Sentiment Variability in Online Communities: A Comparative Analysis of Business-Related Subreddits Using Natural Language Processing

Anna D. Kyosova

Graduate School of Technology Management, Ritsumeikan University, Osaka, Japan
Email: a.kyosova@gmail.com (A.D.K.)

Manuscript received October 20, 2023; revised November 28, 2023; accepted February 19, 2024; published July 15, 2024.

***Abstract*—This research aims to shed light on how community sentiment varies across nine distinct subreddits focused on entrepreneurship and business. To ensure highly reliable data, we used a multi-faceted approach combining automated data extraction with manual curation techniques. Advanced text pre-processing methods were utilized along with the syuzhet package in R for precise sentiment analysis. Upon conducting pairwise hypothesis testing, it was found that there are statistically relevant variations in mood between various subreddit pairs while others demonstrated no notable disparities at all levels analyzed. Our findings provide a comprehensive understanding of how views on business ideas differ across online communities and offer valuable guidance for potential entrepreneurs, investors, and policymakers. Additionally, our approach can easily be adapted for other social media platforms but limitations like external factors must be considered when interpreting the results obtained solely from Reddit data. To expand upon this research further exploration into longitudinal analysis is suggested along with delving into the impact that external events have on community perception towards certain subjects/subreddit topic areas may also warrant investigation. Our research bears weight beyond the sphere of entrepreneurs, reaching a wider audience that includes investors, educators, and policymakers with an interest in assessing community perceptions regarding entrepreneurial patterns. Our findings are pivotal to advancing conversations surrounding how big data and machine learning can facilitate grasping and interpreting entrepreneurial sentiment within our digital age.**

***Keywords*—business ideas and trends, pairwise hypothesis testing, Reddit, sentiment analysis**

I. Introduction

In 2006, Jeff Howe's introduction of crowdsourcing brought about a fundamental change in how organizations seek out ideas, services and content. The concept taps into the collective intelligence of large communities to address problems and drive innovation. Crowdsourcing initiatives have proven effective due to "the wisdom of the crowd", surpassing results that could be achieved by individual experts (Surowiecki, 2004). Social media platforms' widespread use has transformed crowdsourcing significantly as it overcomes geographical, temporal, and linguistic obstacles that previously posed significant barriers.

With regards to business discussions, online forums can provide valuable perceptions on public sentiments; providing aid to entrepreneurs and investors with their decision-making processes. However popular these platforms may be and despite their potential value; there remains a need for thorough investigations into the variability of opinions found between various communities centered around business-related subjects.

By employing a rigorous methodology that involves both automated data extraction and manual curation, this study endeavours to fill the void by conducting an extensive sentiment analysis across several Reddit subreddits[1] related to business idea generation and entrepreneurship. The main objective of this research is to address the following research inquiries: Do the sentiments expressed vary significantly among various subreddits related to businesses based on statistical analysis? Which factors have the potential to affect these fluctuations in sentiment? Gaining a comprehensive understanding of the subtle variations in public perception among distinct online communities can provide essential direction to stakeholders operating within various business ecosystems. The knowledge obtained from this examination is not solely for scholarly use, as it holds noteworthy pragmatic consequences and applies equitably to entrepreneurs and investors alike.

II. Literature Review

The extensive use of social media platforms has brought about a paradigm shift in how people interact, discuss, and share information on various subjects, including business and entrepreneurship. The current literature demonstrates burgeoning interest in employing Natural Language Processing (NLP) and sentiment analysis techniques to gain insights into public opinions, beliefs, and attitudes.

Sentiment analysis or opinion mining is a sub-discipline within NLP that focuses on identifying and categorizing opinions and sentiments expressed in textual data. Early works in the field concentrated mainly on product reviews and social media content (Pang & Lee, 2008). The methodologies involved range from simple lexicon-based approaches to more sophisticated machine learning models, including but not limited to Naive Bayes, Support Vector Machines, and neural networks (Cambria *et al.*, 2013; Zhang *et al.*, 2018).

Reddit, one of the foremost platforms for community-driven discussions, has recently gained attention from scholars for its diverse range of subreddits that allow targeted studies across different topics, including politics, healthcare, and business (Paul & Dredze, 2011; Althoff *et al.*, 2014). Multiple studies have utilized Reddit data to study phenomena like political polarization, mental health issues, and market trends (Massa & Avesani, 2015; Choudhury & De, 2014).

[1] In this study, the prefix 'r/' has been omitted from the subreddit names for stylistic consistency and to enhance readability

doi:10.18178/joebm.2024.12.3.804

While there is abundant research on sentiment analysis in online platforms, the application of such methodologies to specific business and entrepreneurship contexts is still relatively unexplored. A handful of studies have ventured into analyzing startup pitches, business announcements, and investment trends using sentiment analysis (Loughran & McDonald, 2011; Antweiler & Frank, 2004). However, comprehensive investigations that target specific Reddit subreddits related to business ideas and entrepreneurship remain limited.

What appears to be missing is an integrative study that captures the nuances of sentiment variations across multiple, closely related, business-focused communities on Reddit. Such a study could not only deepen our understanding of public sentiment but also provide invaluable insights for entrepreneurs and investors alike.

III. Data and Methods

A. Collection of Data

We utilized a two-stage approach to gather information from Reddit, which is widely recognized for its community-driven conversations as a renowned social media platform. We utilized a two-stage approach to gather information from Reddit using the RedditExtractor R package. After retrieving a broad range of subreddits, we manually selected the nine most relevant for business idea generation, ensuring the removal of irrelevant or spam content to maintain data quality. Finally, these curated subreddits were manually scrutinized once again and cleaned off irrelevant posts and spam content so that only high-quality data could be obtained through our efforts.

B. Pre-processing

Using the readxl package, R was used to import cleaned subreddit data. For pre-processing the subreddit data, we used the readxl, tm, and textclean packages in R, applying standard NLP normalization techniques such as stop-word removal and stemming.

C. Analysis of Sentiments

Sentiment analysis was conducted using the syuzhet package in R, employing a combination of lexical dictionaries and machine learning algorithms to evaluate sentiments in posts and comments. We aggregated these scores to create subreddit-level sentiment indicators and conducted pairwise hypothesis testing to identify differences, with statistical analyses performed in IBM SPSS Statistics 27.

IV. Initial Data

Table 1 shows the quantity of posts for each of the selected subreddits, their score the number of generated comments as well as the number of subscribers at the time of extraction.

Generally, the table shows that the most active business subreddits in the dataset are the ‘Entrepreneur’ and ‘smallbusiness’ subreddits with the highest number of posts (998). ‘Entrepreneur’ also leads in comments (31791) and subscribers (1186541), signaling a highly active community and the broad appeal of the subjects discussed.

Table 1. Main information

Subreddits	Posts	Score	Comments	Subscribers
Entrepreneur	998	33278	31791	1186541
smallbusiness	998	12898	17567	801205
business	685	58734	13027	709327
ecommerce	481	2363	4173	198760
EntrepreneurRideAlong	406	3133	2496	205628
Business_Ideas	384	2157	2919	126989
startups	147	3731	4495	973672
sidehustle	61	1317	1652	102524
Lightbulb	30	778	198	101110
Total	4190	118389	78318	4405756

Table 1 summarizes the activity across selected subreddits, highlighting ‘Entrepreneur’ and ‘smallbusiness’ as the most active in terms of posts, comments, and subscribers. Notably, ‘business’ shows high engagement despite fewer posts, indicating quality discussions. In contrast, ‘ecommerce’ and smaller subreddits like ‘Lightbulb’ show lower engagement metrics, suggesting more niche or specialized content.

Based on the data we can conclude that ‘Entrepreneur’ is a hub for entrepreneurial discussions, however, ‘business’ despite fewer posts, has high engagement, indicating the quality of discussions.

V. Results

A. The Score of the Posts

The Reddit score is a measure of popularity, it is formed by the sum of the positive reactions to a post or comment (upvotes) minus the negative (downvotes) that the readers have. Table 2 below shows the main descriptive statistics of the score indicator, depending on the subreddit of the posts.

Table 2. Score for all subreddits

Subreddit	N	Mean	Median	Std. Deviation	Minimum	Maximum	Range
business	685	85.74	2.00	235.761	0	1918	1918
Business_ Ideas	384	5.62	2.00	20.837	0	315	315
ecommerce	481	4.91	3.00	7.473	0	88	88
Entrepreneur	998	33.34	3.00	129.311	0	1716	1716
Entrepreneur RideAlong	406	7.72	2.00	21.774	0	262	262
Lightbulb	30	25.93	9.50	52.009	0	283	283
sidehustle	61	21.59	13.00	29.998	0	173	173
small business	998	12.92	2.00	39.691	0	532	532
startups	147	25.38	13.00	35.675	0	193	193
Total	4190	28.26	2.00	119.802	0	1918	1918

Subreddits like ‘business’ and ‘Entrepreneur’ exhibit high variability in scores, suggesting diverse engagement, while ‘Business_Ideas’ and ‘EntrepreneurRideAlong’ show more consistency, indicative of focused topics.

Based on the calculated averages, it can be assumed that, depending on the subreddit of the post, the value of the score indicator changes. It is also seen that the value of the average is higher than the median. This suggests that in each subreddit, there are individual posts that have been given very high ratings. Let's formulate statistical hypotheses:

H0: The value of the score indicator does not change depending on the subreddit of the post.

H1: The value of the score indicator varies depending on the subreddit of the post.

Before the analysis, the data were checked for the normality of the distribution. The assumption of the normality of the distribution was not confirmed using the Kolmogorov-Smirnov criterion (*p*-value < 0.05). Therefore, the nonparametric Kruskal-Wallis test will be used for further analysis.

The results of hypothesis testing are presented in Table 3.

Table 3. Independent-Samples Kruskal-Wallis test

Independent-Samples Kruskal-Wallis Test Summary	
Total N	4190
Test Statistic	276.505[a]
Degree Of Freedom	8
Asymptotic Sig.(2-sided test)	0.000

a. The test statistic is adjusted for ties.

The results of the analysis show that the null hypothesis should be rejected (*p*-value < 0.05) and an alternative hypothesis accepted. The value of the score indicator varies depending on the subreddit of the post.

Table 4 presents the outcomes of pairwise multiple comparisons, indicating which comparison pairs display statistical significance.

The table shows that most comparison pairs differ statistically significantly (*p*-value < 0.05). This means that the rating that people give to the posts varies depending on the subreddit of the post as well as its content.

There is no significant difference between 'Business_Ideas' and 'Entrepreneur RideAlong' (Sig.=0.642). However, there is a positive test statistic indicating significance when comparing 'Business_Ideas' to the 'business' subreddit (Sig.=0.000). When compared to other subreddits such as 'smallbusiness', 'ecommerce', and 'Entrepreneurship' among others, there are significant differences with negative test statistics for each category (Sig.=0.000 for all comparisons).

The data indicates a noteworthy contrast between the outcomes of 'EntrepreneurRideAlong' and 'business' (Sig.=0.000). When compared to others, 'EntrepreneurRideAlong' generally exhibits notable negative test statistics; however, it shows significantly positive results when compared with 'ecommerce', or 'Entrepreneur' (Sig.=0.000).

When comparing the 'business' subreddit to the other subreddits in our dataset no significant negative difference between it and 'smallbusiness' (Sig.=0.067) was detected. However, 'business' displays a noteworthy negative distinction compared to subreddits such as 'ecommerce', or 'Entrepreneurship' with Sig. values ranging from 0.000 to 0.013.

While 'smallbusinesses' and 'ecommerce' exhibit a positive difference, it is not statistically significant (Sig. = 0.100). However, when compared to other subreddits, 'smallbusinesses' generally display a noteworthy distinction.

Table 4. Pairwise comparisons of subreddits

Subreddit Pairs	**Test Statistic**	**Std. Error**	**Std. Test Statistic**	**Sig.**
Business_Ideas-Entrepreneur RideAlong	−39.561	85.144	−0.465	0.642
Business_Ideas-business	338.803	76.252	4.443	**0.000**
Business_Ideas-smallbusiness	−447.377	71.828	−6.228	**0.000**
Business_Ideas- ecommerce	−556.537	81.854	−6.799	**0.000**
Business_Ideas-Entrepreneur	−720.303	71.828	−10.028	**0.000**
Business_Ideas- Lightbulb	−895.439	226.749	−3.949	**0.000**
Business_Ideas-startups	−1314.855	116.010	−11.334	**0.000**
Business_Ideas-sidehustle	−1345.938	164.862	−8.164	**0.000**
EntrepreneurRideAlong-business	299.242	74.916	3.994	**0.000**
EntrepreneurRideAlong-smallbusiness	−407.817	70.409	−5.792	**0.000**
EntrepreneurRideAlong -ecommerce	516.976	80.612	6.413	**0.000**
EntrepreneurRideAlong-Entrepreneur	680.742	70.409	9.668	**0.000**
EntrepreneurRideAlong-Lightbulb	−855.878	226.303	−3.782	**0.000**
EntrepreneurRideAlong-startups	−1275.295	115.136	−11.076	**0.000**
EntrepreneurRideAlong-sidehustle	−1306.377	164.249	−7.954	**0.000**
business-smallbusiness	−108.574	59.348	−1.829	0.067
business-ecommerce	−217.734	71.155	−3.060	**0.002**
business-Entrepreneur	−381.500	59.348	−6.428	**0.000**
business-Lightbulb	−556.636	223.110	−2.495	**0.013**
business-startups	−976.052	108.725	−8.977	**0.000**
business-sidehustle	−1007.135	159.820	−6.302	**0.000**
smallbusiness-ecommerce	109.159	66.392	1.644	0.100
smallbusiness- Entrepreneur	272.926	53.545	5.097	**0.000**
smallbusiness- Lightbulb	448.062	221.637	2.022	**0.043**
smallbusiness-startups	−867.478	105.670	−8.209	**0.000**
smallbusiness-sidehustle	898.561	157.757	5.696	**0.000**
ecommerce-Entrepreneur	−163.767	66.392	−2.467	**0.014**
ecommerce-Lightbulb	−338.902	225.086	−1.506	0.132
ecommerce-startups	−758.319	112.725	−6.727	**0.000**
ecommerce-sidehustle	−789.401	162.567	−4.856	**0.000**
Entrepreneur-Lightbulb	−175.136	221.637	−0.790	0.429
Entrepreneur-startups	−594.552	105.670	−5.627	**0.000**
Entrepreneur-sidehustle	−625.635	157.757	−3.966	**0.000**
Lightbulb-startups	−419.416	239.629	−1.750	0.080
Lightbulb-sidehustle	−450.499	266.727	−1.689	0.091
startups-sidehustle	31.083	182.171	0.171	0.865

There is a noteworthy adverse gap between 'ecommerce' and 'Entrepreneur' (Sig.=0.014). However, the difference in negativity between 'ecommerce' and 'Lightbulb' is not statistically significant (Sig.=0.132).

No significant negative difference between 'Entrepreneur' and 'Lightbulb' (Sig. = 0.429) were observed. There is a non-substantial favorable variation between 'startups' and 'sidehustle' (Sig. = 0.865).

Statistically speaking, numerous paired samples show notable discrepancies with p-values (Sig.) below 0.05, except for subreddits such as 'Business_Ideas' and 'EntrepreneurRideAlong', 'business' and 'smallbusiness', 'ecommerce' and 'Lightbulb', 'Entrepreneur' and 'Lightbulb' or 'startups' and 'sidehustle' where the p-values surpass 0.05 signifying insignificant differentiation.

In terms of business idea crowdsourcing, this information can provide guidance for deciding which subreddit to choose as a vehicle for content development, marketing plans, and making operational choices in the entrepreneurship and business industries. Recognizing which subreddits have higher levels of community engagement can help in forming the crucial component in crowdsourcing–forming the wise crowd.

B. The Comments of the Posts

Let's look at the number of comments that people have written under each post and analyze whether there are differences in the number of comments depending on the respective subreddits. Table 5 shows the main descriptive statistics for the number of comments, depending on the topic of the post.

Table 5. Comments per posts

Subreddit	N	Mean	Median	Std. Deviation	Minimum	Maximum	Range
business	685	19.02	2.00	57.961	0	571	571
Business_Ideas	384	7.60	5.00	11.460	0	134	134
ecommerce	481	8.68	5.00	10.858	0	90	90
Entrepreneur	998	31.85	9.00	82.318	0	950	950
Entrepreneur RideAlong	406	6.15	3.00	9.900	0	66	66
Lightbulb	30	6.60	4.00	9.690	0	48	48
sidehustle	61	27.08	22.00	23.214	0	102	102
smallbusiness	998	17.60	6.00	40.905	0	819	819
startups	147	30.58	21.00	34.700	0	280	280
Total	4190	18.69	6.00	52.264	0	950	950

Based on the calculated averages, it can be assumed that depending on the subreddit of the post, the number of comments varies. It is also seen that some values of the mean are higher than the median. This suggests that in some subreddits of posts, there are individual posts that have been given a lot of comments. We have formulated the following statistical hypotheses:

H0: The number of comments does not change depending on the subreddit of the post.

H1: The number of comments varies depending on the subreddit of the post.

Before the analysis, the data were checked for the normality of the distribution. The assumption of the normality of the distribution was not confirmed using the Kolmogorov-Smirnov criterion (p-value < 0.05). Therefore, the nonparametric Kruskal-Wallis test will be used for further analysis.

The results of hypothesis testing are presented in Table 6.

Table 6. Independent-samples Kruskal-Wallis test

Independent-Samples Kruskal-Wallis Test Summary	
Total N	4190
Test Statistic	580.811[a]
Degree of Freedom	8
Asymptotic Sig (2-sided test)	0.000

a. The test statistic is adjusted for ties.

The results of the analysis show that the null hypothesis should be rejected (p-value < 0.05) and an alternative hypothesis accepted. The number of comments varies depending on the subreddit of the post.

By conducting pairwise multiple comparisons, we can see which pairs are statistically significantly different (Table 7).

Table 7. Pairwise comparisons of subreddit

Subreddit Pairs	Test Statistic	Std. Error	Std. Test Statistic	Sig.
EntrepreneurRideAlong-business	44.175	75.620	0.584	0.559
EntrepreneurRideAlong-Lightbulb	−158.418	228.430	−0.694	0.488
EntrepreneurRideAlong-Business_Ideas	367.537	85.944	4.276	**0.000**
EntrepreneurRideAlong-ecommerce	478.715	81.369	5.883	**0.000**
EntrepreneurRideAlong-smallbusiness	−809.637	71.070	−11.392	**0.000**
EntrepreneurRideAlong-Entrepreneur	1011.043	71.070	14.226	**0.000**
EntrepreneurRideAlong-sidehustle	−1632.872	165.792	−9.,849	**0.000**
EntrepreneurRideAlong-startups	−1701.842	116.218	−14.644	**0.000**
Business-Lightbulb	−114.242	225.206	−0.507	0.612
Business-Business_Ideas	−323.361	76.968	−4.201	**0.000**
Business-ecommerce	−434.540	71.823	−6.050	**0.000**
Business-smallbusiness	−765.462	59.905	−12.778	**0.000**
Business-Entrepreneur	−966.867	59.905	−16.140	**0.000**
Business-sidehustle	−1588.697	161.321	−9.848	**0.000**
Business-startups	−1657.667	109.746	−15.105	**0.000**
Lightbulb-Business_Ideas	209.119	228.879	0.914	0.361
Lightbulb-ecommerce	320.298	227.201	1.410	0.159
Lightbulb-smallbusiness	−651.219	223.719	−2.911	**0.004**
Lightbulb-Entrepreneur	852.625	223.719	3.811	**0.000**
Lightbulb-sidehustle	−1474.454	269.233	−5.477	**0.000**
Lightbulb-startups	−1543.424	241.880	−6.381	**0.000**
Business_Ideas- ecommerce	−111.179	82.623	−1.346	0.178
Business_Ideas-smallbusiness	−442.100	72.503	−6.098	**0.000**
Business_Ideas-Entrepreneur	−643.506	72.503	−8.876	**0.000**
Business_Ideas-sidehustle	−1265.335	166.411	−7.604	**0.000**
Business_Ideas-startups	−1334.305	117.100	−11.395	**0.000**
ecommerce-smallbusiness	−330.921	67.016	−4.938	**0.000**
ecommerce-Entrepreneur	−532.327	67.016	−7.943	**0.000**
ecommerce-sidehustle	−1154.157	164.095	−7.033	**0.000**
ecommerce-startups	−1223.126	113.784	−10.750	**0.000**
emallbusiness-Entrepreneur	201.406	54.048	3.726	**0.000**
smallbusiness-sidehustle	823.235	159.239	5.170	**0.000**
smallbusiness-startups	−892.205	106.662	−8.365	**0.000**
Entrepreneur-sidehustle	−621.829	159.239	−3.905	**0.000**
Entrepreneur-startups	−690.799	106.662	−6.476	**0.000**
sidehustle-startups	−68.970	183.883	−0.375	0.708

The table shows that most comparison pairs differ statistically significantly (p-value < 0.05). This means that the number of comments that people put to the post varies depending on the subreddit of the post.

However, there are several comparison pairs where these differences are statistically insignificant (p-value > 0.05). No statistically significant differences were found for the following comparison pairs: 'EntrepreneurRideAlong' and 'business'; 'EntrepreneurRideAlong' and 'Lightbulb'; 'business' and 'Lightbulb'; 'Lightbulb' and 'Business_Ideas'; 'Lightbulb' and 'ecommerce'; 'Business_Ideas' and 'ecommerce'; 'sidehustle' and 'startups'. This means that the number of comments does not depend on the subreddit of the post in these comparison pairs.

The standardized test statistics for subreddits such as 'EntrepreneurRideAlong-Entrepreneur' and 'Business-Entrepreneur', with values of 14.226 and −16.140, respectively, indicate substantial differences between the groups; this is supported by a p-value of 0.000 which denotes high statistical significance. Samples such as 'EntrepreneurRideAlong-business' and 'business-Lightbulb' possess p-values exceeding 0.05, suggesting that there exists no significant statistical difference between these pairs of data.

This information can aid in the selection of the subreddit most suitable for the crowdsourcing project.

C. Sentiment Analysis of the Posts

We will examine the sentiment of the posts and analyze whether the differences in the sentiment of the text are dependent on the subreddit (Table 8).

Table 8. Sentiment text

Subreddit	N	Mean	Median	Std. Deviation	Minimum	Maximum	Range
business	685	1.122	0.000	2.198	−5.25	16	21.25
Business_Ideas	384	2.590	2.000	2.799	−2.5	18.45	20.95
ecommerce	481	2.393	1.600	3.483	−2.2	38.75	40.95
Entrepreneur	998	3.628	2.600	4.559	−7.35	51.2	58.55
Entrepreneur RideAlong	406	4.094	3.100	4.500	−2.3	31.05	33.35
Lightbulb	30	2.070	1.225	3.208	−1	16.2	17.2
sidehustle	61	3.018	2.100	4.168	−1.35	27.8	29.15
smallbusiness	998	2.651	2.050	2.973	−5.1	21.95	27.05
startups	147	4.224	3.750	3.091	−2.5	15.5	18
Total	4190	2.795	1.900	3.667	−7.35	51.2	58.55

Based on the calculated average values, it can be assumed that, the sentiment value of the post changes depending on the subreddit. The results show that the average value is greater than the median. This means that in all subreddits the sentiment has more positive values than negative.

Formulating statistical hypotheses:

H0: The value of the sentiment does not change depending on the subreddit.

H1: The value of the sentiment varies depending on the subreddit.

Before the analysis, the data were checked for the normality of the distribution. The assumption of the normality of the distribution was not confirmed using the Kolmogorov-Smirnov criterion (*p*-value < 0.05). Therefore, the nonparametric Kruskal-Wallis test will be used for further analysis. The results of hypothesis testing are presented in the Table 9.

Table 9. Independent-samples Kruskal-Wallis test

Independent-Samples Kruskal-Wallis Test Summary	
Total N	4190
Test Statistic	454.410[a]
Degree Of Freedom	8
Asymptotic Sig.(2-sided test)	0.000

a. The test statistic is adjusted for ties.

The results of the analysis show that the null hypothesis should be rejected (*p*-value < 0.05) and an alternative hypothesis accepted. Therefore, we have proven that the sentiment changes depending on the subreddit as well as the topic of the post.

Table 10 shows that most comparison pairs differ statistically significantly (*p*-value < 0.05). This means that the sentiment of the text differs depending on the subreddit of the article. However, there are several comparison pairs where these differences are statistically insignificant (*p*-value > 0.05). No statistically significant differences were found for the following comparison pairs: 'Lightbulb-ecommerce'; 'Lightbulb- Business_Ideas'; 'Lightbulb-smallbusiness'; 'Lightbulb-sidehustle'; 'ecommerce-Business_Ideas'; 'ecommerce-sidehustle'; 'Business_Ideas-smallbusiness'; 'Business_Ideas-sidehustle'; 'smallbusiness- sidehustle'; 'sidehustle-Entrepreneur'; 'sidehustle-EntrepreneurRideAlong'; 'Entrepreneur-Entrepreneur Ride Along'. This means that the value of the post's sentiment does not depend on the subreddit of these comparison pairs.

Table 10. Pairwise Comparisons of Subreddits

Subreddit Pairs	Test Statistic	Std. Error	Std. Test Statistic	Sig.
Business-Lightbulb	−484.223	225.370	−2.149	**0.032**
Business-ecommerce	−678.536	71.876	−9.440	**0.000**
Business-Business_Ideas	−823.701	77.024	−10.694	**0.000**
Business-smallbusiness	−857.502	59.949	−14.304	**0.000**
Business-sidehustle	−868.350	161.439	−5.379	**0.000**
Business-Entrepreneur	−1091.109	59.949	−18.201	**0.000**
Business-Entrepreneur ridealong	−1150.266	75.675	−15.200	**0.000**
Business-startups	−1522.303	109.827	−13.861	**0.000**
Lightbulb-ecommerce	194.312	227.367	0.855	0.393
Lightbulb-Business_Ideas	339.478	229.047	1.482	0.138
Lightbulb-smallbusiness	−373.279	223.883	−1.667	0.095
Lightbulb-sidehustle	−384.127	269.429	−1.426	0.154
Lightbulb-Entrepreneur	606.885	223.883	2.711	**0.007**
Lightbulb-Entrepreneur RideAlong	666.043	228.596	2.914	**0.004**
Lightbulb-startups	−1038.080	242.057	−4.289	**0.000**
ecommerce-Business_Ideas	145.165	82.684	1.756	0.079
ecommerce-smallbusiness	−178.966	67.065	−2.669	**0.008**
ecommerce-sidehustle	−189.815	164.215	−1.156	0.248
ecommerce-Entrepreneur	−412.573	67.065	−6.152	**0.000**
ecommerce-Entrepreneur ridealong	−471.730	81.428	−5.793	**0.000**
ecommerce-startups	−843.767	113.867	−7.410	**0.000**
Business_Ideas-smallbusiness	−33.801	72.556	−0.466	0.641
Business_Ideas-sidehustle	−44.650	166.533	−0.268	0.789
Business_Ideas-Entrepreneur	−267.408	72.556	−3.686	**0.000**
Business_Ideas-Entrepreneur RideAlong	−326.565	86.007	−3.797	**0.000**
Business_Ideas-startups	−698.602	117.185	−5.962	**0.000**
smallbusiness-sidehustle	10.849	159.356	0.068	0.946
smallbusiness-Entrepreneur	233.607	54.088	4.319	**0.000**
smallbusiness-Entrepreneur RideAlong	292.764	71.122	4.116	**0.000**
smallbusiness-startups	−664.801	106.740	−6.228	**0.000**
sidehustle-Entrepreneur	222.758	159.356	1.398	0.162
sidehustle-Entrepreneur RideAlong	281.915	165.913	1.699	0.089
sidehustle-startups	−653.952	184.017	−3.554	**0.000**
Entrepreneur-Entrepreneur RideAlong	−59.157	71.122	−0.832	0.406
Entrepreneur-startups	−431.194	106.740	−4.040	**0.000**
EntrepreneurRideAlong-startups	−372.037	116.303	−3.199	**0.001**

D. Sentiment Analysis of the Comments

Examining the sentiment indicator for the comments left on posts, we can analyze whether there are differences in the values of the sentiment of the comments, depending on the subreddit of the post (Table 11).

Table 11. Comments sentiment

Subreddit	**N**	**Mean**	**Median**	**Std. Deviation**	**Minimum**	**Maximum**	**Range**
business	12363	0.246	0.000	1.387	−13.30	16.45	29.75
Business_ Ideas	2775	0.980	0.600	1.730	−11.30	21.15	32.45
ecommerce	4044	1.119	0.675	1.811	−6.85	16.10	22.95
Entrepreneur	28972	1.106	0.600	1.927	−11.75	31.55	43.30
Entrepreneur RideAlong	2411	1.278	0.750	1.935	−4.80	18.15	22.95
Lightbulb	198	0.456	0.250	1.149	−2.35	4.85	7.20
sidehustle	1128	1.094	0.750	1.673	−2.00	11.00	13.00
smallbusiness	16789	1.141	0.750	1.988	−9.35	27.85	37.20
startups	4112	1.776	1.150	2.335	−4.40	18.30	22.70
Total	72792	1.005	0.600	1.907	−13.30	31.55	44.85

Based on the calculated average values, it can be assumed that, depending on the subject of the article, the value of the comments sentiment changes. The calculations show that the average value of the indicator is greater than the median. This means that in all subreddits there are more positive comments than negative.

Therefore, we can formulate the following statistical hypotheses:

H0: The value of the sentiment of the comment does not change depending on the topic of the post.

H1: The value of the sentiment of the comment varies depending on the topic of the post.

Before the analysis, the data were checked for the normality of the distribution. The assumption of the normality of the distribution was not confirmed using the Kolmogorov-Smirnov criterion (p-value < 0.05). Therefore, the nonparametric Kruskal-Wallis test will be used for further analysis. The results are presented in Table 12.

Table 12. Independent-samples Kruskal-Wallis test

Independent-Samples Kruskal-Wallis Test Summary	
Total N	72792
Test Statistic	3616.763a
Degree of Freedom	8
Asymptotic Sig (2-sided test)	0.000

a. The test statistic is adjusted for ties.

The results of the analysis show that the null hypothesis should be rejected (p-value < 0.05) and an alternative hypothesis accepted. The value of the sentiment of the comments varies depending on the subject of the post.

By conducting pairwise multiple comparisons, we can see which comparison pairs are statistically significantly different. The results of the check are presented in Table 13.

The table shows that most comparison pairs differ statistically significantly (p-value < 0.05). This means that the sentiment of the comments differs depending on the posts as well as the subreddit. However, there are several comparison pairs where these differences are statistically insignificant (p-value > 0.05).

No statistically significant differences were found for the following comparison pairs: ‘Business_Ideas-sidehustle’; ‘Entrepreneur-sidehustle’; ‘Entrepreneur-ecommerce’; ‘sidehustle-ecommerce’; ‘sidehustle-smallbusiness’; ‘ecommerce-smallbusiness’. This means that the value of the sentiment of the comments does not depend on the subreddits of the posts in these comparison pairs.

In summary, the data provides a robust statistical comparison between different subreddits, but the interpretation would benefit from additional context and possibly, further statistical adjustments for multiple comparisons.

Table 13. Pairwise comparisons of subreddits

Subreddit Pairs	**Test Statistic**	**Std. Error**	**Std. Test Statistic**	**Sig.**
business-Lightbulb	−3158.818	1501.644	−2.104	**0.035**
business-Business_ Ideas	−9801.955	440.342	−22.260	**0.000**
business-Entrepreneur	−10860.694	225.194	−48.228	**0.000**
business-sidehustle	−11169.608	652.011	−17.131	**0.000**
business-ecommerce	−11302.906	379.749	−29.764	**0.000**
business-smallbusiness	−11754.744	248.433	−47.316	**0.000**
business-entrepreneur RideAlong	−13053.500	466.700	−27.970	**0.000**
business-startups	−17771.840	377.375	−47.093	**0.000**
Lightbulb-Business_Ideas	6643.138	1541.994	4.308	**0.000**
Lightbulb-Entrepreneur	7701.876	1494.844	5.152	**0.000**
Lightbulb-sidehustle	−8010.790	1615.229	−4.960	**0.000**
Lightbulb-ecommerce	8144.089	1525.796	5.338	**0.000**
Lightbulb-smallbusiness	−8595.927	1498.521	−5.736	**0.000**
Lightbulb-entrepreneur RideAlong	9894.682	1549.727	6.385	**0.000**
Lightbulb-startups	−14613.023	1525.207	−9.581	**0.000**
Business_Ideas-Entrepreneur	−1058.739	416.562	−2.542	**0.011**
Business_Ideas-sidehustle	−1367.653	740.223	−1.848	0.065
Business_Ideas-ecommerce	−1500.951	516.741	−2.905	**0.004**
Business_Ideas-smallbusiness	−1952.789	429.570	−4.546	**0.000**
Business_Ideas-Entrepreneur RideAlong	−3251.544	583.627	−5.571	**0.000**
Business_Ideas-startups	−7969.885	514.999	−15.476	**0.000**
Entrepreneur-sidehustle	−308.914	636.193	−0.486	0.627
Entrepreneur-ecommerce	442.212	351.898	1.257	0.209
Entrepreneur-smallbusiness	−894.050	203.327	−4.397	**0.000**
Entrepreneur-Entrepreneur RideAlong	−2192.806	444.333	−4.935	**0.000**
Entrepreneur-startups	−6911.146	349.335	−19.784	**0.000**
sidehustle-ecommerce	133.298	705.860	0.189	0.850
sidehustle-smallbusiness	−585.136	644.785	−0.907	0.364
sidehustle-entrepreneur RideAlong	1883.892	756.200	2.491	**0.013**
sidehustle-startups	−6602.232	704.586	−9.370	**0.000**
ecommerce-smallbusiness	−451.838	367.204	−1.230	0.219
ecommerce-Entrepreneur RideAlong	−1750.593	539.378	−3.246	**0.001**
ecommerce-startups	−6468.934	464.254	−13.934	**0.000**
smallbusiness-Entrepreneur RideAlong	1298.756	456.551	2.845	**0.004**
smallbusiness-startups	−6017.096	364.749	−16.497	**0.000**
Entrepreneur RideAlong-startups	−4718.341	537.710	−8.775	**0.000**

VI. Discussion

The present study employed a multi-faceted approach to analyze community-driven sentiments related to business ideas across nine distinct subreddits. Utilizing both automated data extraction and manual curation, the methodology ensured a dataset of high reliability and relevance. Advanced text pre-processing techniques were applied to this dataset to prepare it for sentiment analysis, which was conducted using the syuzhet package in R. Subsequently, pairwise hypothesis testing was conducted to evaluate the statistical differences in sentiments across different subreddits.

Most subreddit comparisons uncovered distinct disparities in sentiment scores, as shown by p-values under 0.05. For instance, the ‘Entrepreneur-startups’ and ‘Business_Ideas-startups’ pairs displayed significant differences in tone. These variations could be attributed to the divergence of conversation focus within these forums: high impact technological innovations are often discussed on “startups,” while broader low-cost business concepts may feature heavily on “Entrepreneur” and “Business Ideas.” Nevertheless, specific combinations of subreddits such as ‘Business_Ideas-sidehustle’ and ‘Entrepreneur-ecommerce’, did not exhibit noteworthy variations. The probable explanation for this observation is the convergence in subject matter between these subreddit pairs.

By introducing a highly reliable and adaptable methodology designed for subreddit data, this study makes significant progress in the field of sentiment analysis. Moreover, by showcasing how attitudes towards business concepts can vary considerably amongst online communities, it expands on previously established insights.

Although this study has numerous strengths, it is important to acknowledge some limitations. External factors, such as economic conditions or global events that could have potentially impacted the opinions and attitudes were not considered in the study thereby creating confounding variables. Despite the use of multiple comparison corrections, there is still a possibility for false positives and false negatives to occur.

Future research could focus on longitudinal sentiment analysis to capture temporal variations and explore the influence of external events on subreddit sentiments.

The research effectively displays the diversity of opinions on business concepts among various Reddit communities, highlighting the essentiality of utilizing community perspectives in conceiving ideas. Additionally, it presents a systematic approach that can be employed for comparable studies on different social media platforms.

VII. Conclusion

Our study provides a comprehensive outlook on public perception within various interconnected business and entrepreneurship subreddits, filling an existing void in current literature. The ramifications of our findings extend beyond just prospective entrepreneurs, but also to investors, educators and policymakers who seek to comprehend the potential impact community sentiment may have on business trends and decisions.

In addition, our approach utilizing intensive data cleansing and cutting-edge statistical methods adds to the academic discussion on analyzing sentiment in online platforms. This signifies progress beyond previous studies that frequently failed to achieve rigorousness within their methodologies.

Although rigorous statistical measures were utilized, limitations do exist within this study. The research is limited solely to Reddit usage, it may not be applicable or representative of other online platforms or larger populations. Additionally, incorporating additional factors such as posting frequency, engagement level or more detailed demographic data could offer a fuller understanding of the phenomenon under scrutiny.

The application of NLP and sentiment analysis in evaluating public perspectives on various business and entrepreneurship platforms on Reddit has been shown to be valuable, according to this study. Although the findings demonstrate that there is a fluctuation in people’s sentiments towards different business topics, it also brings attention to areas where emotions are uniform, thereby deepening our comprehension of online conversations within the entrepreneurial setting.

This study paves the way for future research and represents a crucial milestone in revolutionizing how big data and machine learning algorithms are utilized to capture entrepreneurial sentiment in today’s digital era.

Conflict of Interest

The author declares no conflict of interest.

References

Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. 2014. How to ask for a favor: A case study on the success of altruistic requests. ***Proceedings of ICWSM***.

Antweiler, W., & Frank, M. Z. 2004. Is all that talk just noise? The information content of internet stock message boards. ***Journal of Finance***, 59(3).

Brabham, D. C. 2008. Crowdsourcing as a model for problem solving. ***Convergence: The International Journal of Research into New Media Technologies*** DOI: 10.1177/1354856507084420.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. 2013. New avenues in opinion mining and sentiment analysis. ***IEEE Intelligent Systems***, 28(2).

Cannon, E. 2022. An exploration of reddit’s advice communities. ***ScholarWorks at SJSU***.

Chu, B. 2020. Patient crowdsourcing of dermatologic consults on a social media platform. ***NCBI***, PMC7411418.

Cohen, J. 1988. ***Statistical power analysis for the behavioral sciences*** (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. 2013. ***Applied multiple regression/correlation analysis for the behavioral sciences.*** Routledge.

Davidson, C. 2023. Use of reddit for social science research. ***ScholarWorks at WMU***.

De Choudhury, M., & De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. ***Proceedings of ICWSM***.

Estellés-Arolas, E. & González-Ladrón-De-Guevara, F. 2012. Towards an integrated crowdsourcing definition. ***Journal of Information Science,*** 32(2).

Field, A. 2018. Discovering statistics using IBM SPSS Statistics. ***SAGE Publications Limited***.

Howe, J. 2006. The rise of crowdsourcing. ***Wired***, 14. Available: http://www.wired.com/wired/archive/14.06/crowds.html (Retrieved through proxy at: https://12ft.io/proxy?q=http%3A%2F%2Fwww.wired.com%2Fwired%2Farchive%2F14.06%2Fcrowds.html).

Jamnik, M. R. 2017. The use of reddit as an inexpensive source for high-quality data***. ScholarWorks at UMass***.

Kaplan, A. M., & Haenlein, M. 2010. Users of the world, unite! The challenges and opportunities of social media. ***Business Horizons***, 53(1).

Lai, D. 2020. Addressing immediate public coronavirus (COVID-19) concerns through social media: A case study with Reddit. ***PLOS ONE***, DOI: 10.1371/journal.pone.0240326.

Liu, B. 2012. Sentiment analysis and opinion mining***. Synthesis Lectures on Human Language Technologies***, 5(1).

Loughran, T., & McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. ***The Journal of Finance***, 66(1).

Luong, R. 2022. Evaluating Reddit as a crowdsourcing platform for various research projects. ***Sage Journals***, DOI: 10.1177/00986283211020739.

Manning, C. D., Raghavan, P., & Schütze, H. 2008. Introduction to information retrieval. ***Cambridge University Press***.

Massa, P., & Avesani, P. 2015. Controversial users demand local trust metrics: An experimental study on epinions.com community. ***Proceedings of the National Academy of Sciences***, 112(8).

Pang, B., & Lee, L. 2008. Opinion mining and sentiment analysis. ***Foundations and Trends in Information Retrieval***, 2(1–2).

Paul, M. J., & Dredze, M. 2011. A model for mining public health topics from Twitter. ***Health***, 11(16).

Rivera, I. 2021. RedditExtractor: A Python package for extracting Reddit data. ***GitHub Repository***. Available at: https://github.com/ivan-rivera/RedditExtractor.

Surowiecki, J. 2004. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. ***Doubleday & Co***.

Thompson, C. M. 2022. The case of COVID-19 long-haulers: Drawing on uncertainty management theory. ***NCBI***, PMC8805953.

Zhang, L., Wang, S., & Liu, B. 2018. Deep learning for sentiment analysis: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, 8(4).