Understanding and Predicting Post Popularity in Reddit's Political Communities

Master Thesis

by

Leon Holz
3199118

submitted to
Rheinische Friedrich-Wilhelms-Universität Bonn
TODO

in degree course
Computer Science (M.Sc.)

Supervisor: Rafet Sifa
University of Bonn

Bonn, February 9, 2026

Abstract

1 Introduction

Reddit is a social media platform where users engage in discussions across various subreddits—dedicated communities focused on specific topics. Posts within a subreddit can receive upvotes and downvotes, determining their visibility. Meanwhile, karma serves as a numerical representation of a user's contributions through posts and comments, functioning similarly to likes on other social media platforms.

Reddit has over 365 million active weekly users [Inc], with about 70% of users in the U.S. reportedly using it as their primary news source [Cen]. Politicians, including Barack Obama, Donald Trump, and Bernie Sanders, have recognized Reddit's influence by hosting *Ask Me Anything* (AMA) sessions [DHM17].

The ability to predict the popularity of online content has significant implications for marketing, content optimization, and social media engagement strategies. Companies invest heavily in predictive systems to maximize visibility and audience engagement [DHM17]. However, prior studies on online content popularity often focus on other platforms, such as Twitter and Facebook, or rely on limited feature sets that overlook subreddit-specific dynamics.

Moreover, predictive models for content popularity raise ethical concerns. The ability to manipulate a post's visibility could be exploited for misinformation campaigns, artificially amplifying certain narratives under the guise of community approval [DHM17]. Understanding the underlying factors that drive Reddit post popularity is therefore essential, both for refining predictive models and for mitigating potential risks.

This study leverages a neural network model using text embeddings and numerical features, such as author karma and account age, to predict post popularity. We analyze how subreddit-specific characteristics impact predictive performance, revealing key challenges in generalizing across different communities. Our findings contribute to the broader understanding of engagement patterns on Reddit and highlight potential strategies for improving predictive accuracy while addressing biases in content visibility.

2 Basics

2.1 Reddit

generell info regarding reddit

model relevant inputs posts, title, content, score ...

2.2 Behavioral analytics

2.3 Transformer

- history - properties

2.3.1 LLMs

2.4 XGboost/Gradient Boosting

kurz gradient boosting erklären

3 Related Work

Several studies have examined the prediction of online content popularity. Segall et al. [SZ12] utilized features such as the author's identity, post creation time, and selftext content to predict post popularity.

Other research has focused on early engagement metrics. Andrei Terentiev et al. [TT14] predicted post popularity using characteristics of the first 10 comments. Similarly, Fredrik Wigsnes et al. [Wig19] analyzed posts' statistics one hour after publication to make predictions. However, these approaches require a waiting period, limiting their real-time applicability.

Deaton et al. [DHM17] extended the analysis to comment popularity prediction, emphasizing the relevance of understanding how individual contributions perform within discussions.

In this study, we expand on previous works by incorporating additional features related to the author, such as karma and account age, and examining subreddit-specific predictive accuracy.

The sociomateriality of rating and ranking devices on social media: A case study of Reddit's voting practices

Popularity dynamics and intrinsic quality in reddit and hacker news

Hessel and Lee (2019) investigate the early prediction of controversial Reddit posts, defining controversy as community-specific disagreement reflected in mixed upvote and downvote behavior. Using data from multiple subreddits, they show that incorporating features from early user interactions—particularly the textual content and structural properties of initial comment trees—significantly improves prediction performance compared to post-time text and metadata alone. Importantly, they demonstrate that controversy prediction is distinct from popularity predic-

tion: models trained to forecast engagement volume or attention do not transfer well to identifying controversial content, indicating that these outcomes capture fundamentally different dynamics. While their work focuses on controversy rather than popularity, it establishes that early interaction signals provide information beyond post text and that predictive models must be carefully aligned with the specific downstream outcome of interest, a distinction that directly informs pipelines aimed at generating and evaluating posts based on predicted popularity rather than divisiveness [HL19]

4 Methodology

TODO kleine einleitung

4.1 Dataset

Firstly, a large dataset of Reddit posts is required for our study. Initially, we planned to collect data using the Reddit API through the PRAW library [], as it allows us to select specific subreddits freely. However, the API imposes a limitation of retrieving only 1,000 posts per request. Reddit offers different sorting options, such as hot, new, rising, top, and controversial, which can be combined to increase the diversity of the dataset. In theory, this approach could yield up to 5,000 unique posts per subreddit. However, in practice, many posts appear in multiple categories, significantly reducing the actual number of unique posts collected—typically to around 2,000 per subreddit.

volume far too small for training model or making claims using this small subset for millions of posts, necessitating an alternative approach to data collection.

instead of using the api, we used datasets provided by pushshift [Bau+20]

Pushshift is a large-scale data collection and archiving project that provides historical and near real-time access to social media data, most notably from Reddit. It was developed and maintained by Jason Baumgartner with the goal of facilitating social science research, data mining, and natural language processing (NLP) applications.

Pushshift continuously ingested Reddit data via the platform's public API and other data sources, archiving submissions and comments in a structured and queryable format. The data were made publicly accessible through two main interfaces:

The Pushshift API, which allowed users to query Reddit submissions and comments by a wide range of metadata (e.g., subreddit, author, score, creation time, etc.), often returning results faster and with fewer limitations than the official Reddit API.

Public data dumps, periodically released as compressed JSON files, providing bulk access to large portions of Reddit's history for offline processing.

Because of its extensive temporal coverage, Pushshift became a foundational resource in computational social science, linguistics, and political discourse analysis. Researchers used it to study online behavior, polarization, community dynamics, and the diffusion of information across subreddits and time.

However, as of 2023, Pushshift's public API was restricted following policy changes and concerns from Reddit regarding data use and privacy. While historical datasets remain accessible through archives and mirrors, real-time access has become limited. Consequently, many researchers now rely on the official Reddit API (via tools such as PRAW) or third-party mirrors of Pushshift's data for updated collections.

therefore only data up to 2023 is available. also must be mindful of potential biases, such as incomplete coverage, post-deletion timing discrepancies, or missing metadata in later versions of the dataset.

we found that alot of the time the scores noted in the pushshift datasets were slightly off compared to results from the reddit api via praw. this could be due to upvote fuzzing or additional user engagement after the post and its data was saved to pushshift upvote fuzzing adds a random constant to upvotes and downvotes in order to "combat spam and manipulation". this behavior was supposedly stopped 2018 [Sto15]

also there is additional data that was required for analysis and later model training, so we used the reddit api via praw to gather additonal data for each post and all authors contained in it

also information for some posts could not be retrieved via the reddit api because they haeve since been removed, either by author, subreddit moderators or reddit admins, or the Id might not be correct in pushshift therefore those posts were filtered out additionally there were other posts where the user or the post have been banned or deleted, causing the content to not be available anymore. those posts have been filtered too -> possible introduction of bias against controversial or unpopular content Posts where author data could not be retrieved due to API limitations or other reasons were also excluded.

4.2 Subreddit used

for subreddits, since we wanted to see differences and similiaries in the different political landscapes, we chose to pick 5 subreddits of each leaning, meaning left, right, and neutral. Additonally, to have a point of comparison, we chose to also include 5 subreddits that are not inherently political, so that we can see how posts from these subreddits are different to the ones from political subreddits

generel metrics included the size of the sub, what post types they usually use for our analysis, we tried to get subreddit with little images and mostly selftext and link posts, we also tried to find subreddits that were "large enough" after filtering posts under a minimum of either a score of 3 or amount of comments equal to 3

we chose to use keep analysis to 20 subreddits with medium size because of resources used for the encoded datasets. As subreddit as large as relationship_advice for example would take up as much as 20GB using a fairly small dimensional encoder of 1024 dimensions. Ofcourse such size

comes with longer encoding and model training time. WIth our approach each subreddit takes up between 0.15GB and 4GB after being encoded and enables us to test more configurations in the same time. additionally we assumed it was likely that we would try different encoders or prompts which would add even more storage required

to find out which direction a subreddit leans, we decided to use a mixture of the subreddits own descriptions and the work of Waller et al. [WA21], which aligned for the subreddits we analyzed, if the description allowed for clear alignment. in the following x you can see the subreedits chosen, their community title and description and comments on why they were chosen

r/democrats - Current Community Title: Democrats: Building a Better Future; News about Democrats and current events in the United States - Current Community description: The Democratic Party is building a better future for everyone and you can help. Join us today and help elect more Democrats nationwide! This sub offers news about current events in the United States. We are here to get Democrats elected up and down the ballot. - probably the most well known subreddit in favor of the democratic party in the US, categorized by [WA21] as left and officially supports the view of the democratic party.

r/Liberals - Current Community Title: The Liberal Subreddit: News about Liberals and Democrats - Current Community description: A Liberal Subreddit - smaller then democrats and not included in the work of [WA21], but left leaning according to its discription

r/anarchism - Current Community Title: Anarchism: Beneath the pavement, the beach - Current Community description: Anarchism is a social movement that seeks liberation from oppressive systems of control including but not limited to the state, governmentalism, capitalism, racism, sexism, ableism, speciesism, and religion. Anarchists advocate a self-managed, classless, stateless society without borders, bosses, or rulers where everyone takes collective responsibility for the health and prosperity of themselves and the environment. - not included in the work of [WA21], more extreme left leaning subreddit compared to the previous two and not associated with a party

r/racism - Current Community Title: dismantle white supremacy - Current Community description: Reddit's anti-racism community, a safe(r) space for People of Color and their supporters, pre-screens most content for safety. All discussions are expected to be from a post-"racism 101" and postcolonial point of view. We are conscious that race intersects with sex, class, disability, age, and more, and intend this space to be safe(r) for *all* POC. This community was founded by and is actively curated by People of Color. - left leaning according to [WA21], not associated with a party

r/TrueAtheism - Current Community Title: TrueAtheism - Current Community description: A place dedicated to insightful posts and thoughtful, balanced discussion about atheism specifically and related topics concerning irreligion and religion generally. - left leaning according to [WA21], not associated with a party, while not directly about politics, religion is politcal too [Pot20][Per99] [Lan19] [Mad+22]

r/Conservative - Current Community Title: Conservative - Current Community description: https://x.com/rcondiscord https://discord.gg/conservative - right leaning according to [WA21], the counterpart to r/democrats

r/TrueChristian - Current Community Title: A subreddit for followers of Jesus Christ. - Current Community description: A subreddit for Christians of all sorts. We exist to provide a safe haven for all followers of Jesus Christ to discuss God, Jesus, the Bible, and information relative to our beliefs, and to provide non-believers a place to ask questions about Christianity as explained in the scriptures, without fear of mockery or debasement. To post suggestions or ideas for the sub, please go to /r/TrueChristianMeta. Come join us on Discord! https://discord.gg/mGCM9egt77 - right leaning according to [WA21], the counterpart to r/TrueAtheism

r/Libertarian - Current Community Title: r/Libertarian: For a Free Society - Current Community description: Welcome to /r/Libertarian, a subreddit to discuss libertarianism. We are not a generic politics sub. We are a libertarian sub, about libertarianism. We do not owe you a platform to push anti-libertarian ideologies such as socialism/communism. This sub is explicitly against Communism/Socialism as it is antithetical to libertarianism - not included in the work of [WA21], but right leaning according to their own description

r/mensrights - Current Community Title: Men's Rights :: Advocating for the social and legal equality of men and boys since 2008 - Current Community description: At the most basic level, men's rights are the legal rights that are granted to men. However, any issue that pertains to men's relationship to society is also a topic suitable for this subreddit. Men's rights are influenced by the way men are perceived by others. WARNING: Some other subs have bots that will ban you if you post or comment here. - not included in the work of [WA21], anti-feminist communities overlap significantly with alt right communites [MHW21]

r/progun - Current Community Title: Posts must be related to Firearms & Second Amendment Politics. Please engage in Civil Discussion. - Current Community description: This is a place for discussion and debate of Second Amendment related topics, with a Pro-2A emphasis. Civil debate is welcome and encouraged. Even if you're completely opposed to 2A, you're welcome to share your thoughts here, as long as you maintain civility. - right leaning according to [WA21]

r/PoliticalDiscussion

r/PoliticalDiscussion - Current Community Title: Political Discussion - Current Community description: This is a subreddit for substantive and civil discussion on political topics. If you have a political prompt for discussion, ask it here! - not included in the work of [WA21]

r/changemyview - Current Community Title: Change My View (CMV) - Current Community description: A place for people to post an opinion they accept may be flawed, in an effort to understand other perspectives on the issue. Enter with a mindset for conversation, not debate. - not included in the work of [WA21]

r/NeutralPolitics - Current Community Title: Neutral Politics: Evidence. Logic. Respect. - Current Community description: A strictly-moderated community dedicated to evidence-based discussion of political issues. - right leaning according to [WA21]

r/geopolitics - Current Community Title: Geopolitics: Geopolitical news, analysis, & discussion - Current Community description: Geopolitics is focused on the relationship between politics and territory. Through geopolitics we attempt to analyze and predict the actions and decisions of nations, or other forms of political power, by means of their geographical characteristics and

location in the world. In a broader sense, geopolitics studies the general relations between countries on a global scale. Here we analyze local events in terms of the bigger, global picture. - not included in the work of [WA21]

r/DebateReligion - Current Community Title: Discuss and debate religion - Current Community description: A place to respectfully discuss and debate religion - not included in the work of [WA21]

r/DecidingToBeBetter - Current Community Title: Deciding To Be Better - Current Community description: Ready to make a positive change in your life and leave behind habits that no longer serves you? This community is for you! Please read rules before posting. - not included in the work of [WA21]

r/truegaming - Current Community Title: For those who like talking about games as much as playing them. - Current Community description: /r/truegaming is a subreddit dedicated to meaningful, insightful, and high-quality discussion on all topics gaming. - not included in the work of [WA21]

r/Fantasy - Current Community Title: Reddit Fantasy - Current Community description: /r/Fantasy is the internet's largest discussion forum for the greater Speculative Fiction genre. We welcome respectful dialogue related to speculative fiction in literature, games, film, and the wider world. We reserve the right to remove discussion that does not fulfill the mission of /r/Fantasy. - not included in the work of [WA21]

r/Frugal - Current Community Title: Frugal Living: Waste Less, Gain More! - Current Community description: Frugality is the mental approach we each take when considering our resource allocations. It includes time, money, convenience, and many other factors. - neutral leaning according to [WA21]

r/Coffee - Current Community Title: Coffee - Current Community description: /r/Coffee is back - for now - and talking about itself, in addition to coffee. - not included in the work of [WA21]

Initially we also wanted to include infamous and controversial subreddits such as r/The_Donald, r/Mr_Trump, r/uncensorednews or r/new_right. Unfortunately, these subreddit were all banned or quarantined, meaning you can no longer gather data about them via praw. There were also a few subreddits which were too small after filtering out deleted posts and posts with low engagement

as mentioned earlier, we filtered out deleted or removed posts/users. Here we found that left, right and apolitical leaning were relatively close with 16.5 - 20% deleted posts. Neutral leaning however had more than half removed.

4.3 Explorative Analysis

in this chapter we explore the datasets, gathering general statistics in regards to the subreddit as well as a more detailed look into the dataset using sentiment analysis, tfidf, embedding analysis as well as analyzing domains used in link posts

Table 1: *Post Counts and Filtering Losses per Subreddit (Adjusted Original Size)*

Subreddit	Original	Remaining	Deleted	% Filtered
anarchism	182,082	153,447	28,635	15.73
changemyview	255,187	106,222	148,965	58.38
Coffee	208,060	191,772	16,288	7.83
conservative	938,422	772,529	165,893	17.68
DebateReligion	60,222	39,747	20,475	33.99
DecidingToBeBetter	101,165	68,721	32,444	32.06
democrats	175,929	143,813	32,116	18.26
Fantasy	173,384	139,517	33,867	19.53
Frugal	188,139	156,188	31,951	16.99
geopolitics	63,475	46,663	16,812	26.48
Liberal	67,446	51,761	15,685	23.26
Libertarian	525,563	459,502	66,061	12.57
mensrights	308,715	261,078	47,637	15.43
NeutralPolitics	21,728	6,731	14,997	69.03
PoliticalDiscussion	177,351	60,033	117,318	66.14
progun	59,998	50,600	9,398	15.66
racism	52,984	34,731	18,253	34.45
TrueAtheism	20,006	16,102	3,904	19.52
TrueChristian	100,404	68,664	31,740	31.61
truegaming	61,368	39,355	22,013	35.85

Table 2: *Post Counts and Filtering Losses by Political Leaning (Adjusted Original Size)*

Political Leaning	Original	Remaining	Deleted	% Filtered
Right	1,933,102	1,612,373	320,729	16.59
Left	498,447	399,854	98,593	19.78
Neutral	577,963	259,396	318,567	55.12
Apolitical	732,116	595,553	136,563	18.65
Overall	**3,741,628**	**2,867,176**	**874,452**	**23.38**

4.3.1

Basic Information

analyzed content types per subreddit since goal is to create posts via gen ai later and use our model to predict if they will be popular, we focused on link and selftext posts as we can see in Figure 1, most politcal subreddits have an overwhelming amount of link posts, a small amount of images and video posts, that does not exceed 25% when added together and a varying amount of selftext posts.

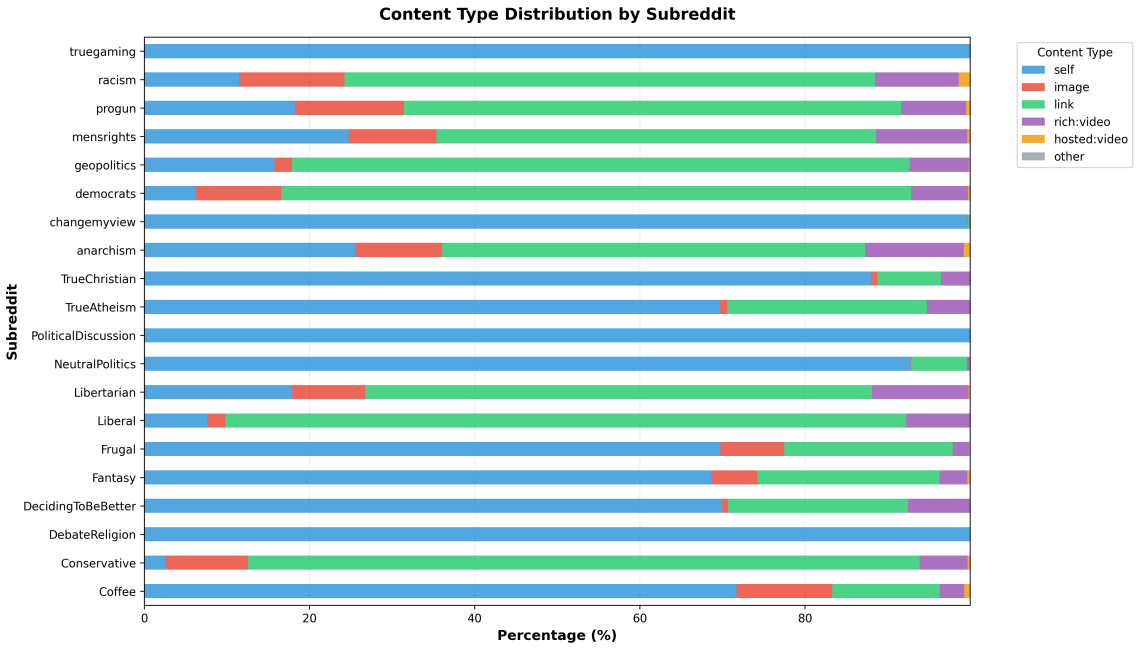

Figure 1: *Post Content Type Distribution per Subreddit*

analyzed posts made per year by subreddit in Figure 2 no clear trend for all subreddit, some appear to rise throughtout the years, some spike and fall of again, some seem to decline. One has to note that depending on the subreddit, a good of amount of posts had to be filtered out. its unclear, where those posts would fall into

next we looked at the score distribution in Figure 3 here we can see that the largest amount of posts are between a score of 1 to 10 and 11-100 respectively, note that a score of 100 as an example might be reached with a upvote ratio of 55%, meaning its not necessarily a popular post taking account Figure 4, we can see that although the amount of posting every year did not have a clear trend (Figure 2), the score seems to increase or roughly stay the same for most subreddits

4.3.2 Sentiment

4.3.3 TFIDF

used tfidf on subreddits added reddit/subreddit specific terms as custom stopwords

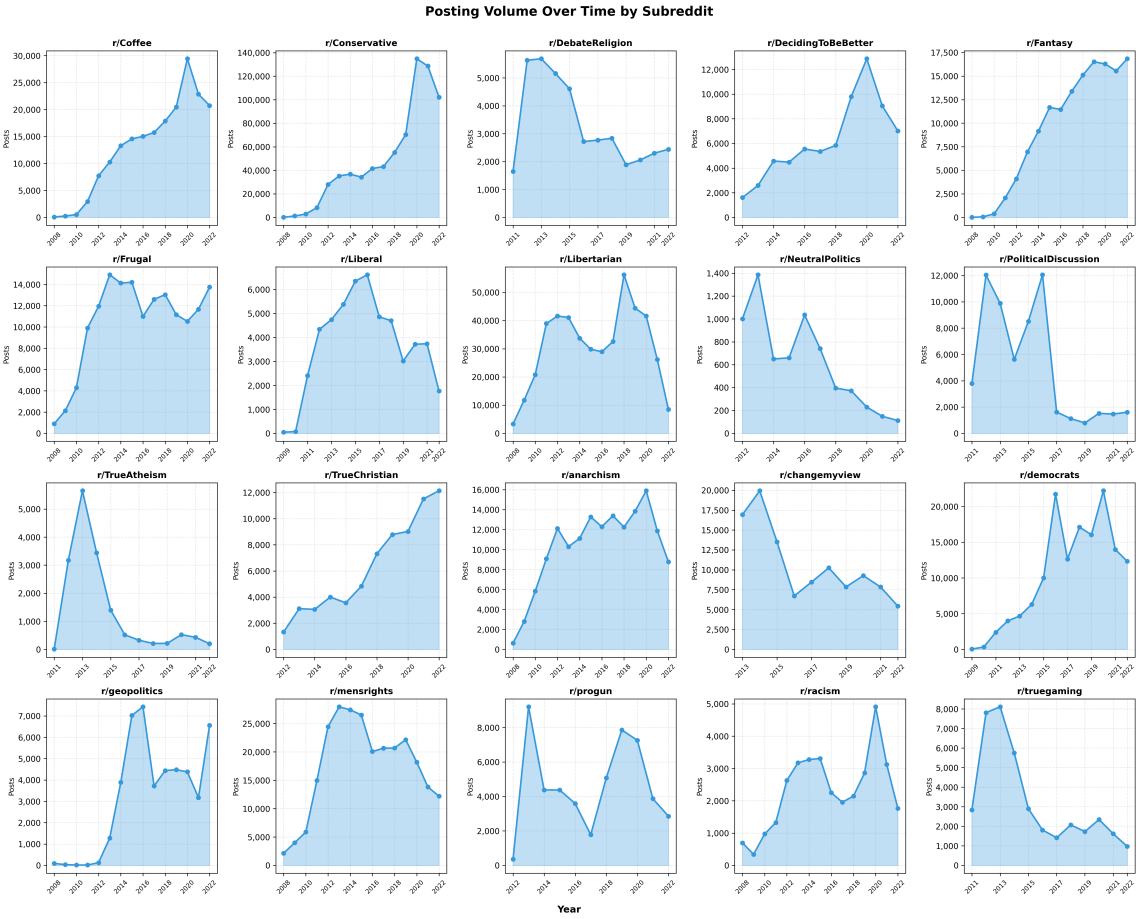

Figure 2: *Posting Volume per year*

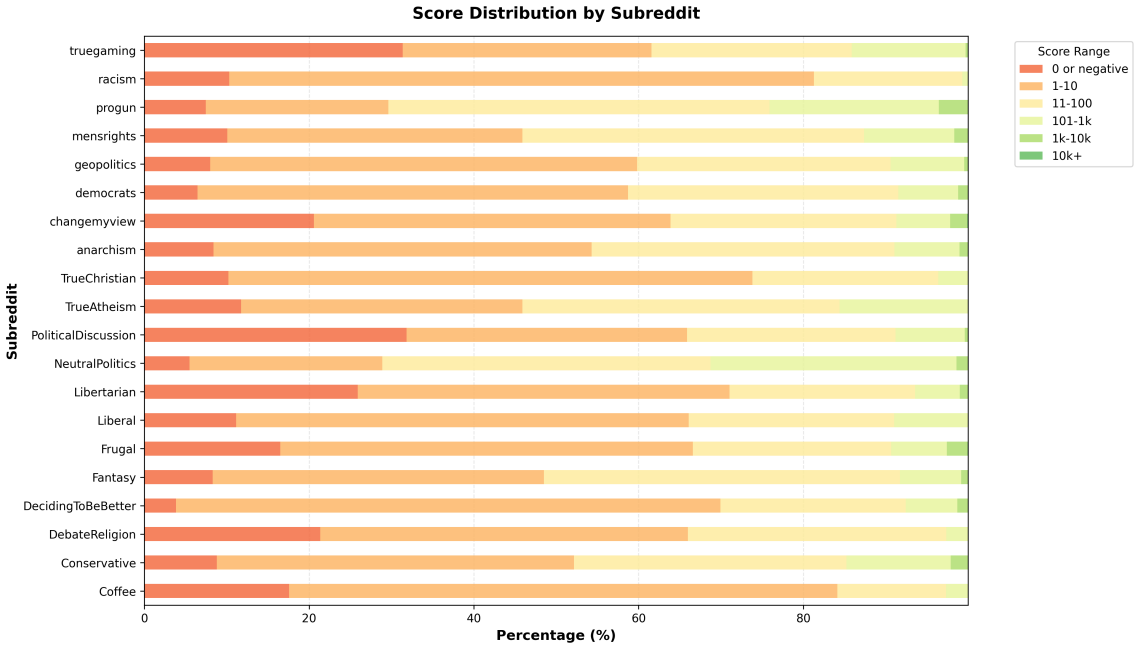

Figure 3: *Score Distribution*

Figure 4: *Score Stastistics over Time*

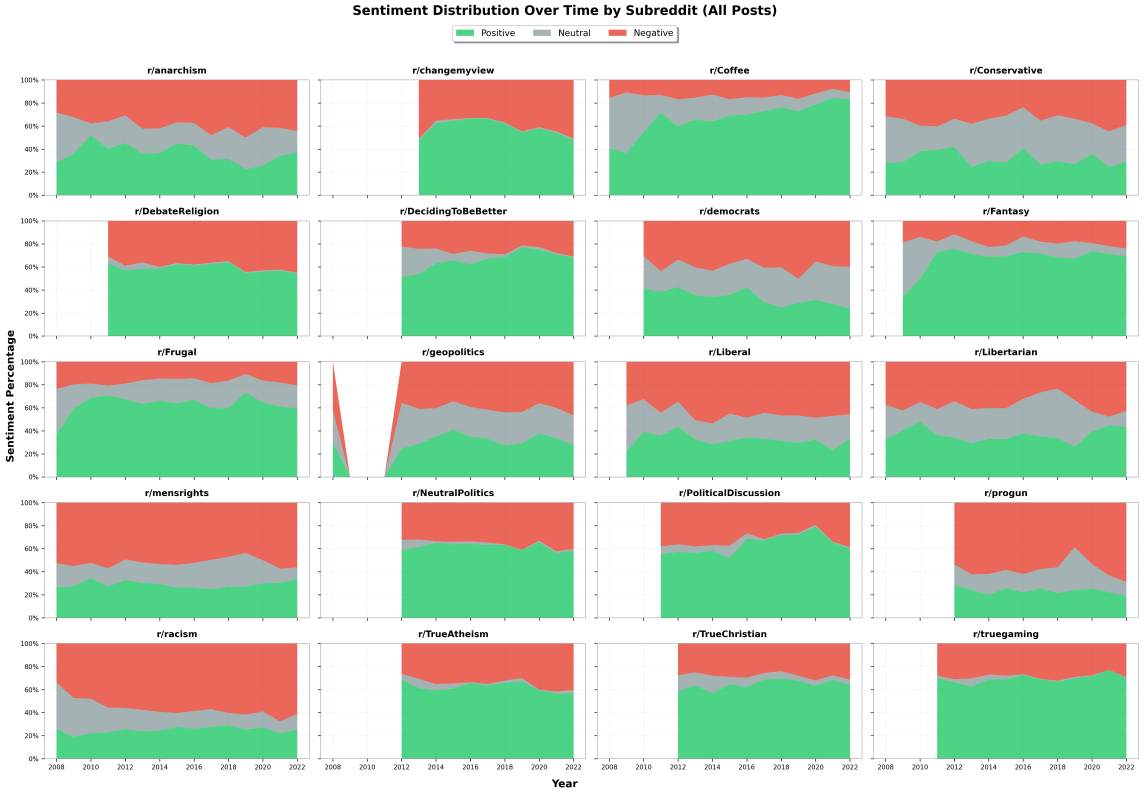

Figure 5: *Analyzing Sentiment per Subreddit over time*

terms such as state, republican, american, political, democrat, president, government were relevant throughout and were not included in above we aggregated the tfidf scores and set a minimum of 5 subreddits that the term should be included in

top terms across subreddits per timespans: 2008-2010: obama, health/obama care 2010-2012: obama, paul, debt, tax 2012-2014: obama, paul romney, war, gun, tax, obamacare 2014-2016: obama, gun, war, hillary clinton, trump 2016-2018: (donald) trump, hillary clinton, obama, bernie sanders 2018-2020: trump, president, democrats, tax, war, free 2020-2022: biden, trump, covid, coronavirus, joe, police

ignoring words that are very specific to the subreddit such as "liberal" for liberal or "conservative" for conservative subreddit or very broad terms such as america, white house, president, conservative, democratic and such

done by calculating the uniqueness: difference between tfidf score of term in subreddit and the average tfidf score of the term overall

distinctive words for subreddit conservative obamacare, media, obama, trump, ted cruz, cnn

distinctive words for subreddit libertarian ron paul, liberty, government, freedom, free market/speech,rights, rand paul, gary johnson,

distinctive words for subreddit progun gun, control, ban, nra, carry, amendment, rights, shooting

distinctive words for subreddit truechristian god, chrisitan, jesus, bible, ...

distinctive words for subreddit liberal obama, trump, torture, naomi klein, blue cross

distinctive words for subreddit democrats obama, trump, pelosi, tax, deficit, hillary clinton, bernie sanders, charlie rangel, ethics, tax,

distinctive words for subreddit progressive health care, bernie sanders, obama, court, public option

distinctive words for subreddit racism racist, racism, black, white, racial, ...

distinctive words for subreddit anarchism anarchist, anarchism, police, capitalism, revolution, ...

distinctive words for subreddit TrueAthemism atheism, religion, rage comics, science, life, ...

distinctive words for subreddit PoliticalDiscussion candidate, election, cote, party, trump, change, support

distinctive words for subreddit ENLIGHTENEDCENTRISM centrists, sides, left, nazi, facist

distinctive words for subreddit changemyview happy, couple, view, broken, free,

though amount posts are political in nature, alot have very different topics not neccessarily limited to politics

distinctive words for subreddit NeutralPolitics article, wikipedia, government, law, gov, usa/united states, state, congress, president, policy, election, vote, news, source,

heavy usage of wikipedia links in content

distinctive words for subreddit geopolitics china, russia, iran, war, oil, east, europe, global, military, nuclear, ...

distinctive words for subreddit DebateReligion theists, debate, answer, human, god, ...

term evolution of specific terms interesting: obamacare + healthcare + care

obama (inflated from obama care) romney biden + joe biden hillary clinton + hillary trump + donald trump

election president

government war

no real real differences aside from some minor fluctuations

analysis was orignally planned with finding datashifts or contradictions bewteen subreddit. For example climate shift vs climate hoax, fake news vs media seems that subreddits like democrats and conservaties are closer than expected

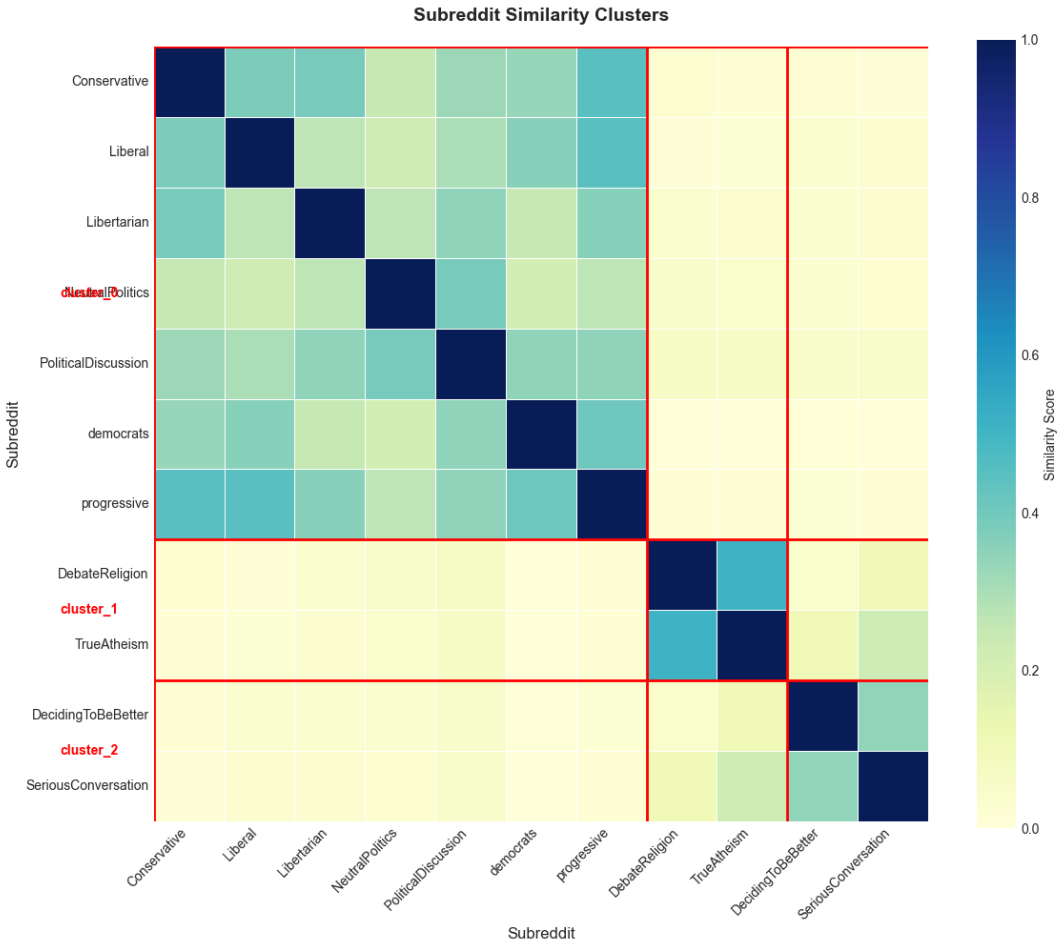

Figure 6: *Clustering Subreddits using TFIDF Similarity*

4.3.4 Link Posts and common URLs

good amount of posts in political subreddits are link posts which only contain a title and a link which often points to a news websites like breitbart, foxnews, dailywire, washingtonpost or social media sites like twitter, reddit itself

sincce link posts are missing the usual post content, they are a valuable indicator as to what the author is communicating with their post. as such, we need to find a way to add such information as an additional input for predicting popularity

4.3.5 Embeddings analysis

clustered and visualized top posts by popular, controversial and unpopular used comments for latter 2 as score cant be below 0 seems like clusters can be seen and subreddits seem to have somewhat individual things that are liked, disliked

determine if that is true by training models that try to determine if a post is from a certain subreddit / political subreddit group used minimum of x score or comments for dataset compare eval on val set from 2008-2015 vs after

4.4 Encoder

looked at Massive Text Embedding Benchmark (MTEB) Leaderboard [Mue+22] for top models

considerations next to performance were hard requirements and first and foremost the embedding dimensions the lower the better, as encoding millions of posts with different encoders or prompts costs alot of time and takes up alot of memory

for initial analysis we employed the *stella_en_400M_v5* model [Zha24] as its dimensions could be set to 1024, meaning smaller encoded datasets and maintained a great perofmenace in clustering and classification tasks while having low hardware requirements

during our later tests, additional models joined the leaderboards, so we also compared other encoders with similar requirements, excluding those that we would have had to pay for like gemini-embedding-001 (https://ai.google.dev/gemini-api/docs/embeddings) to find the one with the best performance for our use case. the resuls of which encoder performed the best for our task can be seen in the results section

using the high-performance computing (HPC) environment, specifically the HPC cluster *Bender* at the University of Bonn [Bon25], which provided the necessary computational resources to handle large-scale data processing.

4.5 Prompts

Different prompts for the best encoder model (stella)

no much research whats prompts are best used for social media or political posts specifically

tried to focus on different aspects

4.6 Gen AI

Schritte:

1. Generate Posts mit GenAI multishot prompting to generate text posts with distinct title and text content

2. encode post encode posts with LLM

3. classify post see if post is classified as unpopular, popular or controversial either throw away everything thats not popular or test prediction for those classes too

4. create reddit Post create post per api save post id and other information, in case post is deleted, in dataset

SLight alterating to the pipeline, each day flip a coin and if head, we dont throw away unpopular posts and just note the results down if its tails we classify and throw away if neccessarily at the end we can see which posts performed better

4.6.1 Generating the Post

first of all we should try to have a consistent persona for the posts depending on subreddit size, users might recognize the username and if two posts contradict each other as an example, the post might indicate the user be a woman in one post and a man in another thereby, they might be detected as written by AI, as there are reddit users that just try to post popular content and generate karma on their account to sell it later a similar issue is not repeating topics. As it stands, there might be a large amount of posts what will be thrown away. therefore it will be difficult to find enough content that does not repeat itself

4.6.2 Creating the Post

problem is the account used to Post people could look at account and see posts -> make account private also subreddits have karma minimums and you cant post otherwise or need to be manually reviewed manual posting using the following prompt to test requirements to post:

Generate a Reddit post. Choose an interesting topic. The post should fit into the subreddit r/truechristian. Make it seem user generated and not like it created by an AI. Give me a Title and Content for the post. Make it suitable for Reddit. Make it seem like a uni student wrote it from the language and sentence structure. Dont use em-dashes.

- Libertarian: post has to be unlocked by moderators (Sorry, this post has been removed by the moderators of r/Libertarian.) - Conservative: Auto mod filters post under karma thresholds - democrats: cant post under r/democrats comment karma threshold - liberal: cant post under r/liberal comment karma threshold - progressive: actually link post only - racism: awaiting moderator approval - metacanada: restricted community. Only approved users can post, but you

can still comment. - changemyview: automatically removed due to lack of posting/comment history - neutralpolitics: needs contain any links. According to the submission rules, this subreddit does not approve posts without links to sources. question needs to be in title, awaiting moderator approval, changed prompt used to generate post according to post requirements - geopolitics: need comment karma to post - decidingtobebetter: need total karma to post - truegaming: removed by reddit filters - fantasy: removed by reddit filters

filters could include karma, account age, posts, comments an such

upvote ratio verteilung angucken daten linearisieren

auf range von 0 bis 1 mappen (regreession)

einmal als perzentil zu definieren(10

Nur die krassesten aha momente in MA TFIDF bspw offen lassen, ob rein oder nicht abhängig, ob genug material die arbeit zu füllen

linear regressor (oder anderer der zwischen 0 und 1 mappt, xgboost oder random forrest) correlations coeffizient (kendal, spearman) einlesen, was gute Werte sind in den oberen 10

upvote ratio irgendwie auf 0 - 1 mappen so dass sinnvoll verteilt ist große terms sonst auf 1 oder 0 clippen, falls wenige daten ausschlagen

5 Results

5.1 Encoder Models

Table 3: *Popularity prediction results for popularity regressor with different encoder models. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
e5 Large Instruct	0.1154	0.1462	0.1680	0.4062	369,834
Stella	0.1153	0.1459	0.1700	0.4086	369,834
Kalm Mini Instruct	0.1158	0.1465	0.1636	0.4011	369,834
Bilingual	0.1165	0.1473	0.1559	0.3912	369,834
Gwen	0.1180	0.1490	0.1361	0.3658	369,834

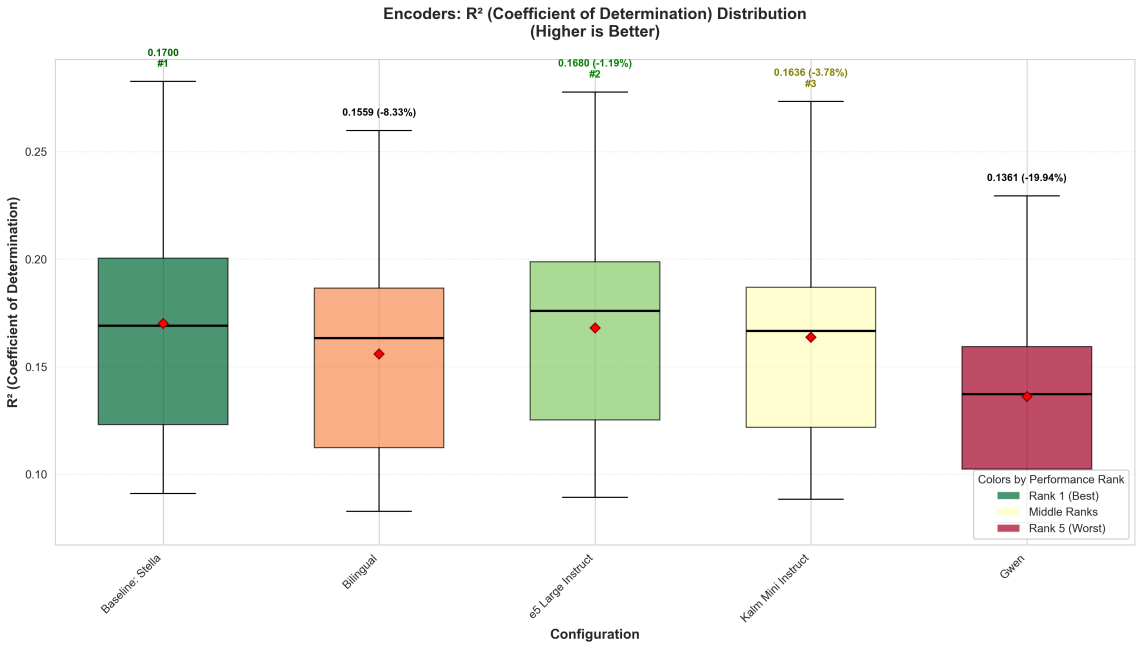

Figure 7: *Popularity prediction results for popularity regressor with different encoder models*

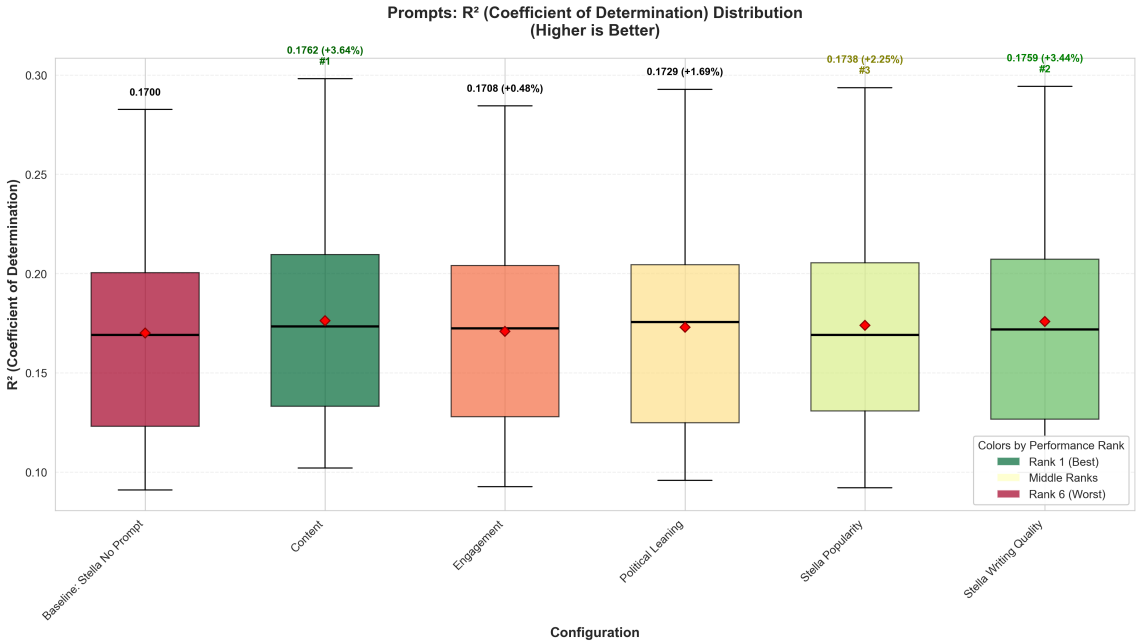

Figure 8: *Popularity prediction results for popularity regressor with different encoder prompts*

Table 4: *Popularity prediction results for popularity regressor with different encoder prompts. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
Stella Popularity	0.1150	0.1456	0.1738	0.4134	369,834
Stella Writing Quality	0.1148	0.1454	0.1759	0.4158	369,834
Stella Engagement	0.1152	0.1459	0.1708	0.4099	369,834
Stella Baseline	0.1153	0.1459	0.1700	0.4086	369,834
Stella Political Leaning	0.1151	0.1456	0.1729	0.4122	369,834
Stella Content	0.1148	0.1454	0.1762	0.4164	369,834

5.2 Prompts

5.3 Inputs

5.3.1 Dimensionality Reduction

Table 5: *Popularity prediction results for popularity regressor with different dimensionality reductions. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
Stella	0.1153	0.1459	0.1700	0.4086	369,834
Per-Subreddit PCA 512	0.1153	0.1459	0.1700	0.4086	369,834
Global PCA 256	0.1156	0.1462	0.1669	0.4049	369,834
Global PCA 128	0.1154	0.1461	0.1680	0.4064	369,834
Global PCA 512	0.1207	0.1519	0.0976	0.2630	369,834

all dimensionality reductions decrease performance it seems the 1024 all contain relevant information, some of which would be lost by reducing them we therefore use full 1024

5.3.2 Aggregation methods

despite being the most naive approach, concat has the best performance by a small margin this aligns with our previous finding that the information in the 1024 is relevant and any method that that tries to makethe information more compact decreases performance

5.3.3 Link vs Self Posts

the next thing we tried was to use separate models for self posts and link posts

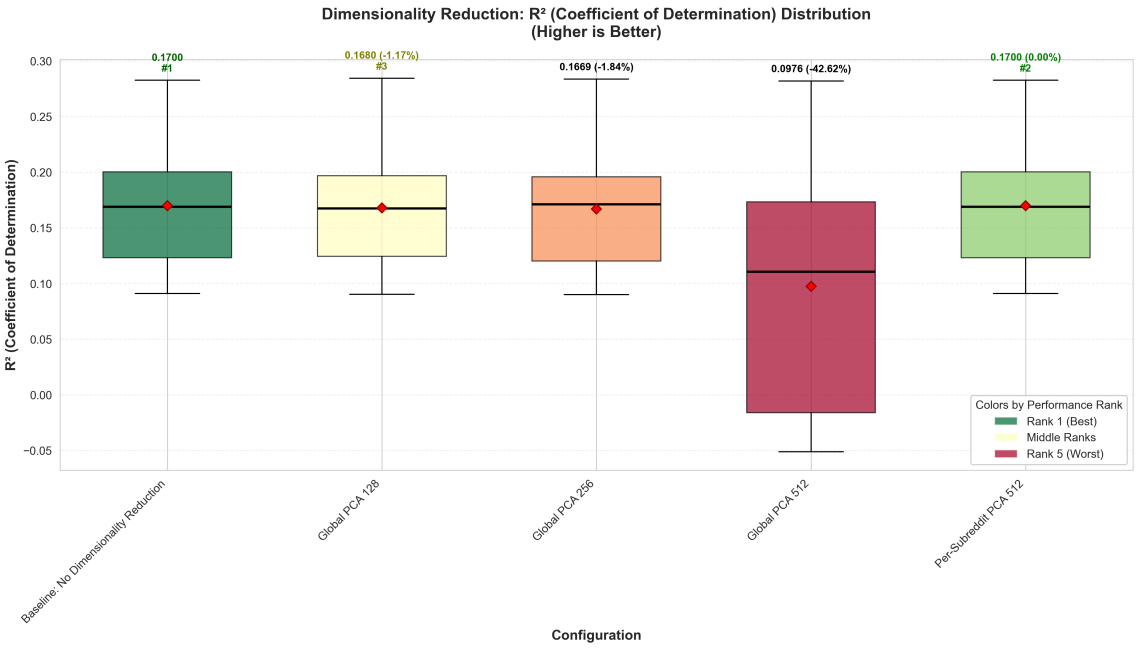

Figure 9: *Popularity prediction results for popularity regressor with different dimensionality reductions*

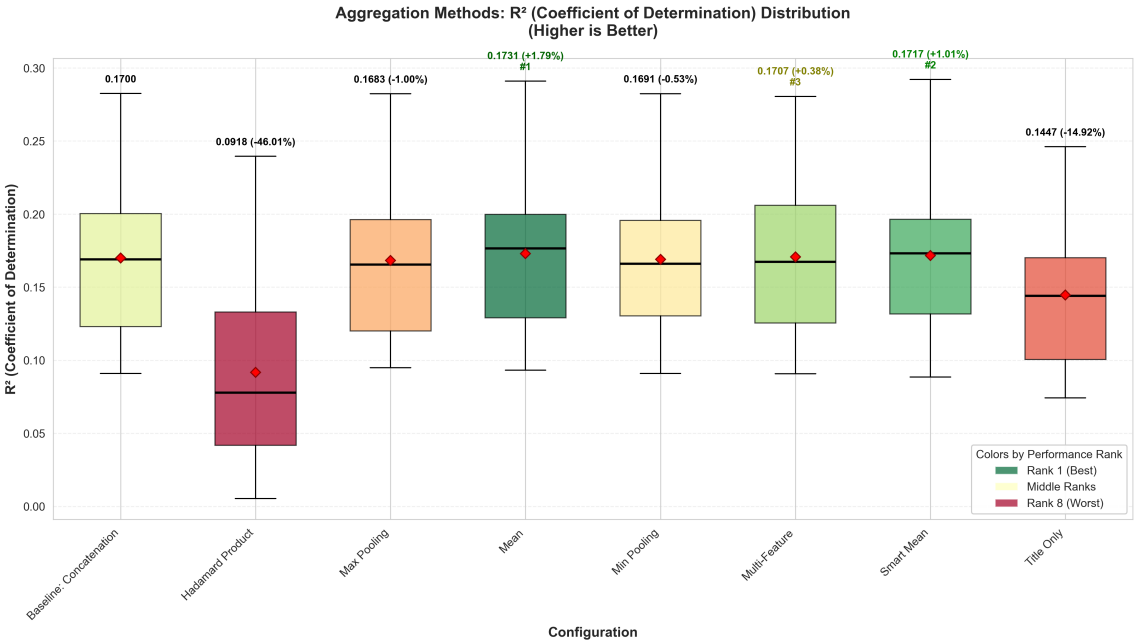

Figure 10: *Popularity prediction results for popularity regressor with*

Table 6: *Popularity prediction results for popularity regressor with different aggregation methods. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
Hadamard Product	0.1211	0.1525	0.0918	0.2843	369,834
Mean	0.1150	0.1457	0.1731	0.4124	369,834
Multi-Feature	0.1153	0.1459	0.1707	0.4094	369,834
Stella	0.1153	0.1459	0.1700	0.4086	369,834
Min Pooling	0.1154	0.1460	0.1691	0.4078	369,834
Max Pooling	0.1154	0.1461	0.1683	0.4066	369,834
Smart Mean	0.1152	0.1458	0.1717	0.4107	369,834
Title Only	0.1173	0.1483	0.1447	0.3763	369,834

5.3.4 Additional Inputs

only use inputs that are available at time of prediction and that can be leveraged for own post creation (no comment amount, no karma or other author information)

top terms coverage tfidf domain analysis

Domain Analysis

Table 7: *Popularity prediction results for popularity regressor with different author feature combinations. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
Self+Link - Domain All	0.1146	0.1452	0.1792	0.4200	369,834
Self+Link - Domain Scores	0.1146	0.1452	0.1793	0.4201	369,834
Self+Link - Domain Binary	0.1153	0.1459	0.1709	0.4097	369,834
Stella	0.1153	0.1459	0.1700	0.4086	369,834

Post Time Features

Author Features

While some author features showed weaker individual performance, the combined feature set achieved optimal results (R^2: 0.2016), suggesting synergistic effects between features. Further ablation of specific feature subsets was deemed unnecessary given computational constraints and the marginal nature of potential gains.

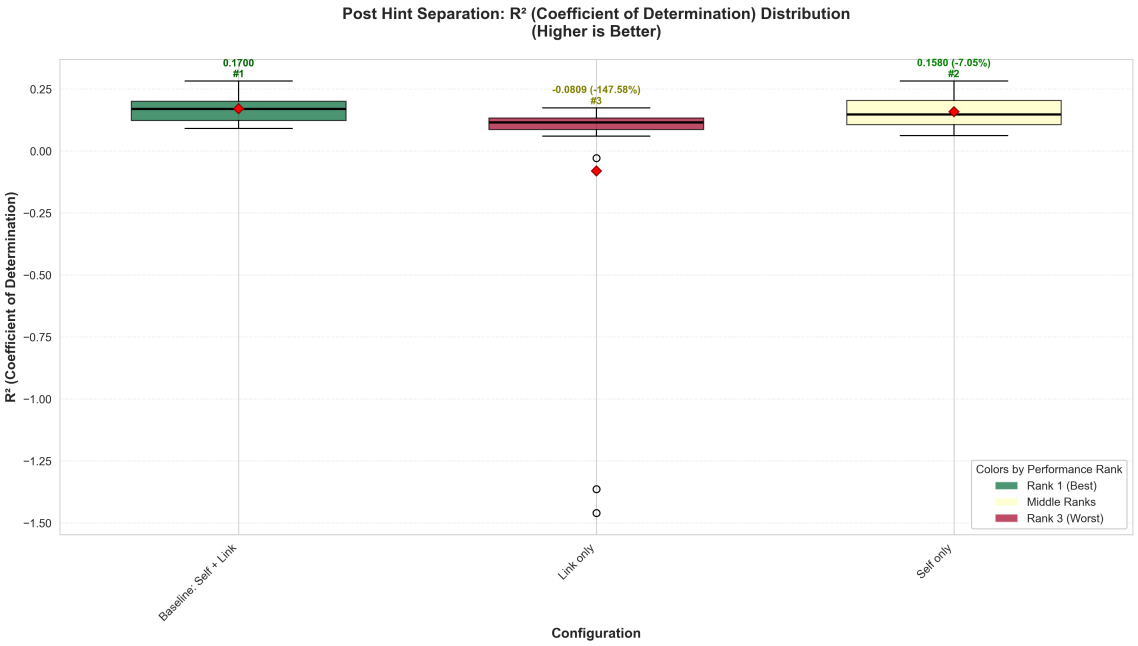

Figure 11: *Popularity prediction results for popularity regressor with Post Hint separation*

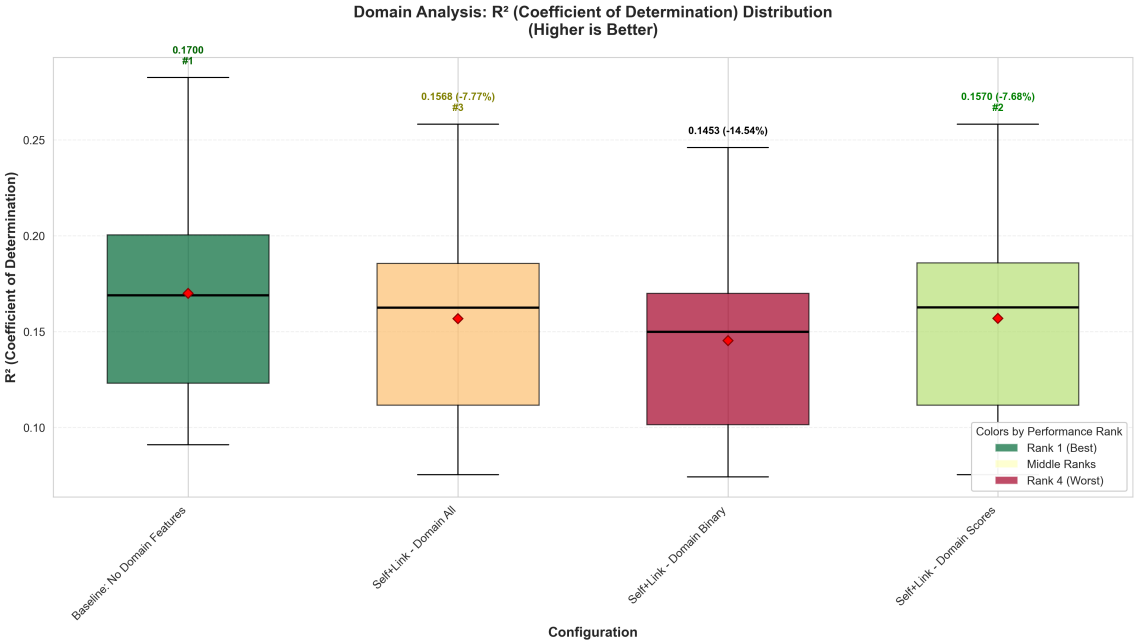

Figure 12: *Popularity prediction results for popularity regressor with domain features*

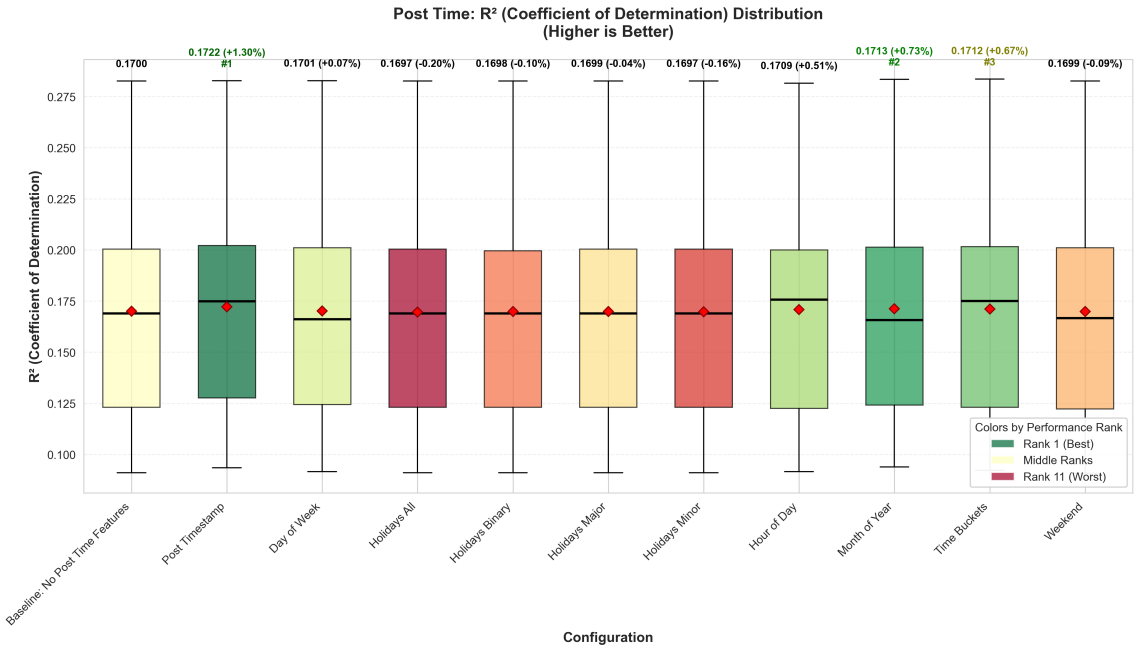

Figure 13: *Popularity prediction results for popularity regressor with post time features*

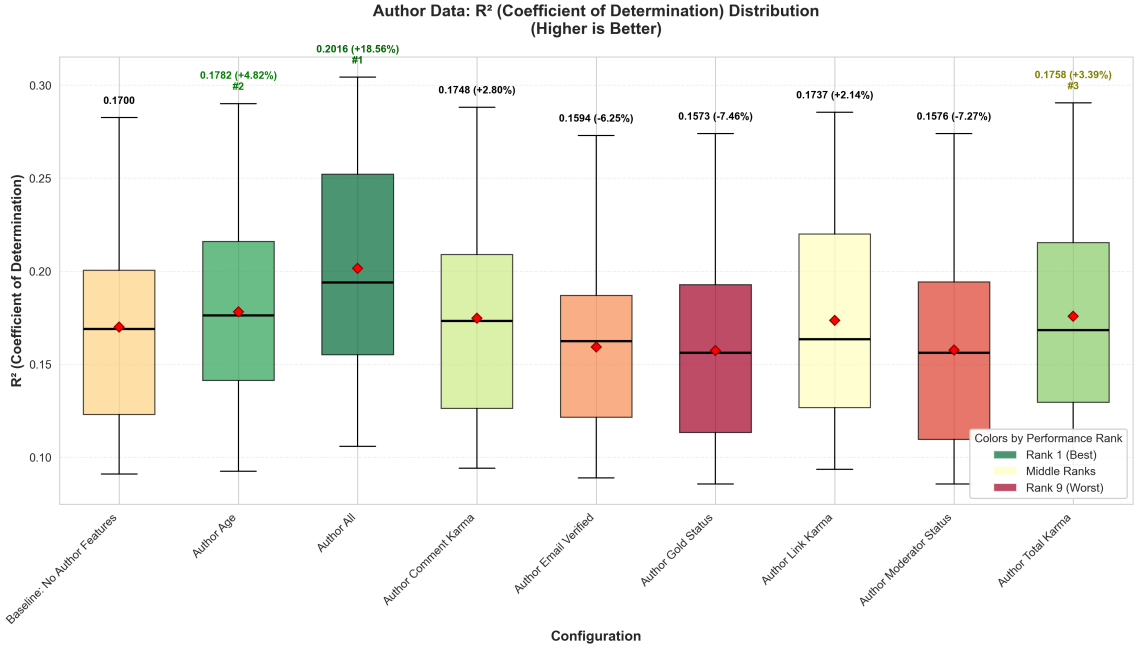

Figure 14: *Popularity prediction results for popularity regressor with author features*

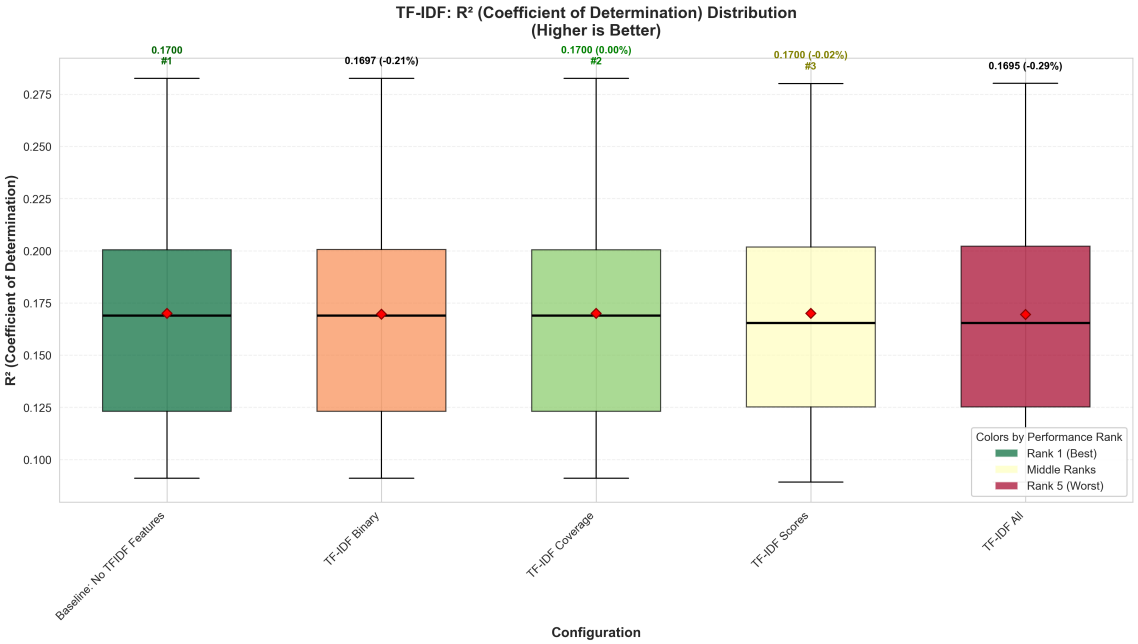

Figure 15: *Popularity prediction results for popularity regressor with tfidf features*

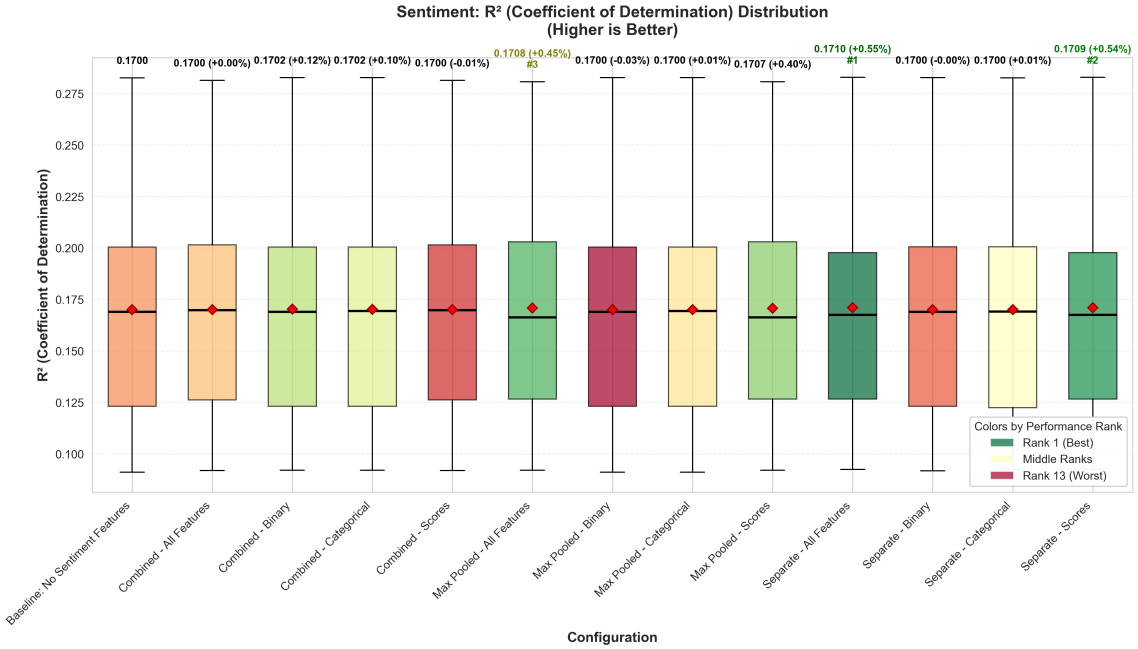

Figure 16: *Popularity prediction results for popularity regressor with sentiment features*

Table 8: *Popularity prediction results for popularity regressor with different author feature combinations. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
Post Timestamp	0.1151	0.1458	0.1722	0.4115	369,834
Hour of Day	0.1152	0.1459	0.1709	0.4097	369,834
Month of Year	0.1152	0.1458	0.1713	0.4104	369,834
Time Buckets	0.1152	0.1458	0.1712	0.4099	369,834
Day of Week	0.1153	0.1459	0.1701	0.4089	369,834
Holidays Minor	0.1154	0.1460	0.1697	0.4082	369,834
Stella	0.1153	0.1459	0.1700	0.4086	369,834
Holidays All	0.1154	0.1460	0.1697	0.4081	369,834
Holidays Binary	0.1153	0.1459	0.1698	0.4084	369,834
Holidays Major	0.1153	0.1459	0.1699	0.4085	369,834
Weekend	0.1153	0.1459	0.1699	0.4085	369,834

TFIDF

Sentiment

5.3.5 Cross-Generalizability

5.4 Gen AI

As a baseline, we calculated the average upvote ratio, score and comments per post for all subreddits in our popular post creation test. we also grouped the posts by users, as a few users might skew the average by creating lots of very popular or unpopular posts. The resulting difference was negligibly though.

6 Discussion

6.1 Learning/Limitation

using chat gpt helped with smaller code fragments alot, but left much to be desired for more complex and large code systems

Table 9: *Popularity prediction results for popularity regressor with different author feature combinations. Results show mean performance across all subreddits.*

Features	Precision	Recall	F1-Macro	Accuracy	Total	% Popular
Author All	0.6952	0.6354	0.6636	0.6478	220,855	50.01%
Author Age	0.6938	0.6268	0.6582	0.6399	220,855	50.01%
Author Link Karma	0.6858	0.6227	0.6524	0.6349	220,855	50.01%
Author Comment Karma	0.6839	0.6234	0.6519	0.6350	220,855	50.01%
Stella	0.6841	0.6223	0.6513	0.6341	369,834	50.00%
Author Total Karma	0.6838	0.6210	0.6505	0.6327	220,855	50.01%
Author Moderator Status	0.6838	0.6182	0.6489	0.6304	220,855	50.01%
Author Email Verified	0.6827	0.6172	0.6479	0.6293	220,855	50.01%
Author Gold Status	0.6823	0.6169	0.6476	0.6289	220,855	50.01%

training model consistency: sometimes hard to gauge how much model improves when certain randomness and noise is introduced during training, causing performance deviations that are not attributed to the "improvements"

GPU histogram building and multi-threading can introduce tiny numeric drift, and you're using stochastic subsampling + early stopping in combination with the above.

therefore decided to use no multithreading and set other specific options to keep it as consistent and deterministic as possible

7 Conclusion

References

[] *PRAW: The Python Reddit API Wrapper*. Accessed: 24.01.2024. URL: https://praw.readthedocs.io/en/stable/.

[Bau+20] Baumgartner, Jason et al.: *The Pushshift Reddit Dataset*. 2020. arXiv: 2001.08435 [cs.SI]. URL: https://arxiv.org/abs/2001.08435.

[Bon25] Bonn, University of: *Bender HPC Cluster*. Accessed: 2025-02-26. 2025. URL: https://www.hpc.uni-bonn.de/.

Table 10: *Popularity prediction results for popularity regressor with different author feature combinations. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
Author All	0.1090	0.1390	0.2016	0.4444	220,855
Author Age	0.1108	0.1412	0.1782	0.4180	220,855
Author Link Karma	0.1113	0.1415	0.1737	0.4128	220,855
Author Comment Karma	0.1113	0.1414	0.1748	0.4141	220,855
Stella	0.1153	0.1459	0.1700	0.4086	369,834
Author Total Karma	0.1112	0.1413	0.1758	0.4152	220,855
Author Moderator Status	0.1126	0.1429	0.1576	0.3931	220,855
Author Email Verified	0.1125	0.1428	0.1594	0.3955	220,855
Author Gold Status	0.1126	0.1429	0.1573	0.3928	220,855

Table 11: *Popularity prediction results for popularity regressor with different TFIDF feature combinations. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
TF-IDF All Features	0.1153	0.1460	0.1695	0.4081	369,834
TF-IDF Binary	0.1153	0.1460	0.1697	0.4081	369,834
TF-IDF Scores	0.1153	0.1459	0.1700	0.4086	369,834
Stella	0.1153	0.1459	0.1700	0.4086	369,834
TF-IDF Coverage	0.1153	0.1459	0.1700	0.4086	369,834

[Cen] Center, Pew Research: *Seven-in-Ten Reddit Users Get News on the Site*. Accessed: 24.01.2024. URL: https://www.pewresearch.org/journalism/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/.

[DHM17] Deaton, Sean ; Hutchison, Scott ; Matthews, Suzanne J: "Using Machine Learning to Predict the Popularity of Reddit Comments". In: *seandeaton. com* (2017).

[HL19] Hessel, Jack ; Lee, Lillian: "Something's brewing! Early prediction of controversy-causing posts from discussion features". In: *arXiv preprint arXiv:1904.07372* (2019).

[Inc] Inc, Reddit: *Reddit Inc.* Accessed: 24.01.2024. URL: https://redditinc.com.

[Lan19] Langsæther, Peter Egge: "Religious voting and moral traditionalism: The moderating role of party characteristics". In: *Electoral Studies* 62 (2019), p. 102095.

[Mad+22] Madrid Jr, Raul et al.: "The relevance of religion for political office: voter bias toward candidates from different religious backgrounds". In: *Political Behavior* 44.2 (2022), pp. 981–1001.

[MHW21] Mamié, Robin ; Horta Ribeiro, Manoel ; West, Robert: "Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube". In: *Proceedings of the 13th ACM Web Science Conference 2021*. 2021, pp. 139–147.

Table 12: *Popularity prediction results for popularity regressor with different TFIDF feature combinations. Results show mean performance across all subreddits.*

Features	MAE	RMSE	R^2	r	Total
Separate - Scores	0.1152	0.1459	0.1709	0.4101	369,834
Separate - All Features	0.1152	0.1459	0.1710	0.4101	369,834
Combined - Binary	0.1153	0.1459	0.1702	0.4089	369,834
Separate - Binary	0.1153	0.1459	0.1700	0.4086	369,834
Combined - Categorical	0.1153	0.1459	0.1702	0.4088	369,834
Max Pooled - Categorical	0.1153	0.1459	0.1700	0.4087	369,834
Max Pooled - Binary	0.1153	0.1459	0.1700	0.4085	369,834
Stella	0.1153	0.1459	0.1700	0.4086	369,834
Separate - Categorical	0.1153	0.1459	0.1700	0.4086	369,834
Max Pooled - All Features	0.1153	0.1459	0.1708	0.4097	369,834
Max Pooled - Scores	0.1153	0.1459	0.1707	0.4096	369,834
Combined - All Features	0.1153	0.1459	0.1700	0.4088	369,834
Combined - Scores	0.1153	0.1459	0.1700	0.4087	369,834

[Mue+22] Muennighoff, Niklas et al.: "MTEB: Massive Text Embedding Benchmark". In: *arXiv preprint arXiv:2210.07316* (2022). URL: https://arxiv.org/abs/2210.07316.

[Per99] Perry, Michael J: *Religion in politics: Constitutional and moral perspectives*. Oxford University Press, 1999.

[Pot20] Potz, Maciej: "Political Science of Religion". In: *Theorising the Political Role of Religion* (2020).

[Sto15] Stoddard, Greg: "Popularity and quality in social news aggregators: A study of reddit and hacker news". In: *Proceedings of the 24th international conference on world wide web*. 2015, pp. 815–818.

[SZ12] Segall, Jordan ; Zamoshchin, Alex: "Predicting Reddit post popularity". In: *nd): n. pag. Stanford University* (2012).

[TT14] Terentiev, Andrei ; Tempest, Alanna: "Predicting Reddit Post Popularity Via Initial Commentary". In: *nd): n. pag* (2014).

[WA21] Waller, Isaac ; Anderson, Ashton: "Quantifying social organization and political polarization in online platforms". In: *Nature* 600.7888 (2021), pp. 264–268.

[Wig19] Wigsnes, Fredrik: "Predicting popularity of Reddit posts using machine learning". MA thesis. University of Stavanger, Norway, 2019.

[Zha24] Zhang, Dun: *stella_en_400M_v5*. 2024. URL: https://huggingface.co/dunzhang/stella_en_400M_v5.

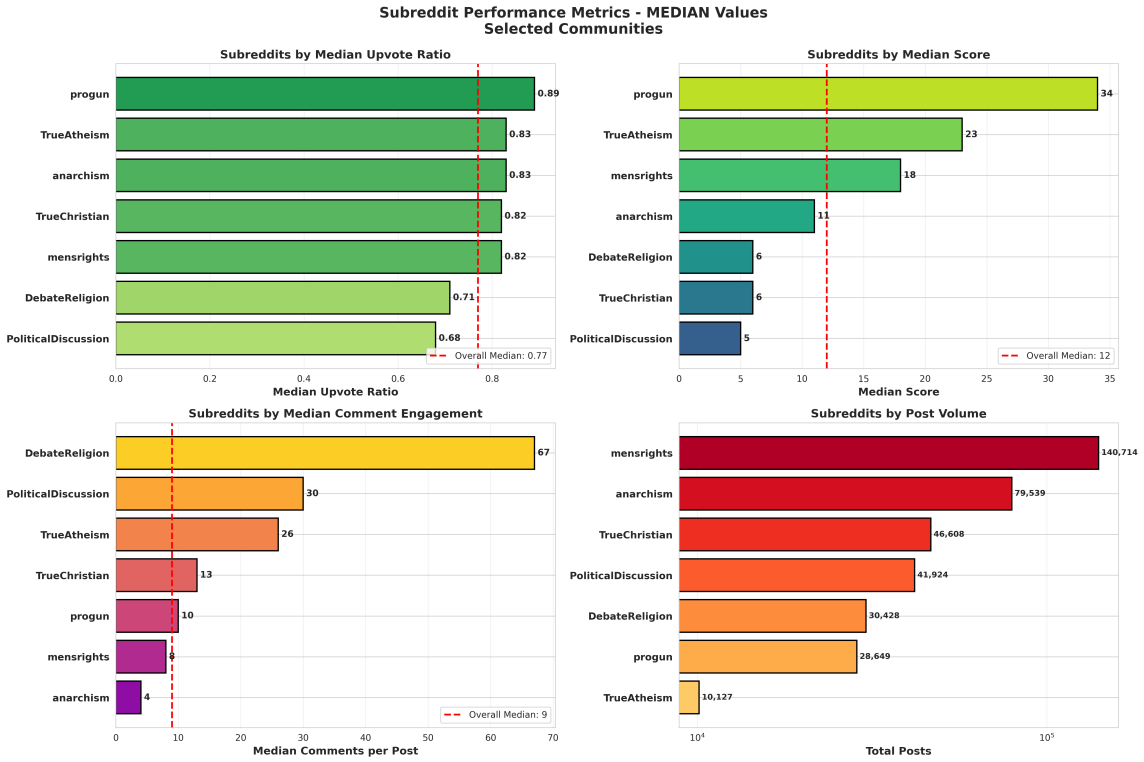

Figure 17: *Average Post Metrics per subreddit*

Table 13: *Subreddit performance statistics showing median and mean values for engagement metrics across selected communities.*

Subreddit	Upvote Ratio		Score		Comments		Posts
	Median	Mean	Median	Mean	Median	Mean	Total
progun	0.89	0.841	34	118.9	10	27.3	28,649
TrueAtheism	0.83	0.783	23	61.8	26	48.7	10,127
anarchism	0.83	0.810	11	39.7	4	12.6	79,539
TrueChristian	0.82	0.790	6	18.6	13	24.1	46,608
mensrights	0.82	0.788	18	78.9	8	22.2	140,714
DebateReligion	0.71	0.679	6	16.1	67	105.1	30,428
PoliticalDiscussion	0.68	0.657	5	49.9	30	101.2	41,924