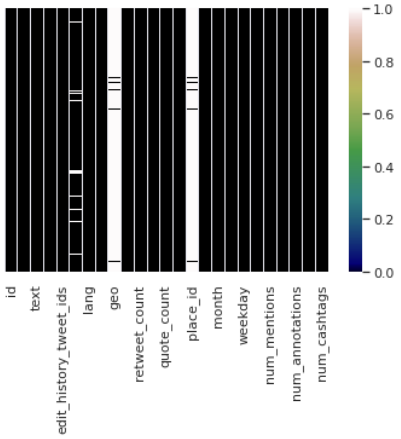
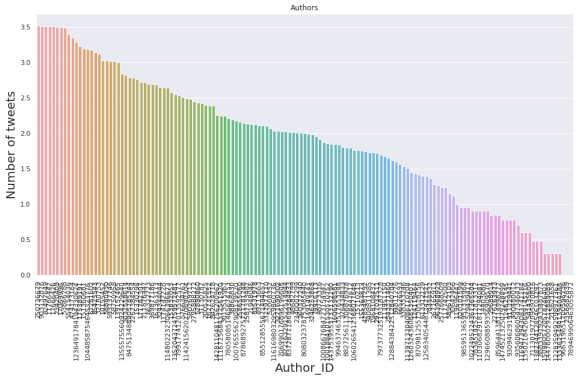
Data Analysis

Exploratory Data Analysis [Colab Notebook]

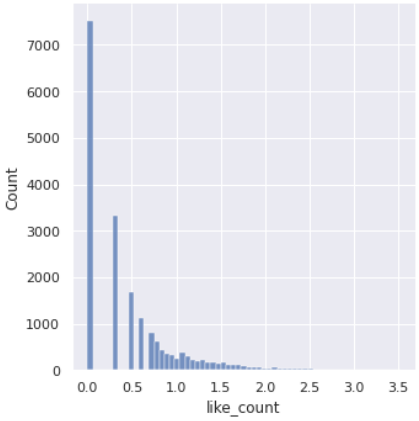
Missing Values



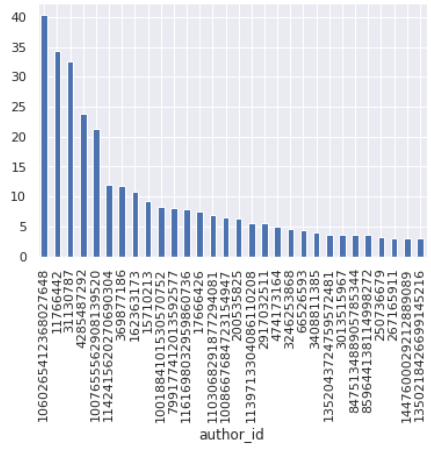
Log of #tweets published by authors



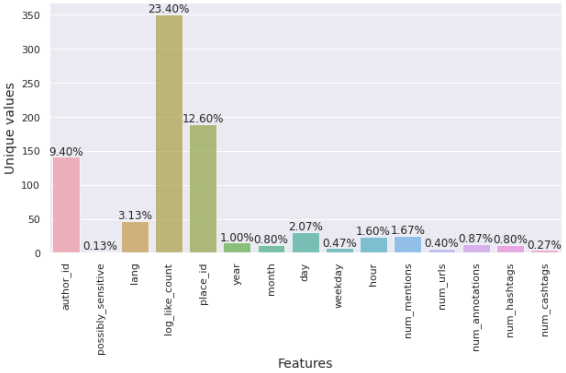
Distribution of Log of like counts



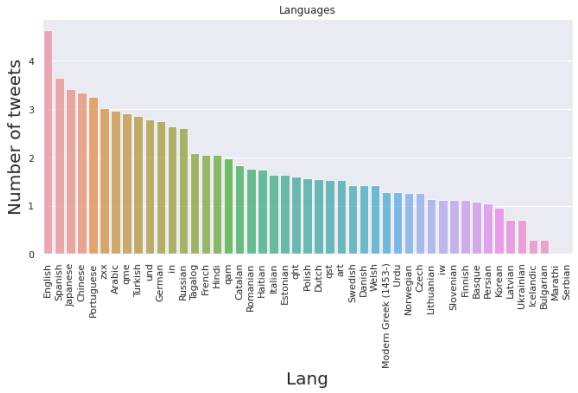
Top 30 authors receiving most likes



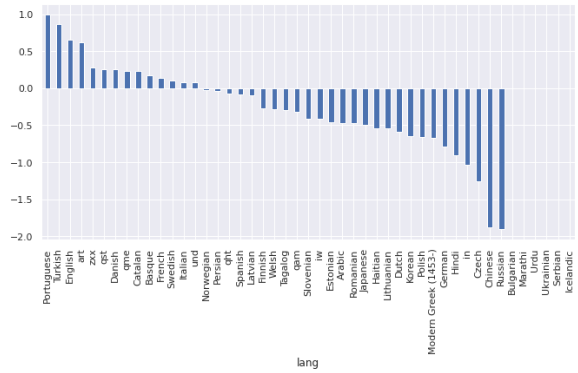
Unique Values



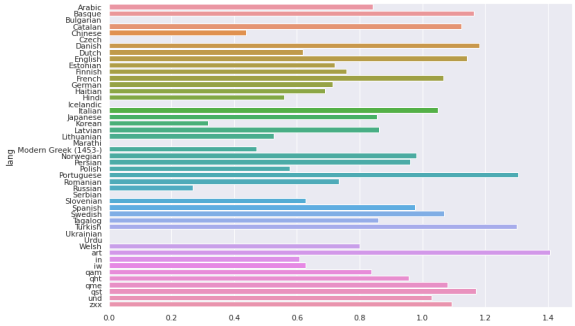
Log of #tweets by language



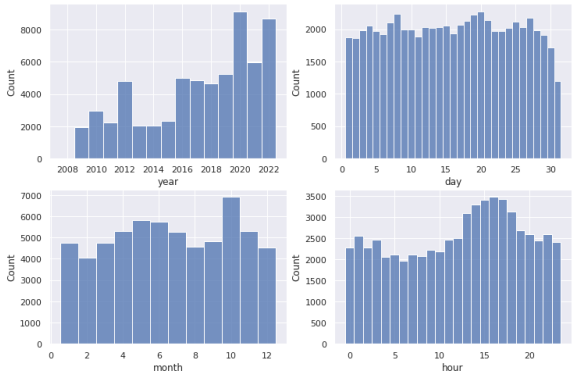
Log of average like counts by language



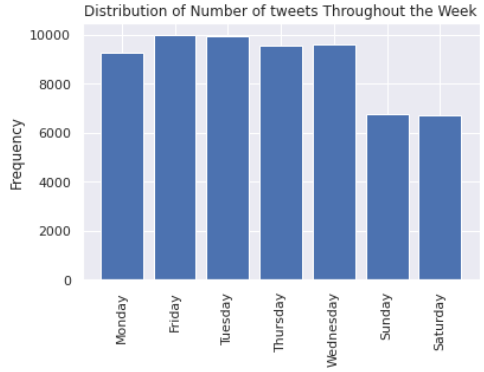
Log #likes / Log #tweets by language



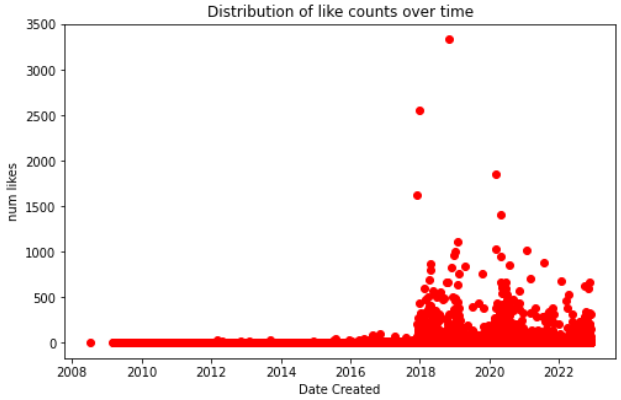
Distribution of #tweets by time



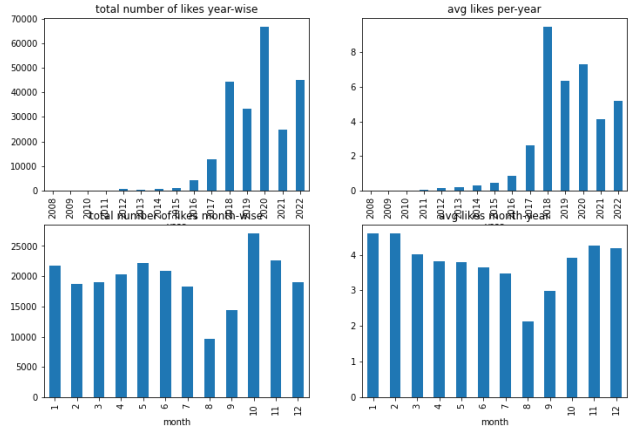
Distribution of #tweets by days in week



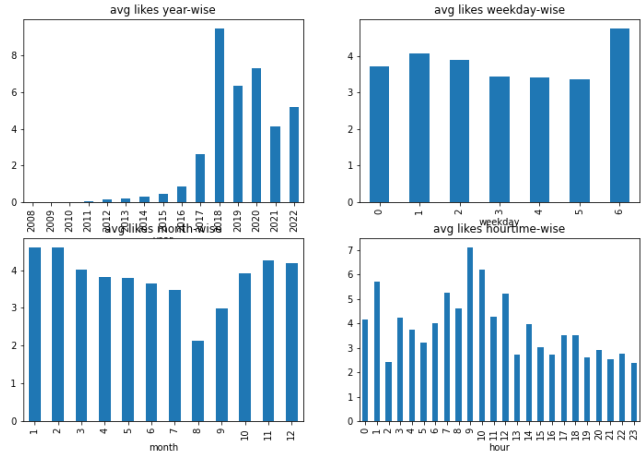
Distribution of like counts over time



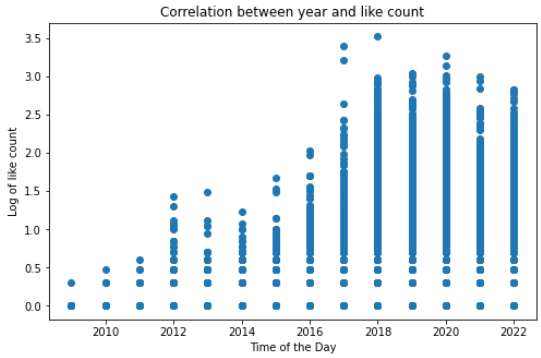
Distribution of average like counts over time



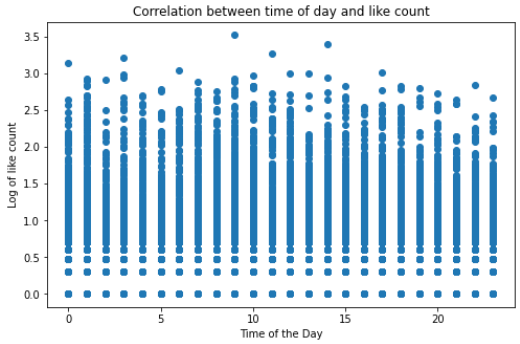
Distribution of average likes over time



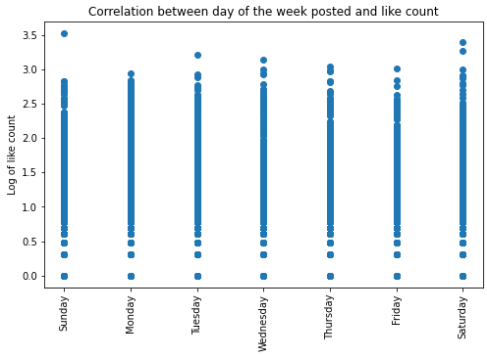
Correlation b/w year and like count



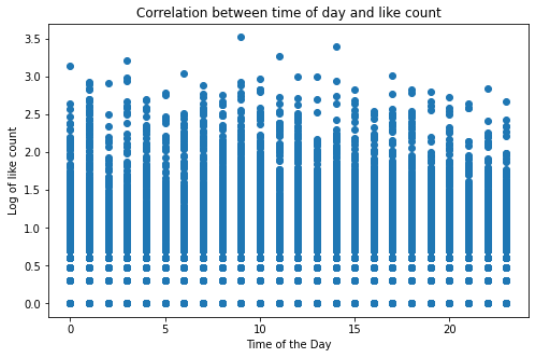
Correlation b/w time of day and like count



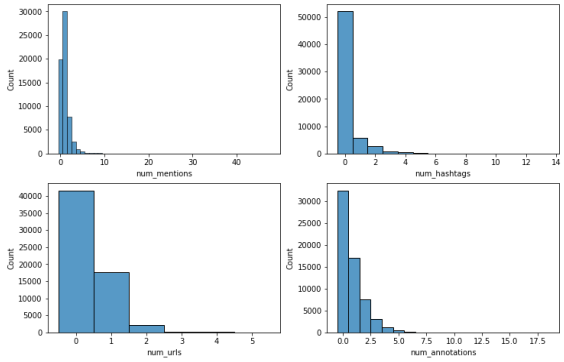
Correlation b/w weekday and like count



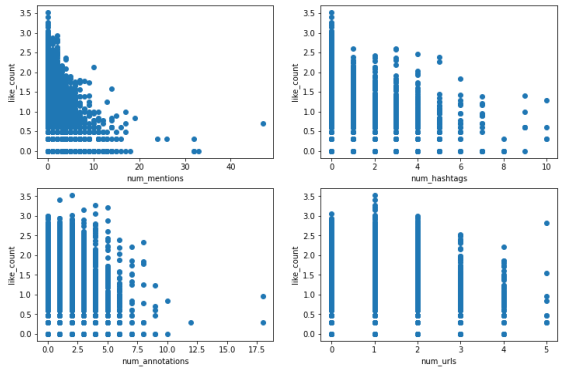
Correlation b/w hour and like count



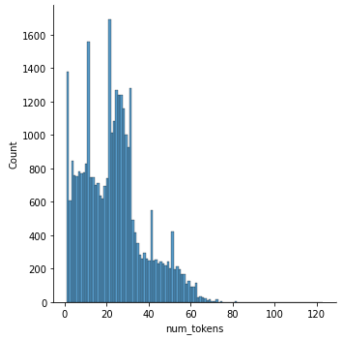
Distribution of mentions, hashtags, URLs, annotations



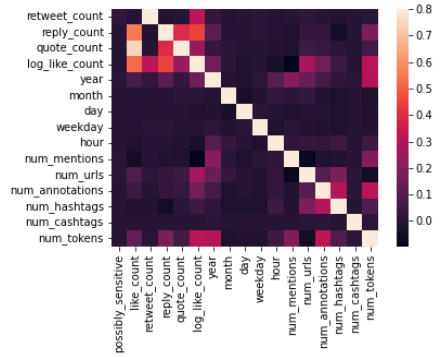
Correlation of mentions, hashtags, URLs and annotations with log of like count



Distribution of #tokens



Feature correlation heatmap



EDA findings

- Location data is mostly not available; cannot use it as a feature
- There are 141 unique authors in the dataset; including author IDs in the model input can help inform predictions as some authors may be more influential and receive more likes
- The like count distribution is skewed; most of the tweets receive 0 likes; very few tweets receive >100 likes
- Out of 141 total users, only 8 users contribute more than 5% of total tweets
- Only few influential users have large number of likes; Only few users tweet a lot
- There seems to be no correlation between number of tweets and number of likes one gets
- Can construct a popularity metric of a user to inform the probability of getting more likes (this popularity metric will not be included as a feature, but can be used for importance sampling - giving more weightage to influential users); this may depend on:
    - avg number of likes per tweet for the user - higher the better
    - max like count of the user - higher the better
    - more total tweets may increase the probability of getting likes but not so much
    - tweet topics - popular tweet topics get higher likes
- Most of the tweets are in English language; Spanish, Japanese, Chinese, Portuguese also constitute majority tweets
- Average like count for Portuguese, Turkish, English and 'art' are the highest while Russian, Chinese, Czech, 'in' have low average like counts
- There are some un-identified languages in the dataset like 'qme', 'in', 'art' etc. which do not make any semantic sense as texts for such tweets usually contains only user mentions or URLs
- Almost all tweets before 2018 get very few likes; tweets and like counts in 2018-2022 period have better distribution for modeling
- Tweets posted in the morning before noon (9am-12pm) and around midnight (12am-1am) seems to get more likes on average
- Tweets at the end and start of the year (Nov-Feb) get more avg likes; it may be the time when users post about their published papers and AI/ML concepts as some conferences happen during this time
- Tweets with more number of user mentions or hashtags usually get less likes
- There seems to be no correlation of number of user mentions, URIs, hashtags or annotations with high like counts

Data Pre-Processing

- Tweets before 2018 and not of English language are filtered out
- Tweets are pre-processed as described in Nguyen et al., 2020
    - tokenized using “TweetTokenizer” from NLTK
    - translate emotion icons into text strings using emoji package
    - normalized by converting user mentions and links into special tokens @USER and HTTPURL, respectively
- Data is split temporally into - training (80%), validation(10%) and testing (10%)

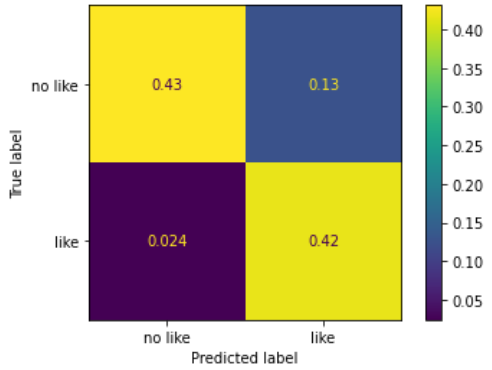
Model Evaluation [Colab Notebook]

Performance

Model	Accuracy	F1 for class=1	Precision for class=1	Recall for class=1
Roberta roberta-base	82.90	0.83	0.75	0.94
DistilBERT distilbert-base-cased	83.83	0.85	0.75	**0.98**
BERTweet **vinai/bertweet-large**	**86.73**	**0.86**	**0.78**	0.94

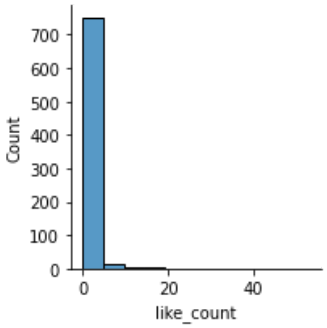
Error Analysis

Confusion Matrix

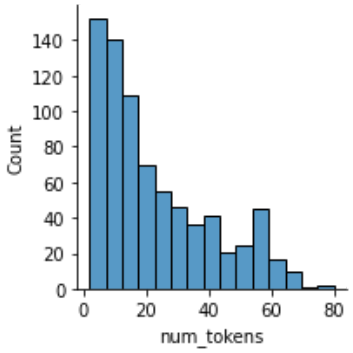


- The model has more false positives as compared to false negatives.
- Overall, longer tweets are classified correctly and shorter tweets are more susceptible to be misclassified. This makes sense as shorter tweets have less context and may contain only URLs, user mentions, emojis etc.
- Tweets lacking context or ambiguous tweets are also more likely to be misclassified.
- Tweets with a very high number of likes are generally correctly classified.

Distribution of like counts for mis-classifications



Distribution of number of tokens in mis-classified tweets



Reasons for misclassification

- Some tweets are duplicate and have different numbers of likes as they come from different users. Example tweets -
    - *STEME 's 4th issue online & openly accessible . Welcoming new submissions ! Embedding opportunities for participation and feedback in large mathematics lectures via audience response systems (HTTPURL) #student #group #lecture #steme #stemeducation HTTPURL*
    - *@USER Thanks !*
- Ambiguity - Short sentences with lack of context
    - *@USER You got it bud!*
    - *@USER what should we do?*
    - *@USER @USER Thought that one might appeal to you!*
    - *@USER Will you burn tokens?*
    - *@USER More seriously, it's a good idea, but I can't think when I 'd have the time.*
- Some need more context to fully understand e.g URLs
    - *#cvpr2022 @USER Impressive Numbers!!! HTTPURL*
    - *@USER If it helps, we took an attempt (by citing the tweet by John Harness) in this paper HTTPURL*
    - *@USER @USER @USER As far as I can tell, the sum totality of human knowledge regarding optimal step size selection for full-batch gradient descent on neural nets is : HTTPURL HTTPURL*
    - *Musk watching advertisers abandon ship HTTPURL HTTPURL*
- Big sentences with math explanation (usually predicted to be liked)
    - *@USER @USER Let's consider a quadratic f (x) = Q x + b. Let g be the current gradient. Exact line search uses the step size g'g / g'Q g (Nocedal and Wright eq 3.26) which is always at least 1 / lambda 1 (Q). Meanwhile the largest possible step size for constant-step-size GD is 2 / lambda 1 (Q)*
    - *Let's say we have a quadratic f (x, y) = x ^ 2 + 1000 y ^ 2. The best step size strategy will depend on the initial position (x0, y0). For which initial positions will line search fail catastrophically?*
- Dependence on number of user mentions; more number of user mentiones (>4) are predicted to be not liked
    - *@USER @USER @USER @USER @USER Thanks Preethi :)* : Not Liked
    - *@USER @USER @USER Thanks Preethi :)* : Liked
    - *@USER @USER @USER @USER @USER Wow what art* : Not Liked
    - *@USER @USER @USER @USER Wow what art* : Liked
- Semantic miscontrue (sarcasm, jokes etc)
    - *@USER " That's why they call me the Joker Batsam, because you're joking with me whether you know it or not! "*
    - *Why Jimmy Kimmel, @USER? Couldn't find a big enough jar of mayo willing to present?*
- Code-switching/ mixing
    - *@USER Yeh to hona he tha.*
- Presence of certain emojis/ hashtags/tokens influence predictions.

    - *:red_heart:, #stemeducation, haha* always Liked
- False Positives
    - Tweets containing *'thank'* have less confidence for smaller and generic tweets; but higher for longer tweets
    - Wordle posts - always predicted positive with high confidence

Other findings

- False negatives
    - Lack of context/Ambiguity
        - *@USER Do you think you could include my info? University of Ottawa Augusto Gerolin Mathematics, Quantum Chemistry and Machine Learning HTTPURL Funding : Yes Ottawa Canada Academic agerolin@uottawa.ca*
    - Small tweets; number of tokens <10
- Tweets where users congratulate usually receive likes if their length is small; longer congrats tweets don't get likes
- RT with thanks do not get likes and correctly predicted
- For high number of like counts (>20), the model predictions are generally correct

Ideas to improve performance

- remove duplicate tweets
- remove short tweets e.g. tweets with less than 10 tokens
- Include metadata like *author_id*, *num_quotes*, *num_mentions* in the text input
- define class more aggressively: Class-1 if *like_count* > 10 else Class-0
- class weighting to account for class imbalance
- importance sampling in accordance to number of likes i.e. tweets with higher *like_count* have more weight or tweets by influential authors have more weight
- tune probability threshold for assigning labels instead of argmax
- increase number of epochs for training

Interpretability Analysis

Interpretability analysis was done using Captum to generate feature importance using Layer Integrated Gradients. This directory contains images of 20 examples with comments for which the analysis was done. A few features/examples on which the model seems to be relying to makes predictions are -

- Wordle tweets

- Presence of URLs
- User talking about personal experience
- Positive words/emotions of the user
- Certain positive actions (eg reading an ML paper) taken by the user
- Posts that give out information other users may be interested in and thus engage more
- Users talking about 'gradient'
- Heavy reliance on the token 'RT' to predict retweets are always liked