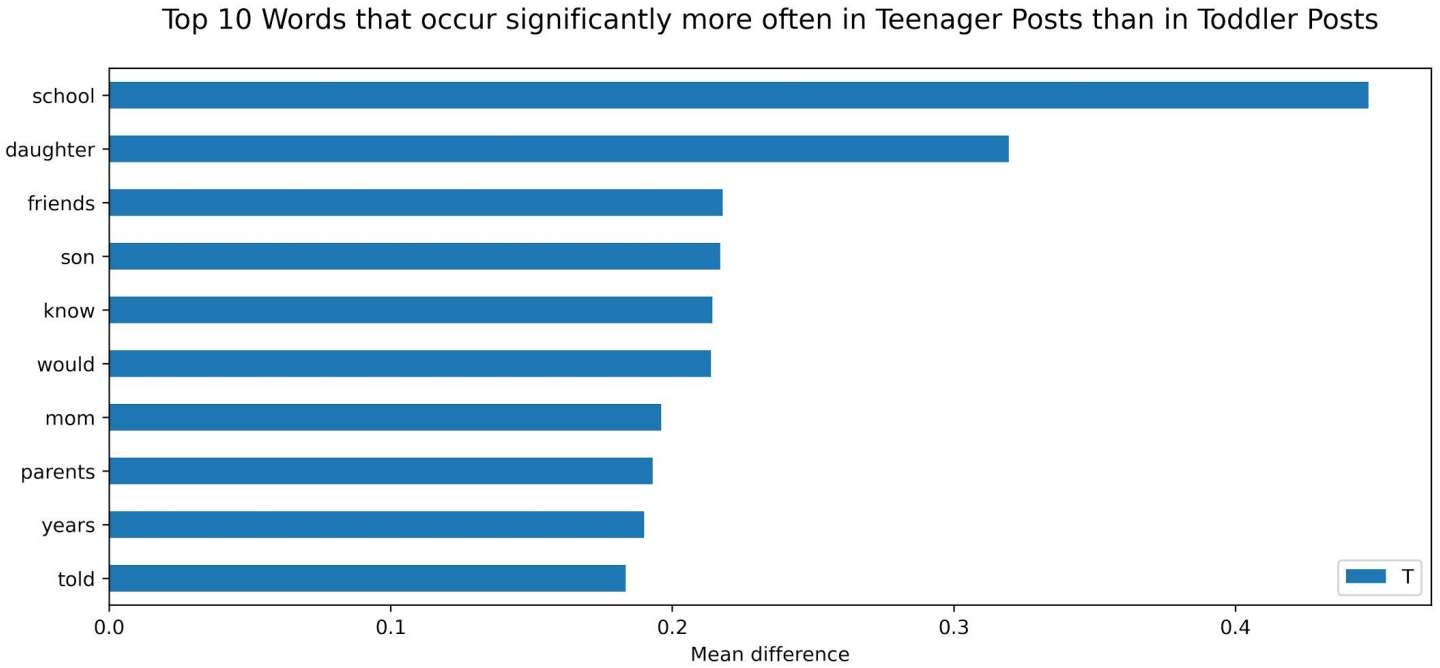
APPENDIX A12.



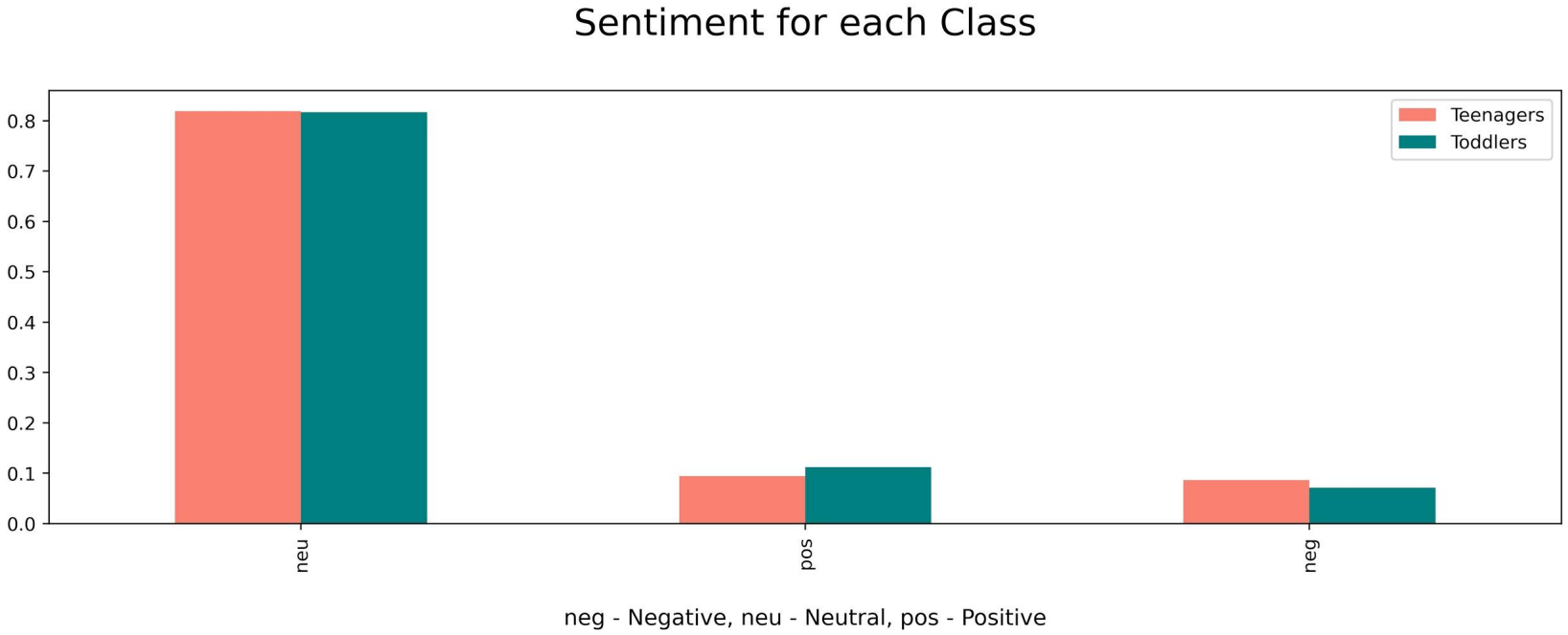
APPENDIX A13.



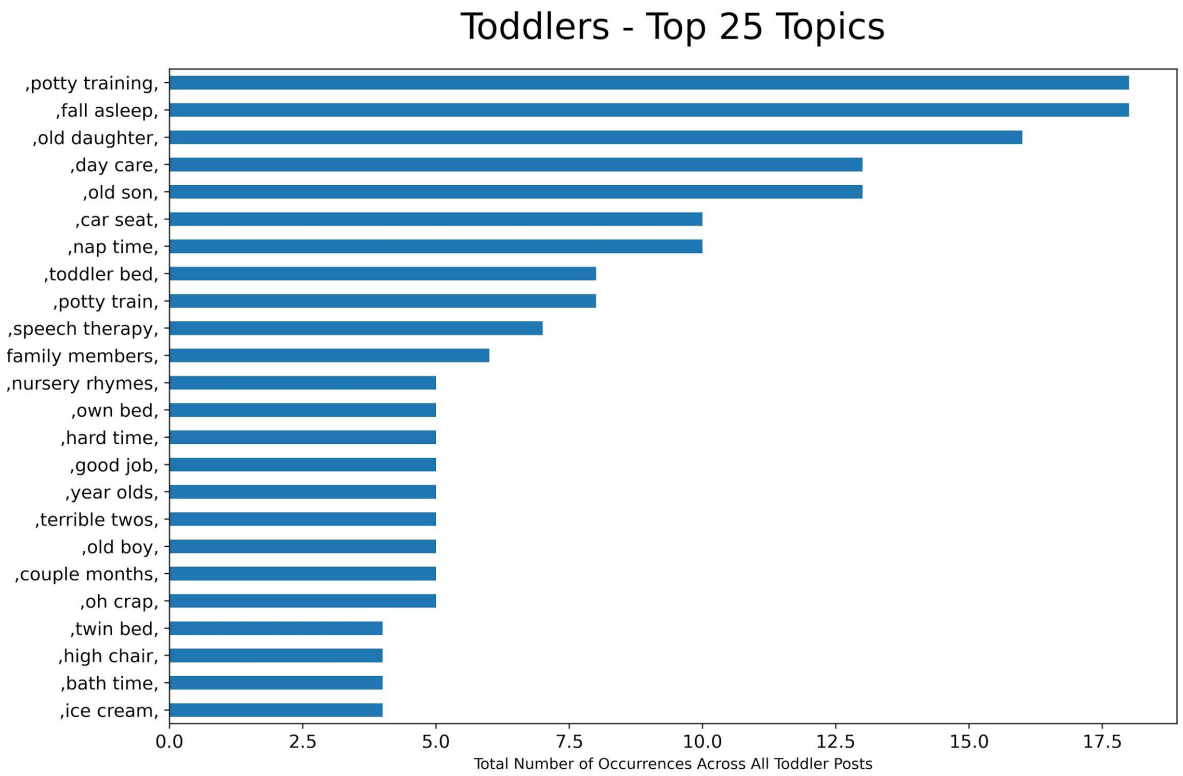
APPENDIX A14.



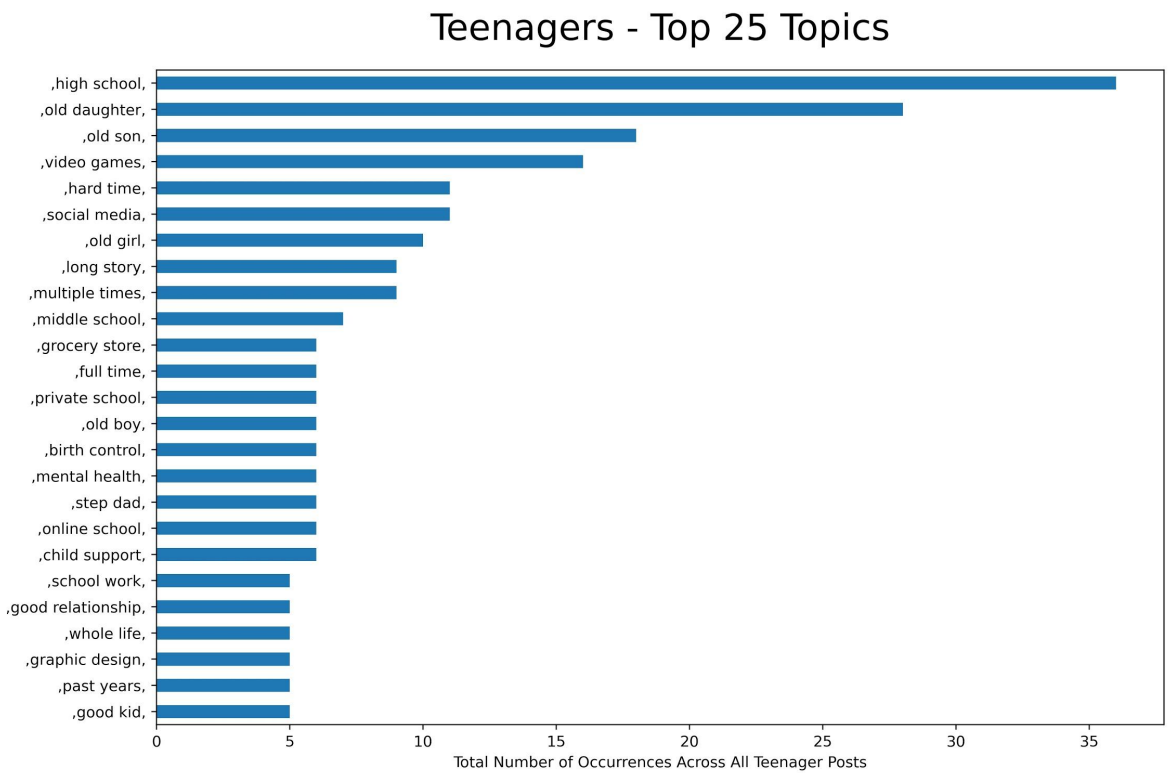
APPENDIX A15.



APPENDIX A16.

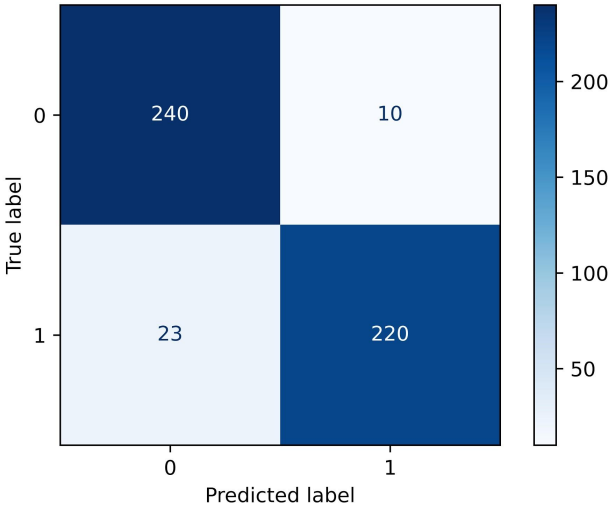


APPENDIX A17.

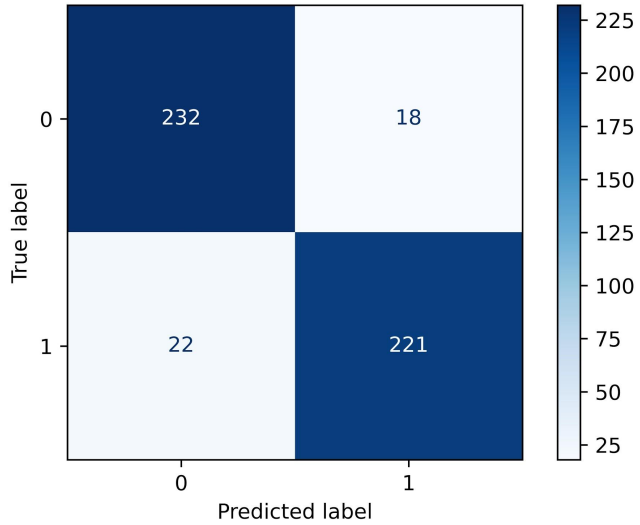


APPENDIX A18.

Multinomial Naive Bayes with Count Vectorizer

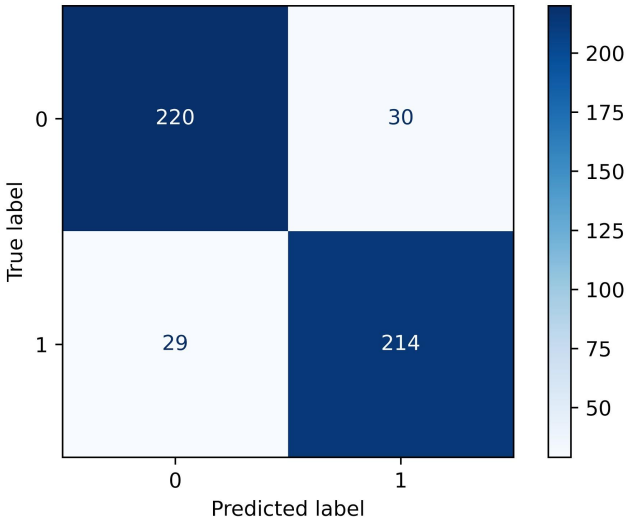


Multinomial Naive Bayes with TFIDF Vectorizer

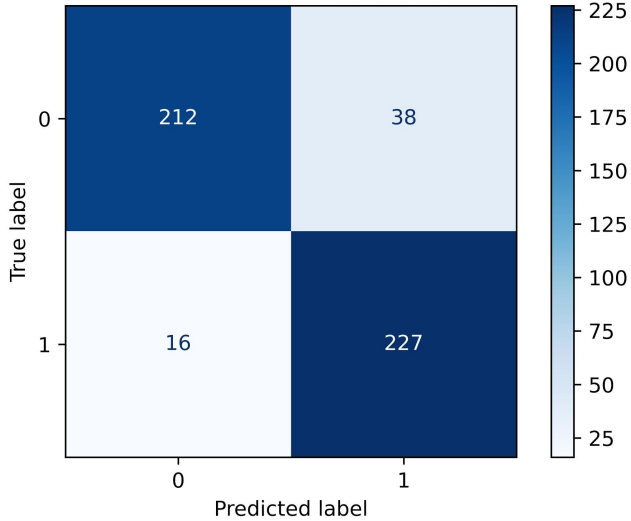


APPENDIX A19.

Random Forest with Count Vectorizer

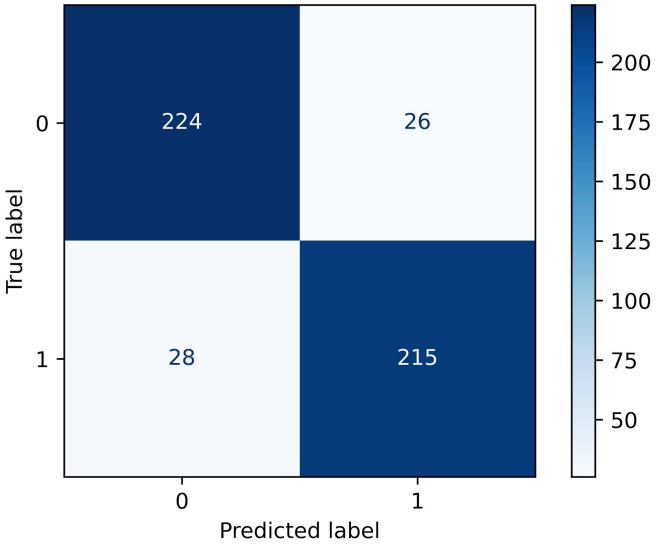


Random Forest with TFIDF Vectorizer

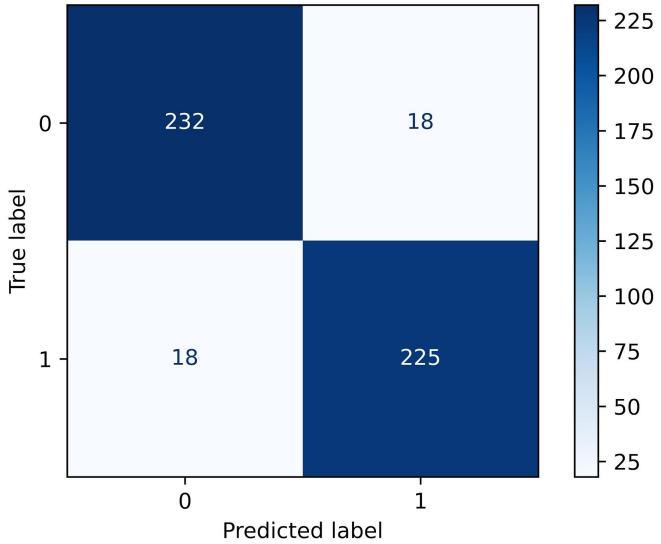


APPENDIX A20.

SVC with Count Vectorizer

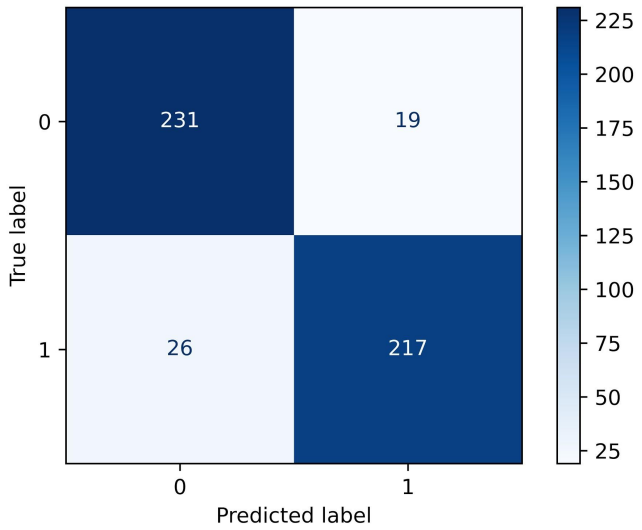


SVC with TFIDF Vectorizer

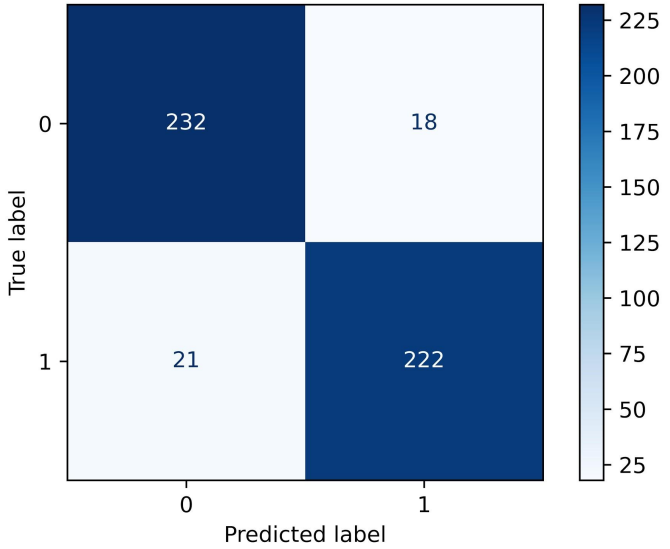


APPENDIX A21.

Logistic Regression with Count Vectorizer



Logistic Regression with TFIDF Vectorizer



APPENDIX A22.

	NB with CV	NB with TFIDF	LR with CV	LR with TFIDF	RF with CV	RF with TFIDF	SVC with CV	SVC with TFIDF
best score	0.95	0.9506	0.9297	0.9432	0.8939	0.9108	0.8783	0.9446
train score	0.9784	0.9797	0.998	0.9885	0.9432	1	0.9675	1
test score	0.9331	0.9189	0.9087	0.9209	0.8844	0.8966	0.8905	0.927
sensitivity	0.9053	0.9095	0.893	0.9136	0.8683	0.9506	0.8848	0.9259
specificity	0.96	0.928	0.924	0.928	0.9	0.844	0.896	0.928
precision	0.9565	0.9247	0.9195	0.925	0.8941	0.8556	0.8921	0.9259
accuracy	0.9331	0.9189	0.9087	0.9209	0.8844	0.8966	0.8905	0.927

APPENDIX A23.

CV	CountVectorizer
NB	Multinomial Naive Bayes
TFIDF	TfidfVectorizer
LR	Logistic Regression
RF	Random Forest Classifier
SVC	Support Vector Classifier