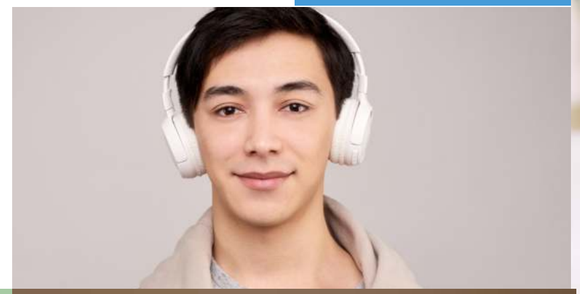
Sentiment Analysis + Subreddit Classification



Data Science Project 3

By Cathy Kam

4 Dec 2022

Project's Aim:

To use Pushshift's API as a form of webscrapping to collect posts from 'Pokemon' and 'Pokemontrade' subreddits, then:

I will apply:

Option 1: Pipeline: CountVectorizer + MultinomialNB
Option 2: Pipeline: CountVectorizer + Term frequency Inverse document frequency (TFIDF) + MultinomialNB

as text classifier models to distinguish which subreddit that a given post should belong to.

After then, an exploratory data analysis will be conducted to find out the sentiment score for each subreddit.

Problem Statements:

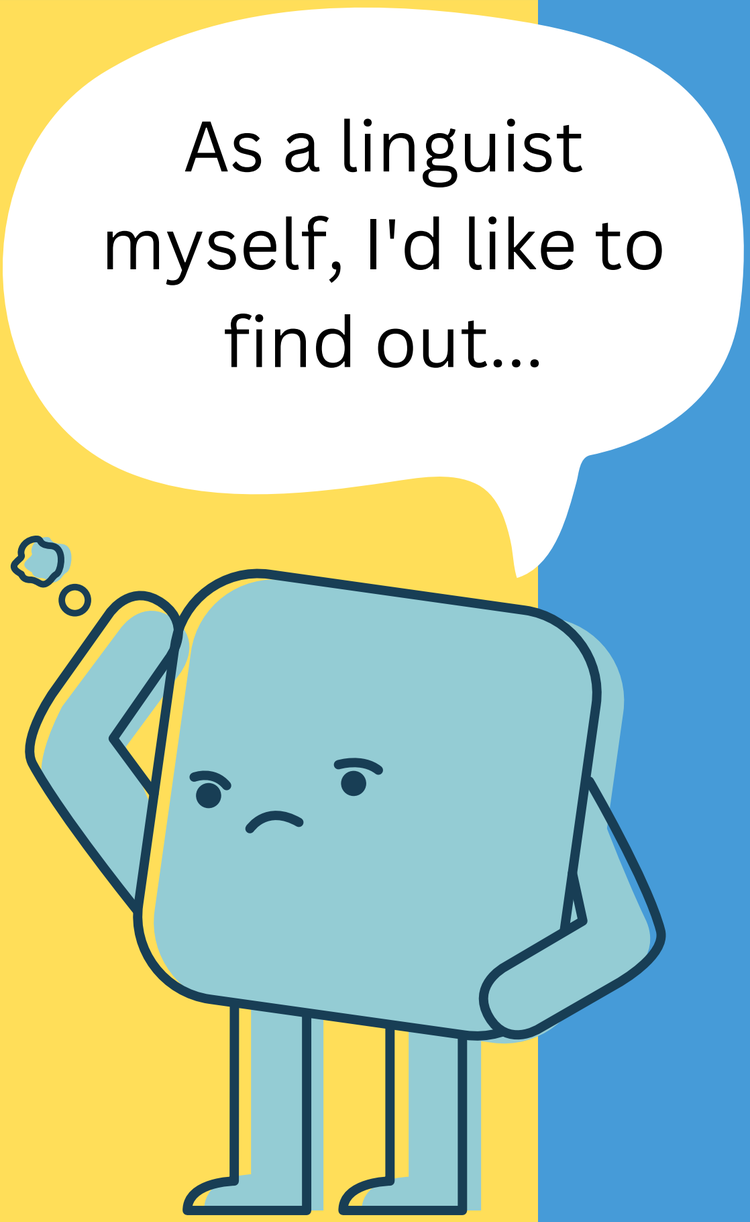
1. What do the users care about in 'pokemontrade' and 'pokemon' subreddits?

2. Are there any differences between those two?

3. What are the heated topics that are discussed among the users in both subreddits?

4. Who are the influencers who have the most subscribers?

5. Are adults the majority of the users in these forums?



1.

Dataset Features

A brief Glimpse into the shape of the datasets after cleaning

Pokemon Subreddit:

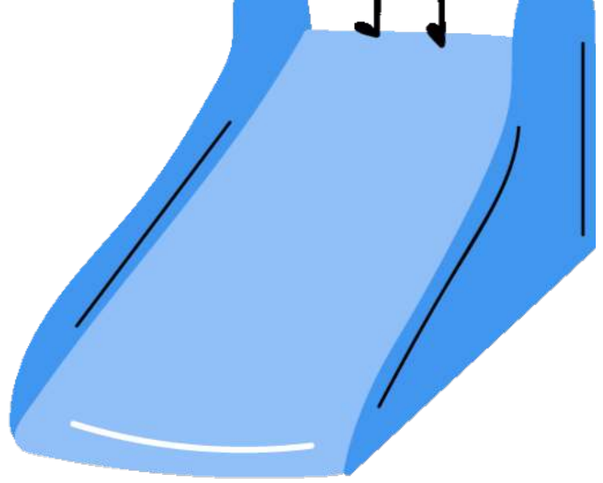
4705 Rows,
107 Columns

Pokemontrade Subreddit:

9215 Rows,
85 Columns

13920 rows in total

Project Structure



Webscrapping
Data Import
Data Cleaning

- Used Pushshift's API to collect data
- Converted the raw data to CSV files
- Deleted all the removed / moderator's / irrelevant posts

Exploratory Data Analysis (EDA) + Sentiment Analysis

- Focused on 'title' and 'selftext' columns
- Got a basic demographic of the users
- Used Afinn score to rate evaluate the sentiment of the subreddits
- Formed some word clouds to highlight the most common words

Classifier Modeling + Evaluation

- Apply and compare 2 pipelines(CountVectorizer + MultinomialNB) and (CountVectorizer + Term frequency Inverse document frequency (TFIDF) + MultinomialNB)

Conclusion / Recommendation

- Solve the problem statements and address the major differences that distinguish the two subreddits
- Select the best model that can run the best prediction

Age Demographic

Users below 18

Pokemontrade 90%
Pokemon 99%

Users over 18

Pokemontrade 9%
Pokemon 1%

It turns out the kids or people under 18 are the majority accounted for these two subreddits.

While only a small minority of adults (>10) were engaging in the 'pokemon' subreddit, more adults are engaging with the 'pokemontrade' subreddit instead. This reflects adults focus more on the practicality and unlocking achievements of the game, while kids spend more time in exchanging thoughts and extending network.

Top 5 Users with Most Subscribers

Pokemon

1. EmiKoizuwumi 46068387.0
2. cshin09 37484929.0
3. Blueeyeswhiteraichu 24890233.0
4. Poll_God 18683148.0
5. Bluecomments 16586987.0

Pokemontrade

1. -Shiny_Star- 11807228.0
2. Porygon-Bot 10919769.0
3. Theduskwolf 8195275.0
4. chenj25 7680800.0
5. hoozayisdead 4543350.0

The Number of Posts

Pokemontrade

9215

Pokemon

4705

According to the post count, 'pokemontrade' subreddit has significant a larger number of post count compared to 'pokemon'. It is almost double of what 'pokemon' subreddit has, implicating a higher user engagement rate overall.

Sentiment Analysis

#AFINN is an English word listed developed by Finn Årup Nielsen.

Each word scores ranging from minus five (negative) to plus five (positive).

The English language dictionary consists of 2,477 coded words.

Average score for Pokemontrade subreddit:

+10

Average score for Pokemon subreddit:

-9

Pokemon Title Sentiment Score

Overall Description:

Mean: 3.39
Maximum: 13
Minimum = -9
Std = 1.75

Example of a (-9) title:

"I bought a new OLED switch for SV, and I made a grave mistake and lost probably 800 hours across Shield, BD, and Arceus. Don’t make my same mistake and put everything into Home right now!"

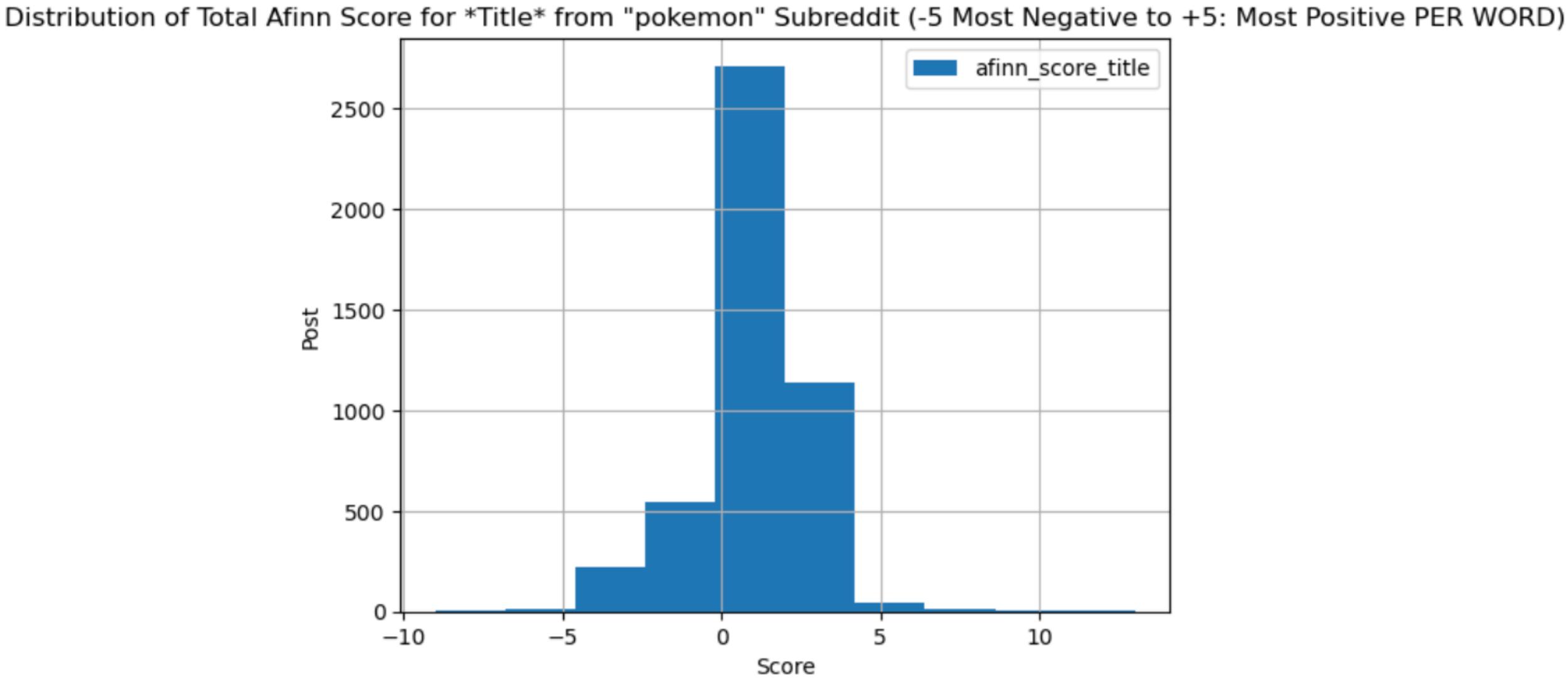
Post Counter:

Positive: 1450
Nagative: 788
Neutral: 2467

Example of a (10) title:

"I just went on a 150 Win Streak at the Battle Maison, far and away my longest one. The team I used really surprised me. What teams got you your best win streaks?"

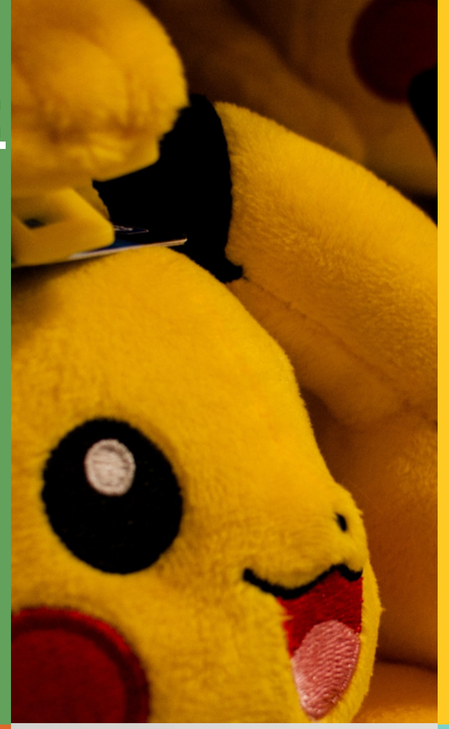
Afinn Score Distribution for 'Title' in Pokemon subreddit
(-5 Most Negative to +5: Most Positive Per Word)



Pokemontrade Title Sentiment Score

Overall Description:

Mean: 0.76
Maximum: 21
Minimum: -12
Std: 1.665



Example of a (-6) title:

"A dumb deal I’m offering for Shield players that you should only consider if you’re an idiot"

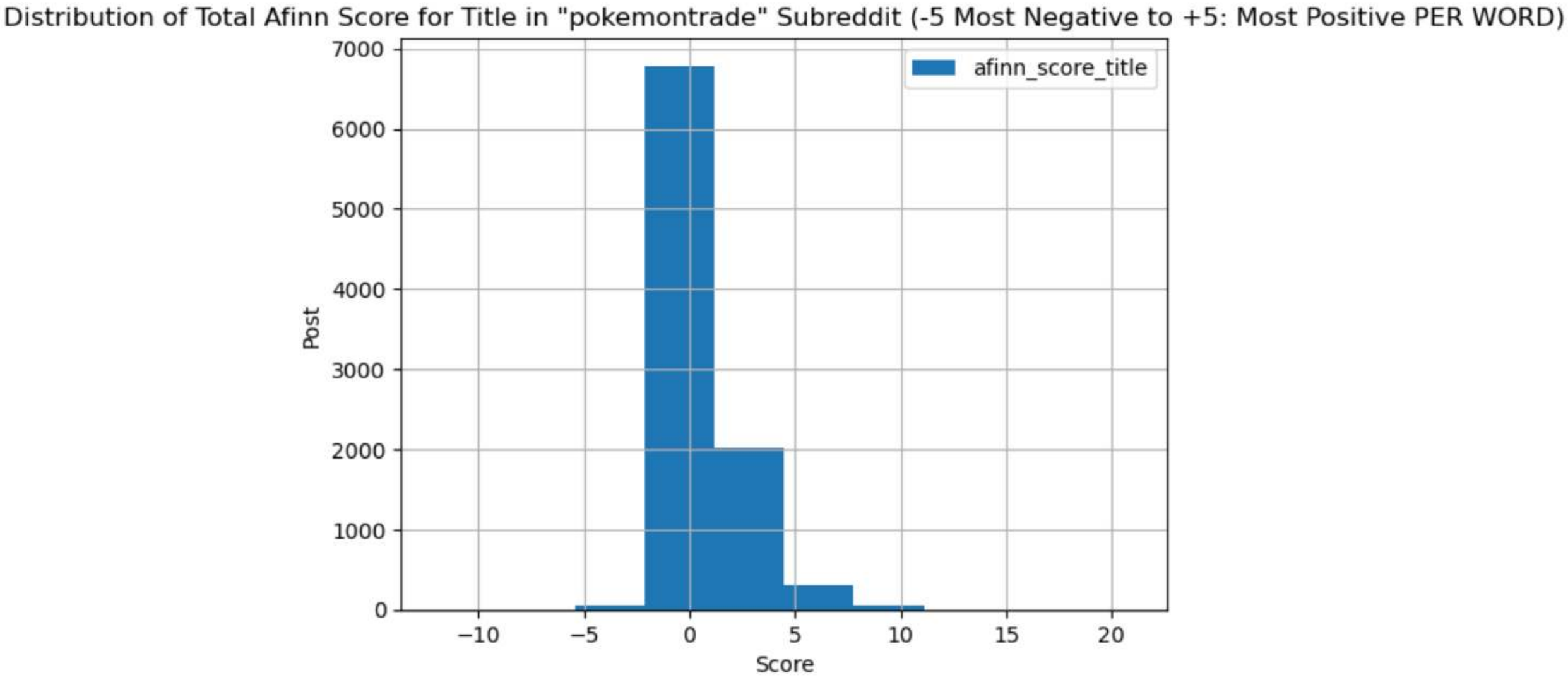
Post Counter:

Positive: 2698
Nagative: 312
Neutral: 6205

Example of a (21) title:

"LF: Fast Ball Shinx (HA preferred), Heavy Ball FT Aprimon: Moon HA Piplup, Lure Piplup, Love HA Torchic, Friend HA Turtwig, Heavy HA Gligar, Heavy Beldum, Moon HA Murkrow, Lure HA Totodile, Level 5IV HA Gible, Friend Larvitar, and MORE IN POST"

Afinn Score Distribution for 'Title' in Pokemontrade subreddit
(-5 Most Negative to +5: Most Positive Per Word)



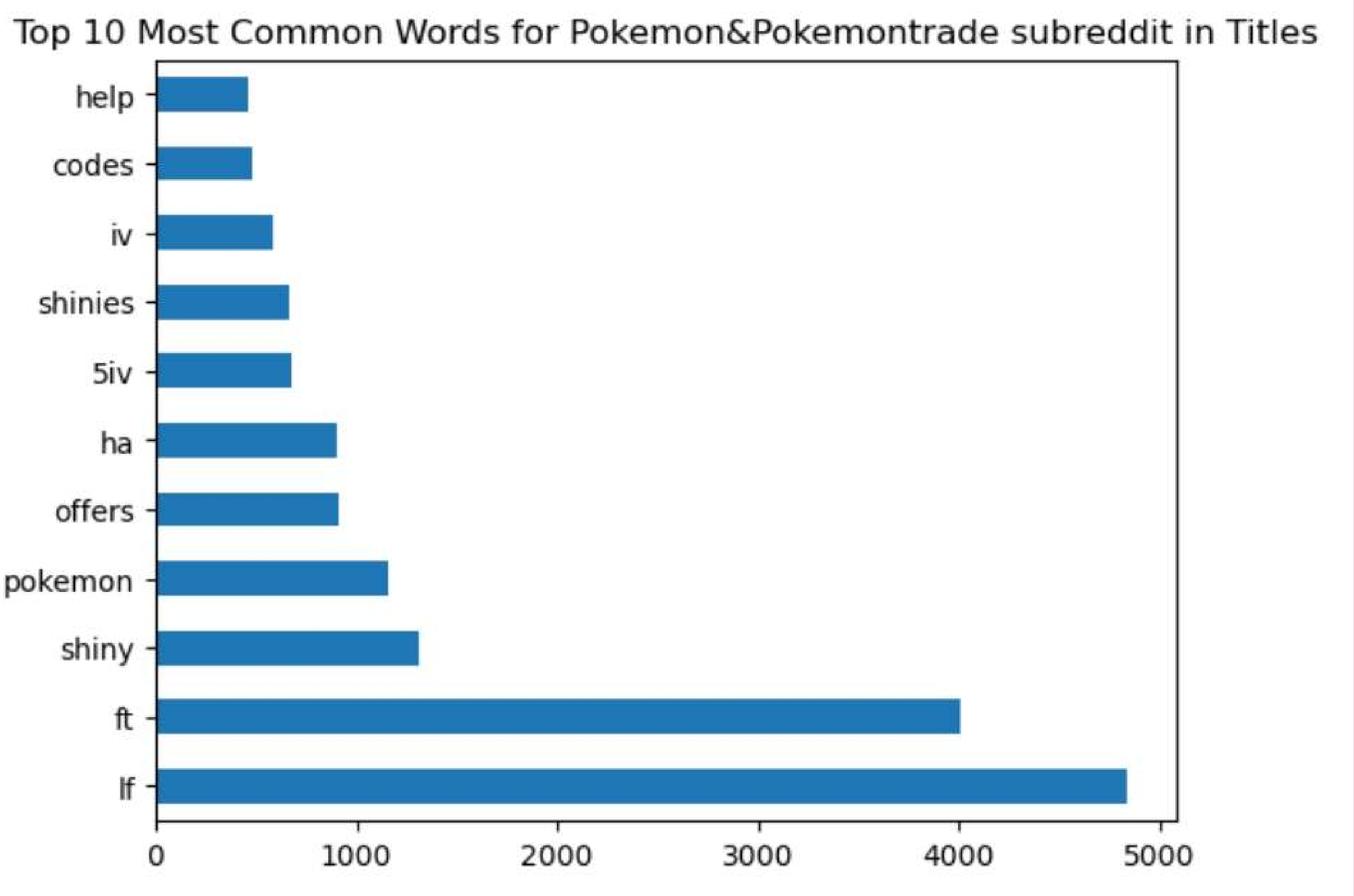
Polarity Score

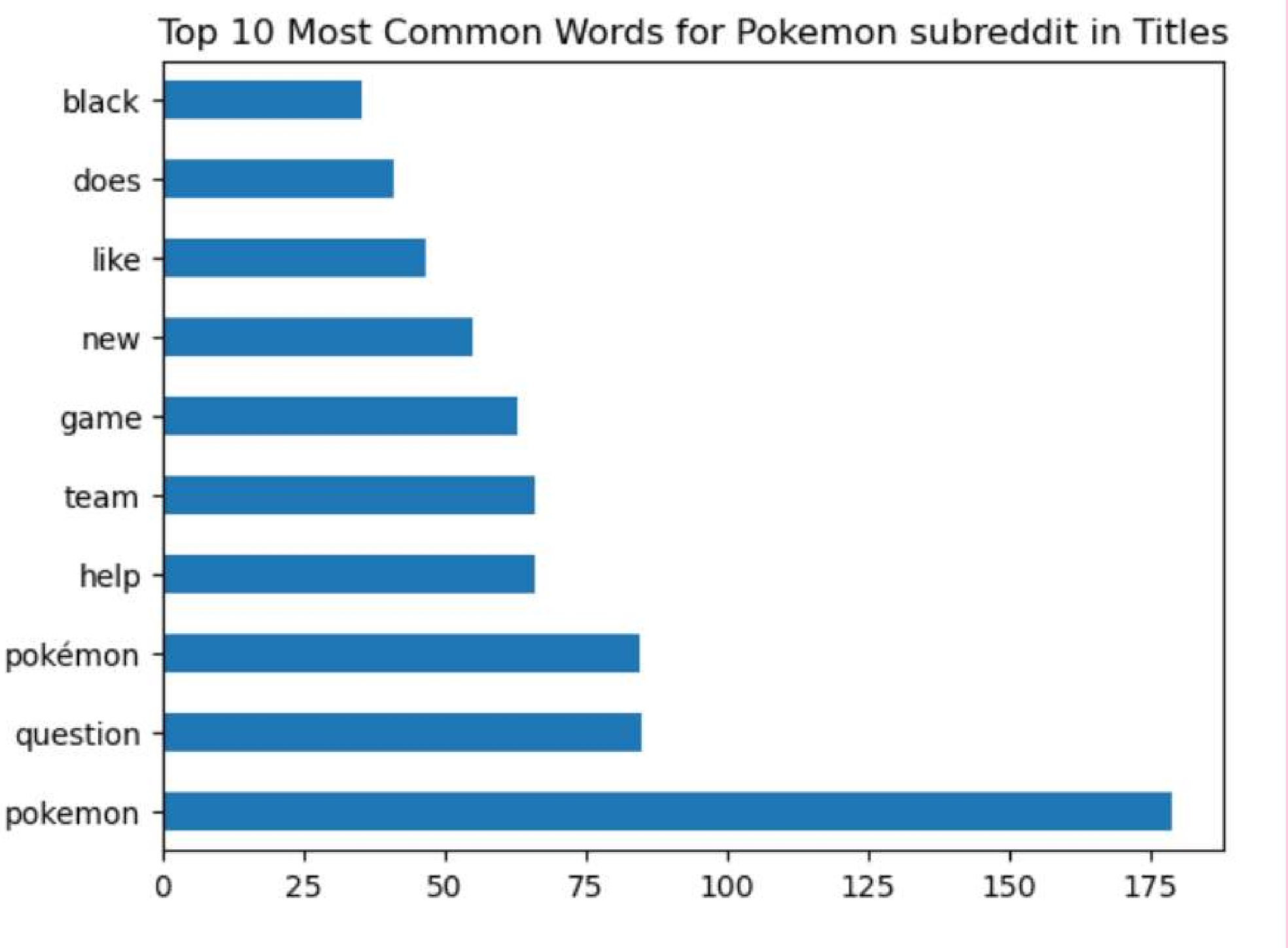
PokemonTrade (Title) + (Selftext) Polarity Score

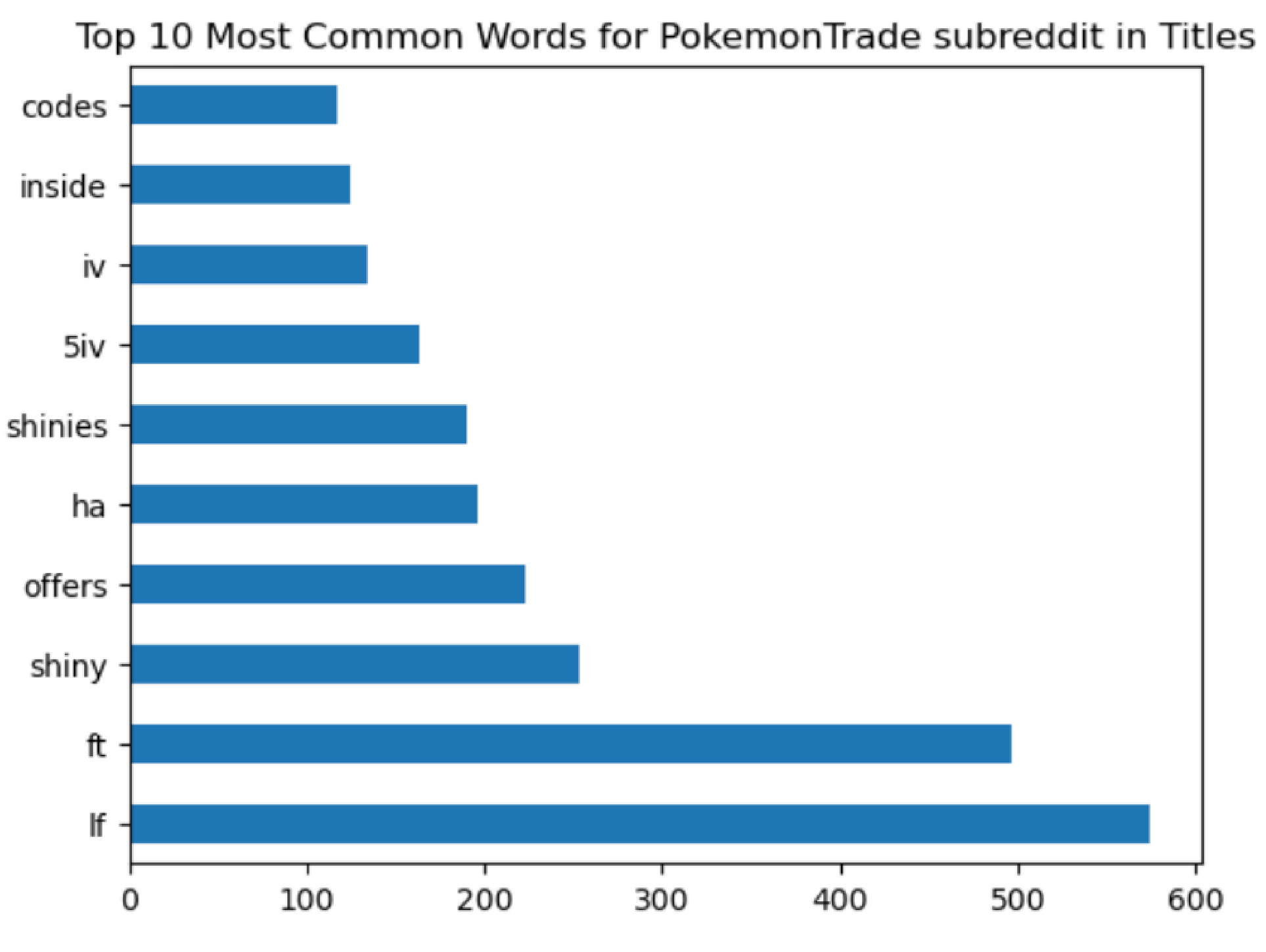
```
: analyzer.polarity_scores(str(poke_trade_text['title']))
: {'neg': 0.0, 'neu': 0.898, 'pos': 0.102, 'compound': 0.8481}

: analyzer.polarity_scores(str(poke_trade_text['selftext']))
: {'neg': 0.014, 'neu': 0.881, 'pos': 0.105, 'compound': 0.8903}
```

Pokemon(Title) + (Selftext) Polarity Score

```
: analyzer.polarity_scores(str(poke_text['title']))
: {'neg': 0.054, 'neu': 0.906, 'pos': 0.041, 'compound': 0.1205}

: analyzer.polarity_scores(str(poke_text['selftext']))
: {'neg': 0.056, 'neu': 0.906, 'pos': 0.039, 'compound': -0.4546}
```


Top 10 Most Common Words for Pokemon&Pokemontrade subreddit in Titles
help
codes
iv
shinies
5iv
ha
offers
pokemon
shiny
ft
lf
0
1000
2000
3000
4000
5000

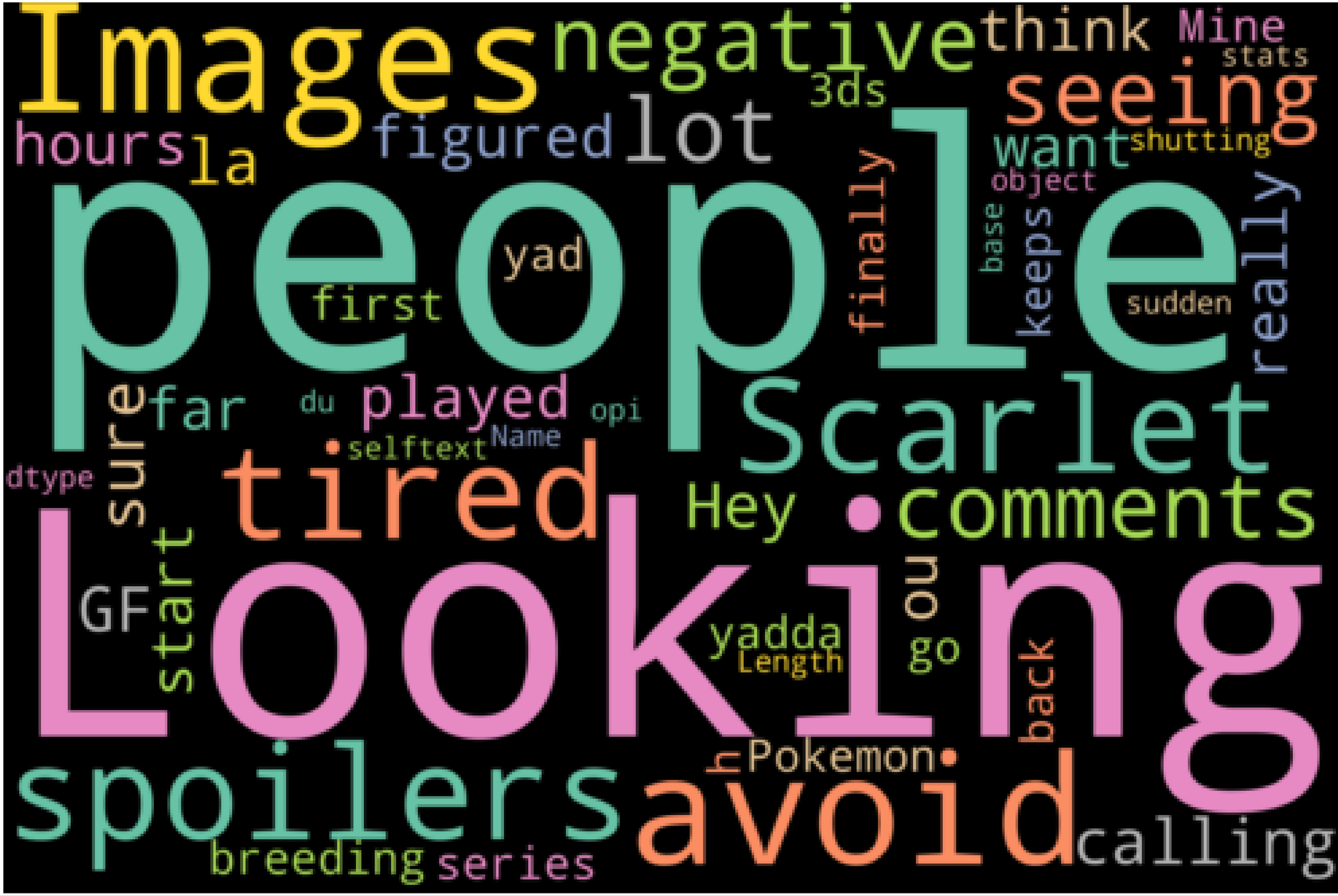

Top 10 Most Common Words for Pokemon subreddit in Titles
black
does
like
new
game
team
help
pokémon
question
pokemon
0
25
50
75
100
125
150
175


Top 10 Most Common Words for PokemonTrade subreddit in Titles
codes
inside
iv
5iv
shinies
ha
offers
shiny
ft
lf
0
100
200
300
400
500
600

Most frequent words in 'pokemon' subreddit

Most frequent words in 'pokemon' subreddit title

Most frequent words in 'pokemon' subreddit selftext



Most frequent words in 'pokemontrade' subreddit title

Most frequent words in 'pokemontrade' subreddit selftext

Conclusion / Recommendation

Conclusion / Recommendation: Differentiating Subreddits using Keywords

For 'pokemon' subreddit:

Keywords are more likely to be generic. Usually it involves some 'conversational starters', 'greetings' as well as some 'personal opinions' and 'general enquiries' on their favorite pokemon or latest generations.

'people', 'hello', 'Scarlett, 'friend', 'pokemon', 'Violet', 'team', 'favorite', 'gen', 'thought', 'like', 'help', 'new', 'question'

For'pokemontrades' subreddit:

Keywords tend to be more specific. Usually it involves pokemon's type/personaility/special ability and particular move sets, and whether they are shiny pokemon or not:

'egg', 'move', 'shiny', 'adament', 'adament', 'timid', 'levitate', 'mirror coat', 'events', 'trading', 'dream ball', 'beast ball', 'looking, 'offers', 'If', 'evolve', '6iv', 'egg moves' 'LF' (Looking For), 'FT' (For Trade) (The most crucial differentiators)

Conclusion / Recommendation:

Users' Major Concerns/ Topics in 'pokemontrade' and 'pokemon'

Pokemontrade subreddit users

- Own mastery of the Pokedex
- Own pursuits of shiny collections
- Exchanging for mutual benefits

Heated discussion topics:

- Rare shiny pokemons
- High IV pokemons
- Special personality traits
- Rare abilities / move sets

Pokemon subreddit users

- Extending their game network by casual greeting
- Exchanging friend codes; opinions about different generations and; new game features such as 'Wonder Trade'.

Heated discussion topics:

- Violet and Scarlett
- Exclusive legendary/special pokemons
- New game features
- General Q&A
- Opinion seeking

Conclusion / Recommendation:
Sentiment Scores in 'pokemontrade' and 'pokemon'

Pokemontrade subreddit:

- tends to indicate higher satisfactory rate and positive emotions.

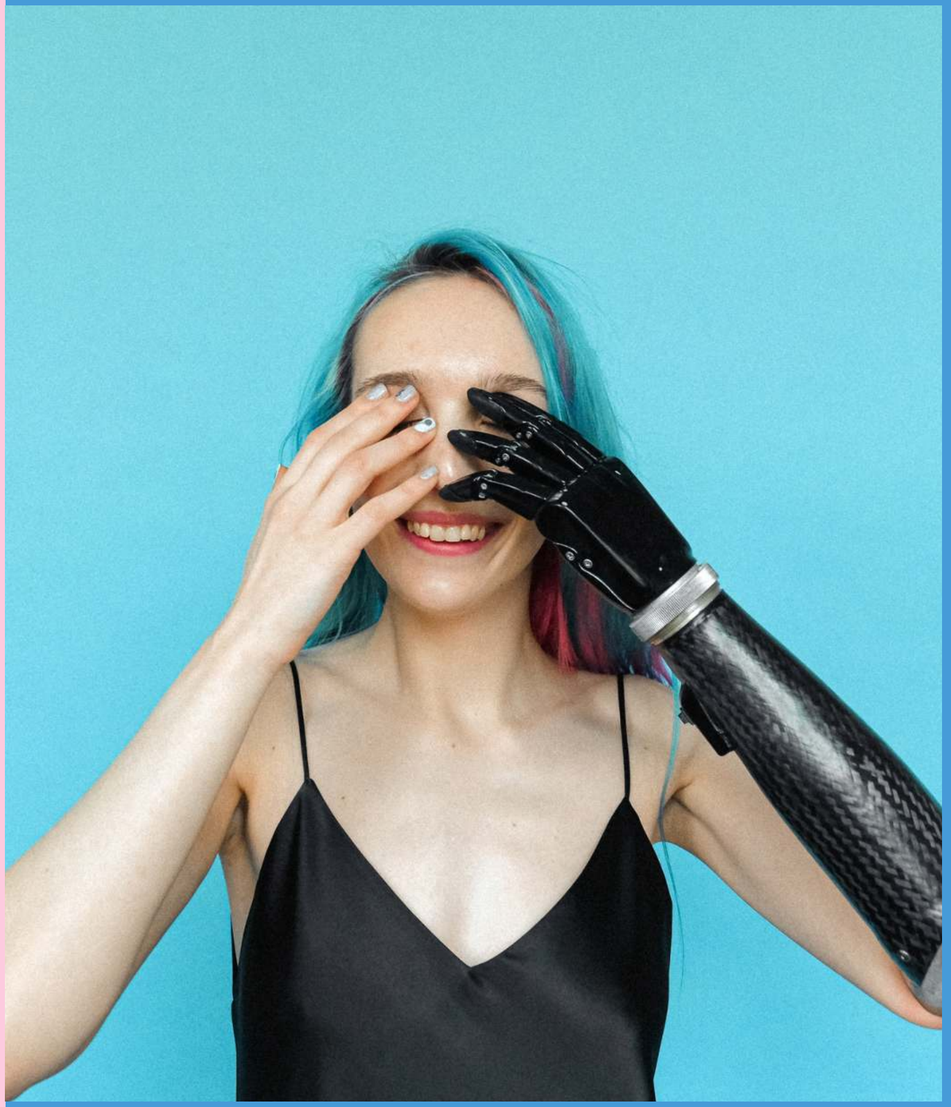
- Pokemontrade seems to be more of a vibrant, energetic, wishful place where users can throw their wishes out and have them fulfilled accordingly.

Pokemon subreddit:

- serves as a more generic discussion forum for users to express their opinions and vent their dissatisfaction.

- For example, the recent Scarlett/ Violet series was critcized negatively due to its awful technical aspects, where the game visual elements were often misplaced and dislocated. Some users are skeptical for its worth, saying it does not match with a well-known franchise's quality.

Recommendations



1. Pokemon franchise marketers can contact the users with the most subscribers to promote their latest products

2. Marketers can use the highlighted keywords to target on their customers base more accurately

3. Instead of issuing questionnaires, stakeholders of Pokemon franchise should take the comments from the pokemon subreddit into consideration when thinking of new games features or improvements to make

Predictive Modeling

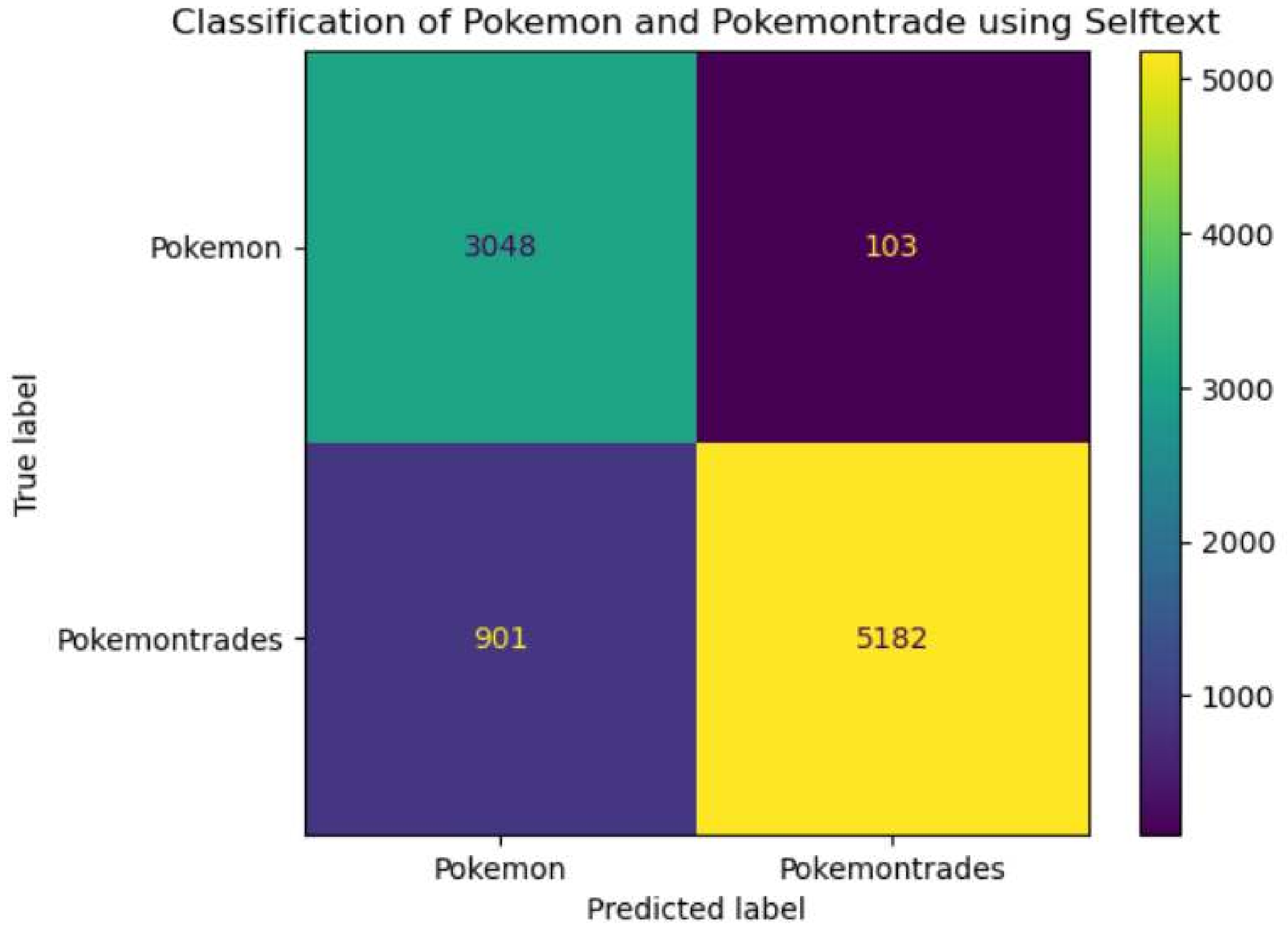
Option 1: Pipeline: CountVectorizer + MultinomialNB
Option 2: Pipeline: CountVectorizer + Term frequency Inverse document frequency (TFIDF) + MultinomialNB

What's inside the magic?

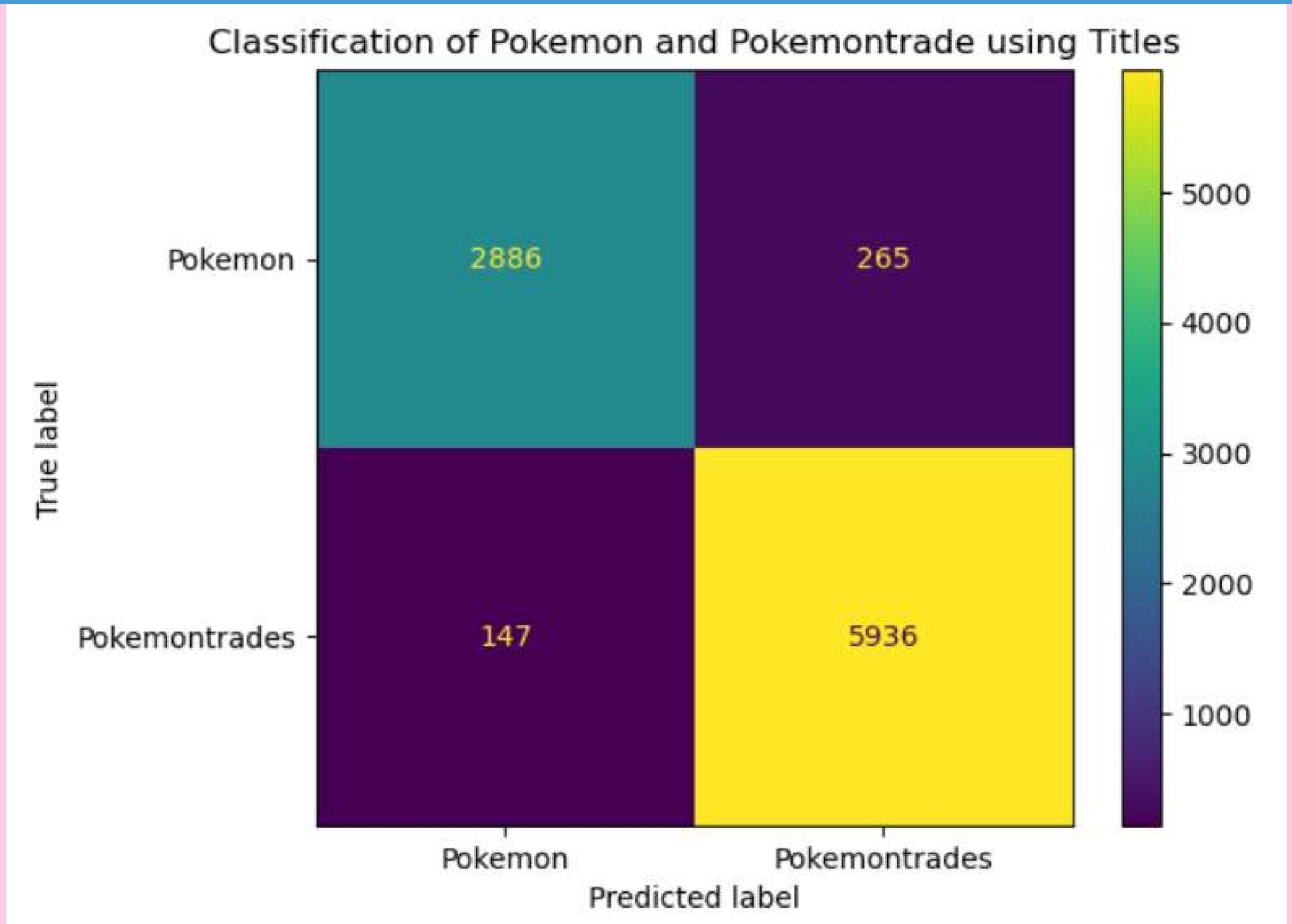
For predictive modeling, this project aims to compare the accuracy of a pipeline that encompass CountVectorizer and MultinomialNB, versus one that encompass CountVectorizer and TfidfVectorizer.

What CountVectorizer does is to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. This provides a foundation to apply both multinomial Naive Bayes classifier and Term frequency Inverse document frequency (TFIDF) at a later stage.

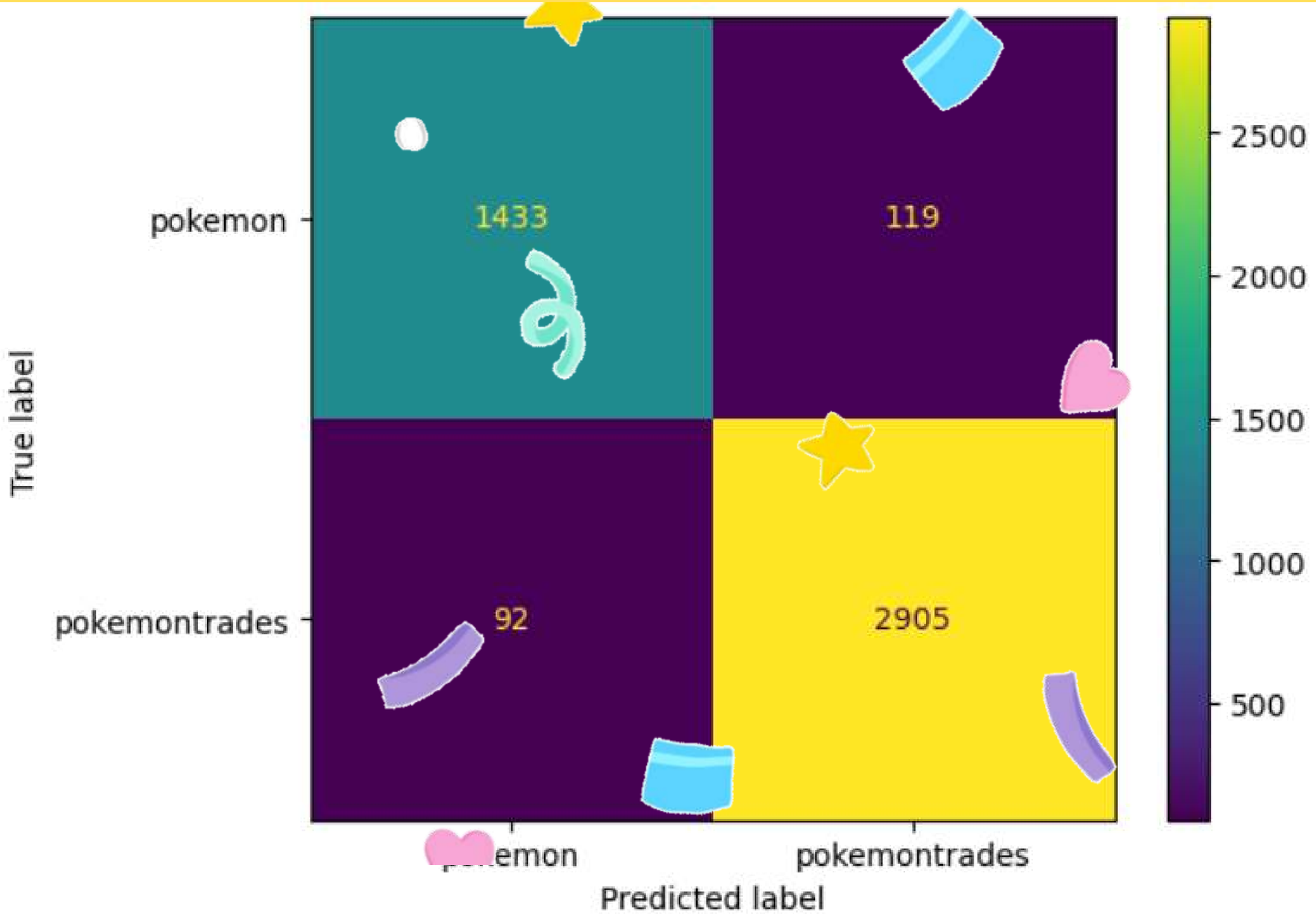
CountVectorizer+ MultinomialNB using *selftext*



CountVectorizer+ MultinomialNB using *title*



TfidfVectorizer + MultinomialNB using *title*



TfidfVectorizer + MultinomialNB using *title*

	precision	recall	f1-score	support
pokemon	0.94	0.92	0.93	1552
pokemontrades	0.96	0.97	0.96	2997
accuracy			0.95	4549
macro avg	0.95	0.95	0.95	4549
weighted avg	0.95	0.95	0.95	4549

Conclusion / Recommendation

Model-wise:
Term frequency Inverse document frequency (TFIDF) crowns a 95% accuracy score testing score, which apparently does a slightly better job on classifying the reddit posts by just MultinomialNB, which has 94% testing score.

The reason why it performs slightly better is because TFIDF converts text documents into vectors based on the relevancy of the word, it takes natural human communication context into account. Comparing to a sole MultinomialNB, which only calculates the probability of each tag for a given sample, it didn't consider any contexts nor considering other features. While probability is its sole priority, it may or may not resemble entirely what a natural context is like for communicating online.