information

MDPI

Article

Prominent User Segments in Online Consumer Recommendation Communities: Capturing Behavioral and Linguistic Qualities with User Comment Embeddings

Apostolos Skotis [1,*] and Christos Livas [2]

[1] Department of Business Administration, University of Piraeus, 18534 Piraeus, Greece
[2] Department of Business Administration, University of Patras, 26504 Rio, Greece; clivas@upatras.gr
* Correspondence: askotis@unipi.gr

Abstract: Online conversation communities have become an influential source of consumer recommendations in recent years. We propose a set of meaningful user segments which emerge from user embedding representations, based exclusively on comments' text input. Data were collected from three popular recommendation communities in Reddit, covering the domains of book and movie suggestions. We utilized two neural language model methods to produce user embeddings, namely Doc2Vec and Sentence-BERT. Embedding interpretation issues were addressed by examining latent factors' associations with behavioral, sentiment, and linguistic variables, acquired using the VADER, LIWC, and LFTK libraries in Python. User clusters were identified, having different levels of engagement and linguistic characteristics. The latent features of both approaches were strongly correlated with several user behavioral and linguistic indicators. Both approaches managed to capture significant variability in writing styles and quality, such as length, readability, use of function words, and complexity. However, the Doc2Vec features better described users by varying level of contribution, while S-BERT-based features were more closely adapted to users' varying emotional engagement. Prominent segments revealed prolific users with formal, intuitive, emotionally distant, and highly analytical styles, as well as users who were less elaborate, less consistent, but more emotionally connected. The observed patterns were largely similar across communities.

Keywords: user segmentation; recommendation communities; information extraction; user embeddings; neural language models; behavioral engagement; linguistic processes

check for updates

Citation: Skotis, A.; Livas, C. Prominent User Segments in Online Consumer Recommendation Communities: Capturing Behavioral and Linguistic Qualities with User Comment Embeddings. *Information* **2024**, *15*, 356. https://doi.org/10.3390/info15060356

Academic Editor: Wei Zhang

Received: 1 May 2024
Revised: 6 June 2024
Accepted: 12 June 2024
Published: 15 June 2024

1. Introduction

Online consumer communities have dramatically evolved in recent years, urging companies to strengthen efforts for extracting actionable insights from user activity and content. Profiling user types or personas based on online behavioral attributes and user-derived content can enhance the understanding of consumer needs and trends [1]. While traditional approaches such as psychographic surveys can be time-consuming, organizations exhibit a growing tendency to utilize online data to reveal behavioral groupings and develop methods to move from raw data complexity to actionable, interpretable outputs, which has become a key challenge for consumer analytics [2]. Several approaches have been observed in recent studies attempting to uncover user roles in online consumer communities [3]. Participation in consumer communities has been analyzed through social exchange and user role theories, explaining multiple aspects of the interactions taking place, undertaking of specific roles, user motivations, and mutual gains received through participation [4,5]. Recent research indicates that a series of moral and social legitimacy criteria emerge strong in consumer conversation communities such as Reddit, challenging marketers' reliability [6]. The growing influence of recommendation communities is highlighted in a recent report [7] by Reddit: While individual influencers in social networking platforms are predominantly viewed as unreliable, a majority of Reddit users (78%) seek recommendations from strangers in conversation communities on a monthly basis, and the vast majority of users are satisfied by these recommendations and willing to consider new ideas or products due to the received suggestions (78% and 75%, respectively). Therefore, the identification of behavioral user groupings is a crucial step toward identifying behavioral patterns and user groups in the process of influential online recommendations.

The writing style characteristics of people can uncover various aspects of their underlying psychological and behavioral patterns, with recent studies placing greater emphasis on how a person writes, not just on what they write (content). A growing number of methodologies and metrics aid experts in this direction [8,9]. Ref. [10] claims that it is feasible to identify truthful responses based solely on linguistic properties. The use of function words has been found to be related to certain psychological traits and social processes of authors [11]. Using methods estimating readability, that is, the difficulty of text comprehension, have shown positive relationships with user engagement and familiarity in social media [12]. Capturing language characteristics and peculiarities is not an easy task, and recent studies place a greater emphasis on the utilization of advanced neural language models, which are commonly used for resolving natural language processing issues [13]. Several recent studies emphasize that readability and text difficulty features can be influential in many contexts. Systems of readability assessment have been proposed, aiming at text simplification and facilitation of learning tasks, in which lexical and syntactic features have a crucial part [14–16]. Readability measurements can be a key factor in meeting people's information needs, and new methods of deep learning neural network mechanisms have been introduced in models of text readability measurement, to improve performance [17].

In contrast to using pre-defined features, high-dimensional neural network models producing embeddings at multiple levels (word, sentence, document, or author) have the edge in finding novel patterns and representing latent features learned directly from textual data [18,19]. Techniques of learning sentence embeddings have been used for capturing linguistic patterns in social media data [20], assessing language quality and correctness, as well as in performing assessments of the preservation of linguistic properties [21]. In a similar manner, embeddings at the user level, representing users as vectors in semantic space, are used to support user analysis tasks such as preferences or personality modeling [19,22]. Neural network deep learning models are increasingly preferred for discovering user personality traits compared to approaches of machine learning using manually extracted features, in terms of performance, time consumption, and ability to capture hidden patterns [23]. Deep learning approaches have also been introduced to overcome the shortcomings of latent factor models in collaborative recommendation systems, which suffer from sparsity problems and a limited ability to extract non-linear features [24]. User embedding methodologies can achieve high performance in author classification tasks, accounting for a wealth of linguistic variability of individual users, yet with a significant drawback; most approaches are finding it hard to determine the degree in which embeddings capture information regarding topical, sentiment, or writing style. Therefore, the evaluation of embedding quality can often be a difficult task [25,26]. A common approach in addressing issues of explainability is using specialized extrinsic evaluation metrics [27]. Additionally, established consumer clustering methods are gradually enriched by using high-dimensional learned customer features in the form of embeddings, using natural language processing technologies [28]. Applying clustering techniques to neural embedding features can achieve high performance standards in a variety of tasks [27].

Acquiring author segments in social network environments can be very useful for assessing online interactions [29], taking into consideration that users can be associated with certain writing styles, instead of topics [26]. Formation of user community interactions can therefore be based on content, as well as on crucial linguistic features such as readability levels and language quality [30]. Acknowledging the limited number of studies on user segmentation in consumer recommendation communities, this study applies two common neural language methods to produce user embeddings, based on comments posted in three suggestion communities in Reddit: Booksuggestions, Suggestmeabook, and Moviesuggestions. Using data spanning a whole year (2021) with 1.3 million comments, we focus on using the learned embeddings to acquire reduced principal factors and examine their capacity to capture variations of linguistic, sentiment, and other behavioral engagement factors. As a next step, we use components to cluster users and assess the acquired segments based on differences in behavior and linguistic traits. Our effort enhances the understanding of users' consumer advice behaviors and the nature of their contributions influencing decision making.

The rest of this paper is organized as follows. Section 2 covers the relevant literature and associated research questions. Section 3 presents the methods followed in detail. Section 4 demonstrates the results and analysis. Section 5 contains a discussion of the main points, as well as the research implications, limitations, and future research suggestions.

2. Relevant Studies and Formulation of Research Questions

Although recent research corroborates the high prediction performance of user embedding features, interpretation of these latent features remains a challenging task. Hence, gaining insight into user behaviors is impacted significantly [19]. Many efforts comprising the training of embeddings in the semantic space are focused on limited tasks and lack a multi-dimensional evaluation of latent features [26]. Several studies have utilized sentence embeddings to predict linguistic phenomena such as sentence lengths, verb tense, or word occurrences [31]. However, explainability and the identification of aspects captured by embeddings at the sentence or user level is usually limited to topic recognition, which is a common anticipation. However, recent studies urge for an increased emphasis on how latent features reflect user linguistic qualities, writing styles, and are related to linguistic structures. Selecting several independent linguistic metrics to perform evaluations is one of the proposed approaches [26,31,32].

Consumer engagement in the online conversation landscape comprises cognitive, emotional, and behavioral concepts, and materializes in several user actions, such as the multitude of contributions, likes, etc. These materializations that are used as measures of engagement can vary between platforms and domains, depending on their structure and context. Therefore, definitions in recent studies can differ [33–35]. Relationships of engagement metrics in social media and text properties of user contributions have been examined by several studies. Characteristics related to text complexity, readability, length, and ease of comprehension can crucially affect message perceptions, popularity, and interactions. Also, engagement phenomena have been associated with writing styles which differ by intensity, depending on the used mix of parts-of-speech, as well as language quality [12,30,33].

Linguistic features and writing styles of users can be indicative of personality aspects and their undertaken roles in the online community. The expression of emotions is associated with higher resilience [36], while tendencies to use certain word categories or parts-of-speech categories can be related to writing styles of formal or informal nature. An increased use of personal pronouns and verbs is indicative of emotional and informal writing styles, while an increased number of prepositions are present in literary or descriptive texts [37]. Ref. [11] associates writing styles with corresponding thinking style dimensions, distinguishing categorical and dynamic thinkers. People of formal and analytical thinking are characterized by emotional distance and hierarchical writing styles, making use of more prepositions and articles. On the other hand, dynamic thinkers demonstrate more socially driven, informal writing tendencies.

Ref. [38] uses the word categories of the Linguistic Inquiry and Word Count (LIWC) library to investigate associations with the Big Five personality traits. Significant correlations with personality dimensions are found regarding the use of pronouns, positive and negative emotions, prepositions, and verb tense. Ref. [9] indicates that language-centered personality research has identified specific patterns regarding extroverted and introverted personalities, with the former exhibiting higher proportions of social words, positive emotions, and first-person pronouns. Similar results are reported by [39], who also emphasize the use of longer words and complexity by personalities of higher intuition. Finally, user texts can be revealing of emotional intelligence traits. Ref. [40] found positive relationships between higher levels of emotional intelligence and the use of positive emotions as well as social-oriented words in social media texts. Advanced neural language methods for the analysis, identification, and classification of individuals' behaviors are an observed trend in many recent studies. The use of approaches like the BERT architecture and similar NLP methods are increasingly used in psychology studies for the classification of behavioral transcriptions, and the analysis of human interactions. The advantages of these methods are emphasized, compared to the use of linguistic and lexical resources like the LIWC library [41,42].

In light of the above, we aim to examine the extent to which user comment embeddings of two suggested neural language methods from the literature (based on Doc2Vec and Sentence-BERT, respectively) can explain variations in user behaviors and writing styles in the context of online recommendation communities. Therefore, we formulate the following research question:

- RQ1: What aspects of user engagement behavior and linguistic characteristics can be sufficiently captured by latent factors of user comment embeddings in consumer recommendation communities?

Language use and style in online environments can be a defining characteristic for author and user profiling. Ref. [43] analyzes the language style of long-term participants in online discussion communities, finding that there is a gradual development of informal, familiar, and emotional language styles among users. However, the authors indicate that language adaptation in different environments and domains can develop different trends, depending on the context. In more recent studies, writing style features are the target of embedding representation efforts for author profiling. Ref. [44] embeds documents in the stylometric space using deep neural models to cluster stylistic clues to reflect authorship in online texts. In a similar direction, ref. [45] uses a BERT-based method to obtain author representations with the aim of identifying writing styles in social media.

Segmenting users according to emerging roles in online conversation communities is commonly approached through the utilization of activity and text content metrics. Ref. [29] assesses the presence of author stereotypes in online conversations according to posting frequency, conversational patterns, and expressions of positivity, negativity, and neutrality. Through this process, user roles such as systematic "answer-persons" or "explorers" are formed. Language features can be used to identify groups of experts, varying by readability levels, text lengths, and lexical categorizations [46]. The ways and styles of communication characterizing distinct user groups can highlight significant aspects of their common values, norms, and psychological trends. This approach puts greater emphasis on communication styles as explaining factors, rather than relying on the differences of expressed topics [47]. Following a similar logic, we formulate the following question:

- RQ2: What are the prominent user segments in consumer recommendation communities, based on user comment embeddings:
  - ○ RQ2a: distinguished by factors of behavioral engagement?
  - ○ RQ2b: distinguished by factors of writing style and quality?

3. Methods

3.1. Data Collection and Preperation

A recent report published by Reddit [7] indicates the growing importance of consumer recommendation communities, which is confirmed by users' levels of preference and trust in conversation communities as sources of advice and suggestions. Reddit conversation communities are organized in a thread-like architecture, allowing for the development of discussion trees. Sub-forums in Reddit are topic-specific (named "subreddits") and users can make submissions (opening a new discussion) and comments. Comments can be replies to a submission or to other comments in the thread. A special consumer-related topic covered by many communities is the one of suggestions, where a submission is posted by a user seeking suggestions from other users. We selected three subforums following a similar discussion logic, namely r/Booksuggestions, r/Moviesuggestions, and r/Suggestmeabook, covering the domains of book and movies suggestions. We used data of submissions, comments, and available metadata regarding the three communities, covering a year, from January to December of 2021. Data were retrieved using the Pushshift dataset [48] from its provided monthly data dumps, which has been consistently used by academic research for its completeness, accuracy, and efficiency [49]. Ref. [50] indicates that using Reddit as a source of data for academic research can pose challenges regarding the generalizability of findings and models, in contexts outside the platform. This could be partly attributable to Reddit demographics, which are predominantly male and belonging to young age groups [51]. However, its characteristics of self-governing nature, accessibility, and richness as a source of complex user public interactions and deep context can be very advantageous for academic studies [39,52]. Reddit's discussion structure has been described as having an edge over other social media platforms regarding the expression of consumer needs [53]. Studies have used suggestions in subreddits to perform deep learning tasks [54]. Additionally, user data in Reddit contains no demographic data and are anonymous. Both submissions and comments can be upvoted or downvoted by other users, producing a net positive or net negative score per contribution.

All tasks regarding data management, cleansing, text pre-processing, retrieval and calculation of evaluation variables, training of embeddings, and dimension reduction were performed in Python (version 3.11.7) using the following modules, packages and libraries: numpy (version 1.24.3), pandas (version 2.1.4), os (built-in module) , re (version 2.2.1), zstandard (version 0.22.0), json (version 2.0.9), gensim (version 4.3.0), nltk (version 3.8.1), spacy (version 3.7.2), sklearn (version 1.3.0), tqdm (version 4.65.0), multiprocessing (version 0.70.15), sentence_transformers (version 2.2.2), vaderSentiment (version 3.3.2), liwc (version 0.5.0), lftk (version 1.0.9). Pearson correlation tables, cluster analysis, and statistics by cluster were performed in IBM SPSS (version 25). We ran the analysis for each of the three communities separately. Datasets of submissions and comments were joined to acquire complete historical data for each thread. We then performed data cleansing by identifying and removing bots and moderator posts, similar to other studies [53] and removing comments not having a corresponding submission (due to occurring outside the 1-year interval). Table 1 presents the basic sample information after data cleansing was performed. After performing the necessary transformations (date conversions, calculation of month and day variables, etc.), we created datasets aggregated at the user level, calculating a series of variables to be used for evaluating the acquired embeddings at a subsequent step.

Table 1. Sample Information.

	Subreddit (Community)			
	r/Booksuggestions	**r/Moviesuggestions**	**r/Suggestmeabook**	**Total**
Comments	303.881	325.955	680.351	1310.187
Threads	29.510	17.576	47.218	94.304
Unique users with min 1 comment	56.138	40.520	119.749	216.407
Unique users >3 comments	15.714	12.490	32.054	60.258

3.2. Evaluation Variables

Using source data (Reddit dataset), we calculated variables related to user activity and engagement. Specifically, we created the following metrics at the user level: length of activity (unique months), length of activity (unique days), average score of comments, average number of replies created by each user comment, aggregate score of comments, aggregate replies created by user comments, total number of comments, number of unique threads participated, number of comments in own submissions (in threads created by the user). Additionally, we complemented the engagement assessment variables by retrieving sentiment scores for each comment using the VADER library [55]. Scores regarding positive, negative, and neutral sentiment, as well as a compound score (normalized, weighted composite score) were retrieved for each comment. Averages were calculated for each user. The VADER model has been used extensively in recent research for sentiment analysis purposes [20].

We also used the relatively new LFTK library [56] to calculate several linguistic metrics for each comment. The LFTK library offers several metrics, but we opted for the variables most closely associated with readability and essay scoring assessment, according to the creators' validation results. Readability in particular has been described as the ease or extent a text is understandable and comprehensible for the reader and is closely dependent on writing style [12]. The variables selected include number of words and number of unique words, average syllables per sentence, average characters per word, average syllables per word, average Kuperman age of acquisition per word, reading time indexes (for fast and slow readers), Coleman–Liau Readability Index, and SMOG Readability Index. In addition, we used the LIWC library [37] to obtain a series of metrics referring to linguistic and psychological processes, for each comment. The LIWC dictionary has been widely used in a variety of text analysis tasks, including sentiment analysis [57], analysis of group linguistic styles [47], and personality prediction [39]. Given that the LIWC library in Python returns counts of word occurrences per document (in our case, per user comment), we divided occurrences of each category by number of words, to neutralize the effect of text length. Descriptions of all the variables used for the evaluation are presented in Appendix A.

3.3. User Embeddings, Dimension Reduction, and Clustering Approach

User embeddings based on text contributions in social media are learned representations that condense text information associated with a single user in vector coordinates in the semantic space [25]. Using high-dimensional models compared to low-dimensional models (based on pre-defined variables) has been found to achieve consistently better performances within domains [18]. We used two different methods to train user-level embeddings. Doc2Vec [58] is based on the Word2Vec method but extends it so that it learns the semantic representation of documents instead of single words, connecting both paragraph and word vectors. It has been used in several studies for a variety of machine learning tasks, among which is content prediction [59–61]. To acquire user embeddings, we used a special functionality of the Gensim Doc2Vec library (named Tagged-Document) in Python to assign the user class in place of document id. We trained the model in 300 dimensions for 30 epochs (iterations) selecting the DBOW method [19,62]. As an alternative approach, we also used a deep learning language model based on the Sentence-BERT architecture [63] using the transformers technique, to obtain sentence level embeddings. We specifically used the "all-MiniLM-L6-v2" pre-trained, general purpose model, (www.sbert.net/docs/pretrained_models.html, accessed on 3 February 2024) which achieves good performance and overall offers good quality results. The model returns embeddings in 384 dimensions. We then averaged the embeddings of each user to obtain embeddings at the user level [64].

As a next step, we performed dimensionality reduction using the method of principal components analysis (PCA). Condensing information to two dimensions can facilitate evaluations using external variables and the subsequent clustering of users. Given the linear transformation nature of the method, we also examined using the non-linear dimensionality reduction approach of UMAP (uniform manifold approximation and projection), commonly used with high-dimensional sentence embeddings, but the clustering results using various clustering methods were not satisfactory. Therefore, we proceeded with using PCA retaining the first two dimensions, which for each community and approach accounted for approximately 25% of the total variation explained by 300 (Doc2Vec) and 384 dimensions (Sentence-BERT), respectively. We subsequently loaded the datasets to SPSS. Pearson correlation tables were created between principal factors and evaluation variables. According to the correlation results, depending on the method (Doc2Vec–S-BERT), the reduced factors captured a high proportion of variability of our dimensions of interest (sentiment, linguistic, and behavioral). Therefore, we proceeded with using principal components of each method to perform cluster analysis using K-Means. Given the nature and purpose of the study, we followed a simplified approach to select the optimal number of clusters. To retain an explainable number of segments, we considered only cases of 3 to 7 clusters. For these cases, we calculated Silhouette scores. In all cases and for both embedding methods, Silhouette scores indicating the best separation (Score >= 0.5) were achieved for 3 clusters. Additionally, we performed a Bonferroni post hoc test of pairwise tests of mean differences between each cluster, and for each variable, found statistically significant results (p-value < 0.00) in all cases, indicating a strong cluster separation. Finally, tables of mean differences between clusters including all evaluation variables were created, to facilitate the interpretation of each cluster and the identification of the main differences between user segments. The significance of differences was evaluated through ANOVA statistic values and effect size of differences through eta-squared estimates. The final datasets used for analysis, containing transformed data at the user level, including PCA factors and cluster memberships can be found in Supplementary Materials. The provided Excel Workbook contains Datasets S1, S2 and S3 in separate spreadsheets, corresponding to each of the three communities. As an extra validation measure, we also performed K-Means without prior dimensionality reduction for the Doc2Vec model to test if the formed clusters were consistent with our main approach. We analyze the results in the next section.

4. Results

Using an extensive set of behavioral and linguistic evaluation metrics (see Appendix A for full descriptions) our first goal was to examine their correlations with the reduced dimensions given by principal components analysis for each method. The PCA method follows an assumption of linear transformation; therefore, observing significant correlations with variables used for the evaluation of learned semantic embeddings should be an indicator that factors retain sufficient variability and information, at least regarding our aimed user analysis dimensions. Also, a secondary but important objective of our study was to understand the differences between the outputs of the user embedding neural methods of Doc2vec and Sentence-BERT. Both are recommended methods for learning user embeddings based on user input texts. However, the two approaches are vastly different in their architecture. Compared to models based on the Word2Vec or Doc2Vec methods, BERT-based deep learning approaches achieve multiple levels of contextualization and are considered a state-of-the-art approach for various tasks [26]. Nevertheless, the usability of delivered representations from both methods can be highly case-specific.

4.1. Correlations of Principal Factors with Evaluation Variables

Table 2 demonstrates Pearson correlations regarding reduced factors of the Doc2Vec approach by community. Only evaluation variables with large and moderate correlations are included for the purposes of analysis. A range of variables exhibit strong correlations with principal factors, capturing a significant part of the behavioral and linguistic variability of users. The results follow similar patterns in all recommendation communities, which is an observation corroborating the selected methods. The Doc2Vec approach appears to be producing representations closely related to behaviors of activity, regarding volume of participation and length of engagement. This is evident by the large effects observed in the number of days active variable, which are 0.65, −0.65, and −0.62, respectively. Strong correlations (>0.5, <−0.5) are also observed for number of months active, number of comments, and aggregate comment score (upvotes minus downvotes). Therefore, factors derived from Doc2Vec embeddings can describe variations of user representations between different levels of activity. In other words, more active users that post more comments and engage for longer time periods with the community have significantly different representations based on their comments' content. Additionally, there is an opposite relationship of engagement variables with positive sentiment, which is observed by the moderate effect correlations having opposite signs (−0.37, 0.33, and 0.32). Regarding writing style and quality metrics, strong correlations of variables related to readability and language quality indicate that the factors are also representative of many aspects of users' linguistic differences. Factors can capture significant differences evaluated by metrics of total unique words per comment (0.56, −0.53, 0.64), reading time indexes (>=0.5, <=−0.5), and average syllables per sentence (0.50, −0.41, 0.51). Additionally, moderate effects are observed for the SMOG readability index and average use of common verbs. Overall, it is evident that factors based on the Doc2Vev method are significant descriptors of posting volume behaviors and of linguistic characteristics related to length and comments' comprehension and overall quality. This observation is not very common in the relevant literature, in which interpretations and analysis of semantic representations are usually topic-based. The range of large effects across various dimensions outlines the wealth of behavioral information captured by user embeddings.

Table 2. Pearson Correlations * of Principal Factors ** with Evaluation Variables–Doc2Vec approach.

	r/Booksuggestions		**r/Moviesuggestions**		**r/Suggestmeabook**	
Behavioral Engagement	PCA 1	PCA 2	PCA 1	PCA 2	PCA 1	PCA 2
Activ Length Months	0.57	−0.05	−0.50	0.08	−0.53	0.48
Activ Length Days	0.65	0.07	−0.65	0.07	−0.62	0.53
Sum Comment Score	0.53	0.05	−0.51	0.03	−0.47	0.44
Sum Replies Created	0.51	0.05	−0.53	−0.01	−0.45	0.41
Sum Comments	0.51	0.08	−0.49	0.02	−0.47	0.42
Sum Unique Threads	0.51	0.08	−0.45	0.06	−0.45	0.39
Comp Sentim Score	−0.09	−0.15	0.01	−0.43	0.07	0.09
Pos Sentim Score	−0.37	0.05	0.16	−0.33	0.32	−0.23
Neu Sentim Score	0.34	−0.05	−0.12	0.28	−0.28	0.19
Affect	−0.32	0.08	0.18	−0.16	0.28	−0.22
Pos Emotion	−0.34	0.08	0.18	−0.21	0.29	−0.24
Linguistic Style						
Tot Word	0.52	−0.12	−0.50	−0.29	−0.47	0.59
Tot Un Word	0.56	−0.15	−0.53	−0.33	−0.50	0.64
Avg Syll PS	0.50	−0.10	−0.41	−0.27	−0.46	0.51
Avg Kup AoA PW	−0.08	−0.16	0.06	−0.49	0.08	0.21
RT_Fast	0.52	−0.12	−0.50	−0.29	−0.47	0.59
RT_Slow	0.52	−0.12	−0.50	−0.29	−0.47	0.59
SMOG Index	0.38	0.04	−0.32	−0.18	−0.35	0.41
Pers Pronouns	−0.19	−0.04	0.05	−0.42	0.18	0.00
1st Pers Singular	−0.20	−0.10	0.05	−0.43	0.18	−0.01
Common Verbs	−0.30	0.03	0.11	−0.44	0.25	−0.08
Past Tense	−0.04	−0.11	0.02	−0.32	0.05	0.08
Present Tense	−0.28	0.05	0.11	−0.33	0.24	−0.11
Adverbs	−0.05	−0.12	−0.03	−0.37	0.05	0.13
Conjunctions	0.19	−0.19	−0.15	−0.31	−0.13	0.28
Cognitive	0.16	−0.16	−0.11	−0.39	−0.10	0.34
Exclusive	0.10	−0.19	−0.08	−0.32	−0.06	0.25

*: Only cases with moderate (0.3 > x < 0.5) and strong (x > 0.5) Pearson correlations are presented. **: PCA 1, PCA 2 are the first two reduced dimensions of principal components analysis.

A similar range of strong effects can be observed in Table 3, demonstrating correlation coefficients of reduced factors based on the S-BERT method. However, regarding behavioral engagement, S-BERT representations appear to be more adjusted to variations in user sentiment. The factors are strongly correlated to indicators of positive sentiment, neutral sentiment, and compound sentiment (the latter being the overall sentiment score per comment, accounting for positive, negative, and neutral score values). Pearson correlations range from 0.55 to 0.73 for the three aforementioned indexes, while it appears that negative sentiment is more moderately associated (~0.3). Additional emotion metrics of affective words and positive emotion words exhibit similarly strong relationships. It would therefore be valid to assume that S-BERT-based factors can account for a very large part of user sentiment variability. Moreover, components appear to be able to describe an extensive range of linguistic differences. Metrics related to readability and text length are strongly correlated (number of unique words, average syllables per sentence) while components seem to strongly reflect the use of verbs and tenses. A series of moderate effects are also observed regarding readability indexes (SMOG, reading times) and writing style indicators such as personal pronouns, articles, prepositions, conjunctions, and cognitive process words. These relations and corresponding directionalities along with emotional indicators can represent rich patterns of behavioral and linguistic properties of users, which are analyzed in more detail in the next sections.

Table 3. Pearson Correlations * of Principal Factors ** with Evaluation Variables–S-BERT approach.

	r/Booksuggestions		**r/Moviesuggestions**		**r/Suggestmeabook**	
Behavioral Engagement	**PCA 1**	**PCA 2**	**PCA 1**	**PCA 2**	**PCA 1**	**PCA 2**
Num of Com Own Threads	0.29	−0.19	0.00	0.37	0.25	0.23
Comp Sentim Score	0.10	−0.60	−0.30	0.58	0.02	0.62
Pos Sentim Score	0.65	−0.44	−0.02	0.73	0.55	0.54
Neu Sentim Score	−0.61	0.35	0.05	−0.59	−0.51	−0.44
Neg Sentim Score	−0.21	0.31	−0.05	−0.29	−0.17	−0.30
Affect	0.60	−0.24	0.13	0.47	0.52	0.33
Pos Emotion	0.63	−0.30	0.12	0.60	0.55	0.40
Linguistic Style						
Tot Word	−0.43	−0.19	−0.58	0.05	−0.45	0.20
Tot Un Word	−0.51	−0.23	−0.67	0.08	−0.52	0.25
Avg Syll PS	−0.59	−0.11	−0.66	0.00	−0.58	0.10
Avg Kup AoA PW	0.06	−0.38	−0.28	0.28	0.01	0.41
RT Fast	−0.43	−0.19	−0.58	0.05	−0.45	0.20
RT Slow	−0.43	−0.19	−0.58	0.05	−0.45	0.20
SMOG Index	−0.37	−0.14	−0.45	0.00	−0.37	0.13
Pers Pronouns	0.22	−0.30	−0.19	0.44	0.17	0.36
1st Pers Singular	0.06	−0.24	−0.28	0.32	0.02	0.29
2nd Pers	0.32	−0.19	0.06	0.36	0.30	0.24
Articles	−0.35	0.15	−0.07	−0.28	−0.27	−0.22
Common Verbs	0.56	−0.39	−0.10	0.69	0.45	0.49
Present Tense	0.60	−0.36	0.01	0.67	0.51	0.47
Future Tense	0.24	−0.33	−0.07	0.37	0.21	0.36
Adverbs	−0.04	−0.17	−0.33	0.17	−0.09	0.21
Prepositions	−0.46	0.08	−0.34	0.00	−0.43	−0.11
Conjunctions	−0.34	−0.11	−0.44	0.06	−0.34	0.11
Cognitive	−0.32	−0.09	−0.42	0.06	−0.31	0.11
Exclusive	−0.23	−0.09	−0.36	0.09	−0.24	0.12

*: Only cases with moderate (0.3 > x < 0.5) and strong (x > 0.5) Pearson correlations are presented. **: PCA 1, PCA 2 are the first two reduced dimensions of principal components analysis.

It should be noted that similar correlation patterns across recommendation communities (for both the Doc2Vec and S-BERT methods) are indicative that representations can reflect behaviors not affected by specific community circumstances, but are typical of how users communicate, interact, and contribute to similar communities. The findings of correlation analysis can also lead to the assumption that users exhibit consistent writing behaviors across different threads. Given that embeddings of both methods were computed at the user level, variability explained by the extracted factors essentially represents differences between users and not between individual comments. Overall, the reduced factors exhibited a strong ability to interpret different aspects of user writing tendencies and patterns, also reflected in their engagement behaviors. So, we proceeded with using latent factors to perform clustering analysis and identify major user segments in all communities.

4.2. Cluster Analysis

As described in detail in Section 3, the reduced PCA factors for each method were used to acquire user clusters through the K-Means method. The derived clusters give a clear and consistent distinction of users by tendencies of behavioral engagement and corresponding language qualities. Mean comparisons of the evaluation variables between clusters were produced, to reveal user segment differences in a more clear and concise manner. For a more comprehensible presentation, the results presented include only variables with mean differences of moderate and large effect sizes, measured by the eta-squared statistic.

Figures 1 and 2 present the percentages of users and comments of each cluster by Reddit community, revealing largely similar distributions, by level of contribution and by level of positive sentiment intensity.

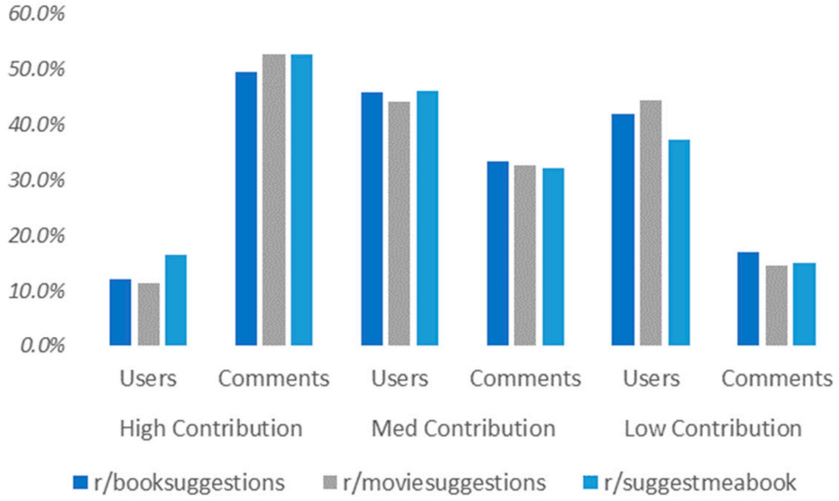


Figure 1. %Users and %Comments by Cluster and Community–Doc2Vec Approach.

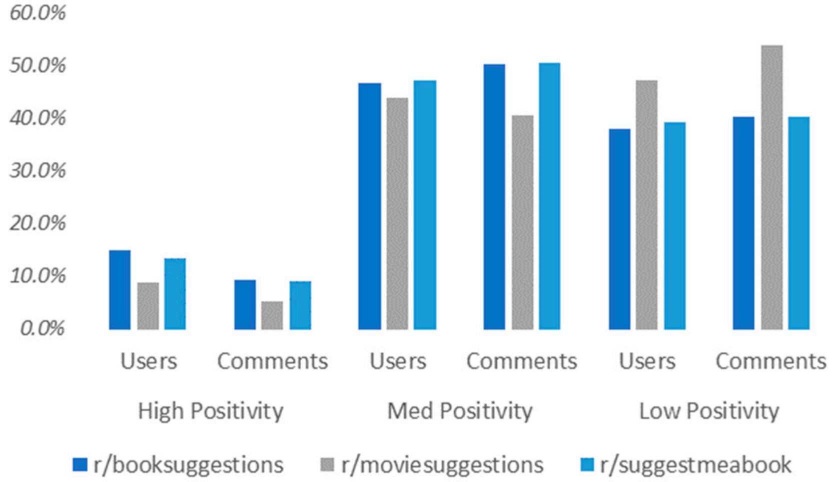


Figure 2. %Users and %Comments by Cluster and Community–S-BERT Approach.

We compare the means of the evaluation variables by cluster, based on the Doc2Vec and S-BERT methods. Both approaches can distinguish user segments with unique writing and behavioral properties. Based on the results, it could be assumed that each segment has highly distinct user representations, capturing the similar writing style tendencies of users.

4.2.1. User Clusters—Different Levels of Contribution

In Table 4, the segment of highly contributing users demonstrates the minority group who tend to provide a high volume of suggestions and are active in the community in the long term. These users are characterized by low positive sentiment, which is a contrasting result in terms of behavioral engagement. However, this should be seen in the context of users highly motivated to provide suggestions. According to language quality metrics, they write lengthier contributions, use a richer vocabulary, and require more effort by the reader in comprehension, according to readability indexes. This segment is a small percentage of users (12–16%) but is by far the most prolific, providing one in two comments on average (~50%, Figure 1). Therefore, due to their long-term engagement, they have a significant part in the overall impact of recommendations (aggregate vote score and replies). On the other hand, less prolific commentators tend to provide less detailed, shorter, and more socially and emotionally driven comments. This is evident in medium and low contribution users, who have proportionally lower values in the metrics of comment length and text complexity, and correspondingly higher values of emotional writing.

Table 4. Means * of Evaluation Variables by User Cluster **—Doc2Vec approach.

	r/Booksuggestions				**r/Moviesuggestions**				**r/Suggestmeabook**			
Behavioral Engagement	**HC**	**MC**	**LC**	**etasq**	**HC**	**MC**	**LC**	**etasq**	**HC**	**MC**	**LC**	**etasq**
Activ Length Months	6.41	3.66	2.35	0.27	6.70	4.09	2.86	0.21	5.90	3.52	2.41	0.24
Activ Length Days	29.76	7.31	3.80	0.31	40.05	9.78	4.84	0.33	23.28	6.74	3.77	0.28
Agg Comment Score	182.97	36.70	16.81	0.19	322.97	52.35	22.16	0.17	180.47	42.07	19.51	0.16
Agg Replies Created	29.55	4.86	2.21	0.16	30.79	4.98	1.90	0.19	23.88	4.71	2.27	0.13
Num of Comments	63.87	11.25	6.23	0.16	98.00	15.57	6.90	0.15	49.37	10.82	6.20	0.14
Num of Unique Threads	51.69	8.32	3.64	0.15	77.06	11.52	5.18	0.13	37.58	7.44	3.67	0.11
Pos Sentiment Score	0.15	0.18	0.25	0.11	0.15	0.17	0.19	0.01	0.16	0.18	0.23	0.07
Neu Sentiment Score	0.79	0.77	0.71	0.09	0.77	0.76	0.75	0.01	0.79	0.77	0.72	0.05
Affect	0.06	0.07	0.11	0.08	0.07	0.08	0.10	0.02	0.06	0.07	0.10	0.05
Pos Emotion	0.05	0.06	0.09	0.09	0.05	0.06	0.07	0.02	0.05	0.06	0.08	0.06
Linguistic Style												
Tot Word	48.36	32.77	16.26	0.22	32.37	21.22	10.44	0.22	46.72	30.39	15.37	0.26
Tot Un Word	37.24	27.88	15.68	0.27	26.22	19.12	10.53	0.26	36.28	26.27	14.96	0.29
Avg Syll PS	19.01	16.44	11.26	0.22	14.04	12.46	8.84	0.17	18.49	15.81	11.09	0.22
RT_Fast	0.16	0.11	0.05	0.22	0.11	0.07	0.03	0.22	0.16	0.10	0.05	0.26
RT_Slow	0.28	0.19	0.09	0.22	0.18	0.12	0.06	0.22	0.27	0.17	0.09	0.26
SMOG Index	2.62	2.12	1.36	0.13	1.88	1.58	1.01	0.10	2.54	2.05	1.30	0.14
Common Verbs	0.09	0.10	0.13	0.07	0.08	0.09	0.09	0.00	0.09	0.10	0.12	0.03
Present Tense	0.06	0.06	0.09	0.06	0.05	0.06	0.06	0.00	0.06	0.06	0.08	0.03

*: Only cases with moderate and strong effect sizes are presented, measured by eta-squared. According to [65], eta-squared values > 0.06 indicate a moderate effect and values above 0.14 indicate a strong effect size. All demonstrated mean differences are statistically significant at the 99.9% level (ANOVA p-value < 0.000). **: HC, MC, and LC represent the high, medium, and low contribution clusters, respectively.

4.2.2. User Clusters—Different Levels of Sentiment

In addition to forming distinct groups of different sentiment engagement, clusters based on S-BERT representations (Table 5) appear to account for a wider range of writing style differences, compared to the Doc2Vec approach. Users exhibiting high positivity in their interactions are less complex in their writings and essentially represent the emotionally driven interactions, predominantly by users socializing in continuance of their own submissions seeking recommendations. This is observed by the mean differences of number of comments in self-initialized threads. However, the relationship between metrics of analytical, mature writing and sentiment does not appear monotonic. The S-BERT method captures users with a similar language style demonstrating significantly higher textual maturity, length, richness, and complexity while being moderately emotional. Overall, the medium positivity segment comprises users with more elaborate and rich suggestions, representing a significant proportion of users (44–47%) in all communities. Nevertheless, linguistic style variables regarding the use of function words and personal pronouns, as well as the use of common verbs and present tense tend to group with less affective and socially focused texts. This relationship is consistent with similar findings from several studies.

Table 5. Means * of Evaluation Variables by User Cluster **—S-BERT approach.

	r/Booksuggestions				**r/Moviesuggestions**				**r/Suggestmeabook**			
Behavioral Engagement	**HP**	**MP**	**LP**	**etasq**	**HP**	**MP**	**LP**	**etasq**	**HP**	**MP**	**LP**	**etasq**
Activ Length Months	1.78	3.70	3.78	0.08	2.00	3.81	4.21	0.05	2.05	3.73	3.71	0.06
Num of Com Own Threads	7.62	1.11	0.96	0.11	9.71	2.21	0.78	0.10	7.80	1.23	0.95	0.10
Comp Sentiment Score	0.50	0.39	0.22	0.21	0.44	0.25	0.09	0.26	0.49	0.37	0.19	0.22
Pos Sentiment Score	0.40	0.18	0.16	0.45	0.39	0.18	0.13	0.39	0.39	0.17	0.15	0.43
Neu Sentiment Score	0.57	0.77	0.78	0.38	0.58	0.75	0.79	0.27	0.59	0.77	0.79	0.34
Neg Sentiment Score	0.02	0.05	0.06	0.08	0.03	0.07	0.08	0.05	0.02	0.05	0.06	0.07
Affect	0.17	0.07	0.07	0.31	0.17	0.08	0.08	0.18	0.16	0.07	0.07	0.29
Pos Emotion	0.16	0.06	0.05	0.36	0.17	0.06	0.05	0.27	0.15	0.05	0.05	0.35
Linguistic Style												
Tot Word	14.25	37.11	21.45	0.16	12.16	26.13	10.87	0.23	15.34	36.84	20.39	0.17
Tot Un Word	13.89	30.97	19.10	0.22	12.17	22.81	10.60	0.31	14.77	30.80	18.24	0.24
Avg Syll PS	8.64	17.57	13.22	0.27	8.24	14.05	8.74	0.29	9.19	17.36	12.86	0.25
Avg Kup AoA PW	4.10	3.90	3.61	0.08	3.99	3.78	3.23	0.12	4.11	3.87	3.53	0.10
RT Fast	0.05	0.12	0.07	0.16	0.04	0.09	0.04	0.23	0.05	0.12	0.07	0.17
RT Slow	0.08	0.21	0.12	0.16	0.07	0.15	0.06	0.23	0.09	0.21	0.12	0.17
SMOG Index	1.21	2.27	1.61	0.11	0.95	1.79	1.03	0.14	1.25	2.24	1.58	0.10
Pers Pronouns	0.09	0.06	0.05	0.09	0.08	0.05	0.03	0.15	0.08	0.06	0.05	0.10
1st Pers Singular	0.05	0.04	0.03	0.03	0.04	0.03	0.02	0.11	0.05	0.04	0.03	0.04
2nd Pers	0.03	0.01	0.01	0.09	0.03	0.01	0.01	0.08	0.03	0.01	0.01	0.09
Articles	0.04	0.07	0.07	0.10	0.04	0.07	0.07	0.04	0.04	0.07	0.07	0.08
Common Verbs	0.20	0.10	0.09	0.34	0.20	0.10	0.06	0.36	0.19	0.10	0.08	0.33
Present Tense	0.16	0.06	0.05	0.35	0.15	0.06	0.04	0.34	0.15	0.06	0.05	0.35
Future Tense	0.02	0.01	0.01	0.13	0.02	0.01	0.00	0.10	0.02	0.01	0.01	0.13
Adverbs	0.04	0.04	0.03	0.02	0.03	0.04	0.02	0.09	0.04	0.04	0.03	0.03
Prepositions	0.07	0.11	0.10	0.13	0.06	0.08	0.06	0.07	0.07	0.11	0.10	0.11
Conjunctions	0.03	0.05	0.04	0.10	0.03	0.04	0.03	0.13	0.03	0.05	0.04	0.09
Cognitive	0.09	0.12	0.11	0.08	0.09	0.12	0.08	0.12	0.09	0.12	0.10	0.07
Exclusive	0.02	0.03	0.02	0.05	0.02	0.03	0.01	0.10	0.02	0.03	0.02	0.05

*: Only cases with moderate and strong effect sizes are presented, measured by eta-squared. According to [65], eta-squared values > 0.06 indicate a moderate effect size and values above 0.14 indicate a strong effect size. All demonstrated mean differences are statistically significant at the 99.9% level (ANOVA *p*-value < 0.000). **: HP, MP, LP represent the high, medium, and low positivity clusters, respectively.

In addition to the examination of large effect mean differences, it is worth noting that user representations of both approaches do not exhibit a significant capability to account for differences regarding other important behavioral engagement metrics, such as average comment score and average replies created. Even though statistically significant differences were found, the effect sizes were small. There is some evidence identifying a tendency of users with higher positivity to exhibit lower average comment scores, but the differences were not consistent across all the clusters and communities.

4.2.3. Testing Cluster Formation without Dimensionality Reduction

As already mentioned in Section 3.3, to test if similar conclusions can be reached regarding cluster results, we performed K-Means without prior dimensionality reduction for one of the two language models, specifically for the Doc2Vec approach, using all originally trained 300 dimensions. The results indicated that clusters were largely similar with very narrow, immaterial differences. Naturally, this approach was computationally more demanding. Also, as it would be expected, regarding metrics used for assessing cluster quality, Silhouette scores were lower compared to the initial approach and a proportion of the 300 dimensions failed to meet the criteria of the Bonferroni post hoc test of pairwise tests of mean differences. However, the K-Means output behaved in a closely similar way, resulting in almost identical clusters. Distributions of users between clusters, as well as differences and effect sizes of difference indicators, were largely the same or very close. We did not display the extra results, as the clusters' interpretation would not gain from the extra information provided. Overall, we conclude that performing clustering without dimensionality reduction offers an additional validation of our results' interpretation, demonstrating both the ability of latent factors to capture significant variability in the users' behavior, writing styles, and quality, as well the overall robustness of our approach.

5. Discussion

This study examined latent factors representing users, acquired by using two alternative neural language methods, solely based on Reddit comments as the input text. We found that based on the correlation results, user embeddings of both methods can describe and represent a significant part of users' writing and behavioral variability. According to the correlation and cluster analysis results, embedding representations account for differences in writing styles between users of different levels of involvement and participation in recommendation communities. Also, between users of varying levels of emotional engagement. We were able to interpret several aspects of their differences by using a large set of independent metrics covering sentiment, linguistic styles, and language quality. It is also important to mention that we ran the process for each community separately. The results were therefore carefully considered and evaluated for three different recommendation communities and for two different domains. This constitutes a triangulation of results that strengthens confidence in the interpretations, since most of the described observed patterns are consistent for users across communities and domains.

5.1. Evaluating User Embeddings (RQ1)

The large number of assessment variables demonstrating strong correlations with user latent factors suggests that user embeddings based on popular approaches (Doc2Vec, S-BERT) can produce high quality representations, explaining various aspects of writing tendencies and behaviors. Taking into consideration that most studies focus on content and topic-centered tasks, we consider our findings to be supportive of additional research toward the direction of writing style and behavioral engagement classification and analysis tasks, using similar embedding methods. In reference to RQ1, we found that depending on the followed method, user embeddings managed to discriminate users of explicit engagement behaviors. The Doc2Vec approach managed to delineate the writing tendencies of users by level of participation in the community. Therefore, it provides an attainable method for the representation of unique writing characteristics of users highly motivated to provide recommendations and engage in suggestion-seeking threads in the long term. On the other hand, embeddings by the S-BERT approach were found to be more adapted to sentiment manifestations, and especially variations of positivity. Using embeddings for sentiment prediction purposes is a common topic in the relevant research. However, utilizing embeddings at the user level describing emotional differences, in conjunction with corresponding language style patterns, is suggested as a deeper approach, to specify user roles having varying levels of sentiment engagement in community interactions. In our examined simplified version, the S-BERT method appears to be able to achieve this convincingly.

Both methods were able to deliver representations accounting for writing style differentiations in multiple dimensions, with S-BERT embeddings being more sensitive to specific linguistic process words such as function words and common verb categories. Strong effects were verified through indicators of comment length, readability indexes, complexity, and use of grammatical forms. So, embeddings can provide a consolidated rendering of user writing style formation that can aid the understanding of user personality, roles, and behavior to a significant extent. Overall, users can clearly be distinguished by level of text rigor, detail, maturity, and comprehension difficulty, exhibiting an inverse relationship with emotionally and socially driven contributions. This can be clearly interpreted by examining correlation coefficients of unique words per comment, syllables per sentence, reading time, and SMOG readability indexes, in combination with correlations of sentiment scores and associated function words and verb categories.

5.2. Prominent User Segments (RQ2)

In response to RQ2, clusters emerging from user comment representations define segments of distinct writing and behavioral patterns. In consequence, several inferences could be formed for user roles and traits by drawing examples from the literature relevant to the interpretation of linguistic styles, in terms of personality type, thinking styles, and psychological inclinations. This can be useful for gaining insights on user segments with different impacts and communication styles in consumer recommendation communities. In Table 4, (Doc2Vec-based cluster results) the cluster of high contribution highlights the characteristics of users that make more recommendations and are engaged for longer time periods with the community. The group exhibits the predominant features of longer, more complex comments, using more rare and longer words, writing texts that demand more effort and maturity from the reader. So, there are properties that can potentially describe intuitive personalities [39]. Also, given the tendency of their comments to be formal and of mainly neutral sentiment, they might refer to the thinking styles which are organized, emotionally distant, and analytical in nature [8,11]. These thinking styles appear to be aligned with a group of dedicated users motivated to provide detailed, informative, and formal suggestions. On the other hand, low contribution users appear to be less elaborate and focus more on the emotional and social exchanges of discussions. Overall, based on the results, consistent participation is closely related to increased quality, formality, and maturity of recommendations. Additionally, the high contribution segment is the most influential through their aggregate recommendation impact in the long term, representing the main source of recommendations (~50% of total comments). Figure A1 in Appendix B demonstrates a few contribution examples by users assigned to the high and low contribution clusters.

In Table 5 (S-BERT-based cluster results), the patterns of the relationship between sentiment and language quality are more clearly delineated, with greater granularity. Clusters describe several differences in linguistic categories as well as in cognitive process words. The medium positivity cluster comprises many of the characteristics of formal, analytical, and categorical thinking mentioned in the previous sections. Compared to the other two clusters, these users have the inclination to write significantly longer, more complex, and sophisticated comments. These are suggested by differences in unique words per comment, syllables per sentence, SMOG readability index, and Kuperman age of acquisition index as demonstrated in Table 5. Moreover, this segment is characterized by more extensive use of articles and prepositions, which are also indicative of logical, formal, less emotionally focused, as well as open-minded authors [8,37,38]. On the other hand, the high positivity cluster comprises all the features pertaining to extroverted, resilient, and emotionally perceptive personalities [9,36,40]. These users employ informal language identified by short and simple comments, a significantly higher use of verbs and present tense, and the use of more personal pronouns. To a large degree, this segment refers to seekers of suggestions and their subsequent social engagements and interactions, according to the "comments in own threads" (comments in self-initiated threads). So, even though they represent a small fraction of users and comments, they are drivers of the social part of discussions that could be characterized by agreeableness [38] and high positive sentiment. Finally, the low positivity segment represents users of low emotional engagement, offering comments of average length and quality. However, they appear more likely to participate in the community for longer time periods. Figure A2 in Appendix B demonstrates a few contribution examples by users assigned to high and low positivity clusters.

5.3. Implications, Limitations, and Future Research

This paper enhances the understanding of user contribution characteristics and dynamics in online recommendation communities, which is an under-researched topic. Our study employed two simplified neural language approaches to learn user embeddings using solely users' comment texts. In the context of the study's aims, user representations and subsequent reduced latent factors demonstrated a significant ability to capture user behaviors and writing styles. Using the same factors, we applied cluster analysis to identify user segments distinguished by unique behavioral and linguistic features. Therefore, our findings are a constructive contribution to analysis efforts regarding user behaviors and roles in recommendation communities. Also, by demonstrating the ability of user embedding representations to describe multiple dimensions regarding how a user writes, contributes, and acts in suggestion communities, we enrich the stream of studies investigating alternative aspects of embeddings at the user level. Additionally, the set of evaluation dimensions used by three different sources can be a useful reference for future research.

Marketing can also benefit from the findings of our study, since it provides an approach to identify and target users with specific properties, monitor their activity over time, and use these segments for achieving marketing targets. For example, the segment of influential, active in the long-term users could potentially be targeted for facilitating promotion efforts, but also be monitored and analyzed for the assessment of trends and information regarding competition. Users with brief social engagement in the community acting mainly as suggestion seekers can also be a significant source of marketing intelligence. Moreover, distinct writing qualities of users can potentially provide useful inputs for the assessment of behavioral, psychological, and personality characteristics, and connection with specific preferences, products, and opinions.

While examining two established methods of acquiring user representations served the purposes of this study, using similar methodological approaches for classification, prediction modeling, and other machine learning tasks would require the incorporation of additional techniques, fine tuning, and possibly complementary data dimensions to serve specific cases and purposes.

Moreover, even though we repeated the analysis for three different recommendation communities, obtaining similar output patterns as a result, our effort remains highly context-specific and limited to a specific time frame (data covers the year 2021). Therefore, user behaviors and segments could differ significantly in recommendation communities with different structural and social characteristics and in different time periods. Communities covered in our sample represent two domains (book and movie suggestions); however, their underlying structure is quite similar and straightforward, given that seekers initializing a thread typically do not use media but plain text. This facilitates responses and helps subsequent discussions to remain focused. However, consumer suggestion communities using more media and having wider social qualities could exhibit different participant behaviors.

Another important limitation of our study concerns our evaluation framework. Although we examined the user representations' information quality in a wide range of dimensions, user embeddings incorporate a considerable contextual depth with a potential to reveal patterns on many levels. This should be evident in a comparison analysis between segments of "High Contribution" and "Medium Positivity" in Tables 4 and 5, respectively. Both segments exhibit similar linguistic characteristics. Although the observed differences represent large effects, it should be assumed that both clusters have additional dimensions to be interpreted and understood. Therefore, more aspects of writing qualities of users participating in recommendation communities should be examined by future studies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info15060356/s1. The provided Excel Workbook contains, in separate spreadsheets, datasets used for analysis at user level and for each community, specifically Dataset S1: user_data_booksuggestions, Dataset S2: user_data_moviesuggestions, and Dataset S3: user_data_suggestmeabook.

Author Contributions: Conceptualization, A.S.; Methodology, A.S.; Validation, C.L.; Formal analysis, A.S.; Resources, A.S.; Data curation, A.S.; Writing—original draft, A.S.; Writing—review & editing, C.L.; Visualization, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Full research data (raw data, Python code scripts) are available upon request from the corresponding author's email address.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. All metrics used for latent factor evaluation. Sources and descriptions.

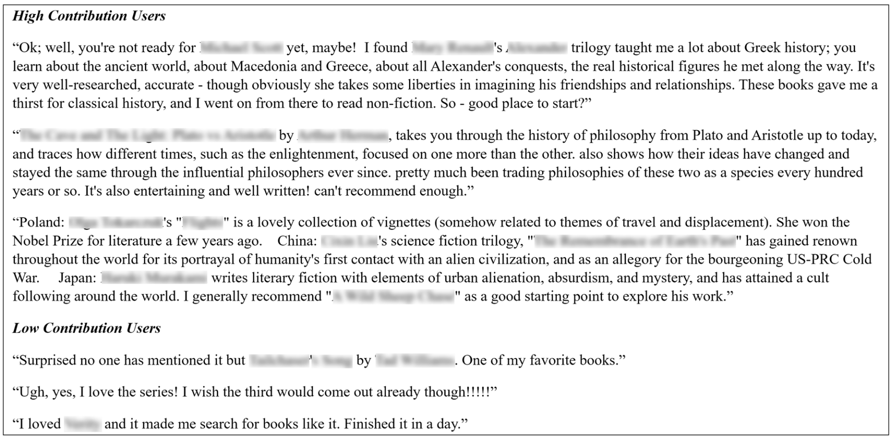
Metrics	Source	Description
Behavioral Engagement		
Activ Length Months	Reddit Dataset	Length of Activity—Total Months
Activ Length Days	Reddit Dataset	Length of Activity—Total Days
Avg Comment Score	Reddit Dataset	Avg Score (Upvotes–Downvotes) per Comment
Avg Replies Created	Reddit Dataset	Average Total Comments (Replies) Created by Each User Comment
Agg Comment Score	Reddit Dataset	Aggregate of Score (Upvotes Minus Downvotes) of Comments
Agg Replies Created	Reddit Dataset	Aggregate Replies Created by User Comments
Num of Comments	Reddit Dataset	Number of User Comments
Num of Unique Threads	Reddit Dataset	Number of Unique Threads Participated
Num of Com Own Threads	Reddit Dataset	Number of Comments in Own Submissions (Threads Created by the User)
Comp Sentiment Score	VADER Library	Compound Sentiment VADER Score per Comment
Pos Sentiment Score	VADER Library	Positive Sentiment VADER Score per Comment
Neu Sentiment Score	VADER Library	Neutral Sentiment VADER Score per Comment
Neg Sentiment Score	VADER Library	Negative Sentiment VADER Score per Comment
Affect [1]	LIWC Library (2007)	Number of Affective Words per Comment
Pos Emotion [1]	LIWC Library (2007)	Number of Positive Emotion Words per Comment
Neg Emotion [1]	LIWC Library (2007)	Number of Negative Emotion Words per Comment
Linguistic Style		
Tot Word [1]	LFTK Library	Words per Comment
Tot Un Word [1]	LFTK Library	Unique Words per Comment
Avg Syll PS [2]	LFTK Library	Syllables per Sentence

Table A1. *Cont.*

Metrics	Source	Description
Avg Char PW [3]	LFTK Library	Characters per Word
Avg Syll PW [3]	LFTK Library	Syllables per Word
Avg Kup AoA PW [3]	LFTK Library	Kuperman Age of Acquisition per Word
RT Fast	LFTK Library	Reading Time of Fast Reader
RT Slow	LFTK Library	Reading Time of Slow Reader
Coleman–Liau Index	LFTK Library	Coleman–Liau Readability Index
SMOG Index	LFTK Library	SMOG Readability Index
Pers Pronouns [1]	LIWC Library (2007)	Personal Pronoun Words per Comment
1st Pers Singular [1]	LIWC Library (2007)	1st Person Singular Words per Comment
1st Pers Plural [1]	LIWC Library (2007)	1st Person Plural Words per Comment
2nd Pers [1]	LIWC Library (2007)	2nd Person Words per Comment
3rd Pers Singular [1]	LIWC Library (2007)	3rd Person Singular Words per Comment
3rd Pers Plural [1]	LIWC Library (2007)	3rd Person Plural Words per Comment
Articles [1]	LIWC Library (2007)	Articles Words per Comment
Common Verbs [1]	LIWC Library (2007)	Common Verbs Words per Comment
Past Tense [1]	LIWC Library (2007)	Past Tense Words per Comment
Present Tense [1]	LIWC Library (2007)	Present Tense Words per Comment
Future Tense [1]	LIWC Library (2007)	Future Tense Words per Comment
Adverbs [1]	LIWC Library (2007)	Adverbs Words per Comment
Prepositions [1]	LIWC Library (2007)	Prepositions Words per Comment
Conjunctions [1]	LIWC Library (2007)	Conjunctions Words per Comment
Negations [1]	LIWC Library (2007)	Negations Words per Comment
Quantifiers [1]	LIWC Library (2007)	Quantifiers Words per Comment
Social Processes [1]	LIWC Library (2007)	Social Processes Words per Comment
Family [1]	LIWC Library (2007)	Family Words per Comment
Friends [1]	LIWC Library (2007)	Friends Words per Comment
Humans [1]	LIWC Library (2007)	Humans Words per Comment
Cognitive [1]	LIWC Library (2007)	Cognitive Processes Words per Comment
Insight [1]	LIWC Library (2007)	Insight Words per Comment
Causation [1]	LIWC Library (2007)	Causation Words per Comment
Discrepancy [1]	LIWC Library (2007)	Discrepancy Words per Comment
Tentative [1]	LIWC Library (2007)	Tentative Words per Comment
Certainty [1]	LIWC Library (2007)	Certainty Words per Comment
Inhibition [1]	LIWC Library (2007)	Inhibition Words per Comment
Inclusive [1]	LIWC Library (2007)	Inclusive Words per Comment
Exclusive [1]	LIWC Library (2007)	Exclusive Words per Comment

[1]: Measured as words per comment. [2]: Measured per sentence. [3]: Measured per word.

Appendix B



High Contribution Users

“Ok; well, you're not ready for [redacted] yet, maybe! I found [redacted]'s [redacted] trilogy taught me a lot about Greek history; you learn about the ancient world, about Macedonia and Greece, about all Alexander's conquests, the real historical figures he met along the way. It's very well-researched, accurate - though obviously she takes some liberties in imagining his friendships and relationships. These books gave me a thirst for classical history, and I went on from there to read non-fiction. So - good place to start?”

“[redacted] by [redacted], takes you through the history of philosophy from Plato and Aristotle up to today, and traces how different times, such as the enlightenment, focused on one more than the other. also shows how their ideas have changed and stayed the same through the influential philosophers ever since. pretty much been trading philosophies of these two as a species every hundred years or so. It's also entertaining and well written! can't recommend enough.”

“Poland: [redacted]'s "[redacted]" is a lovely collection of vignettes (somehow related to themes of travel and displacement). She won the Nobel Prize for literature a few years ago. China: [redacted]'s science fiction trilogy, "[redacted]" has gained renown throughout the world for its portrayal of humanity's first contact with an alien civilization, and as an allegory for the bourgeoning US-PRC Cold War. Japan: [redacted] writes literary fiction with elements of urban alienation, absurdism, and mystery, and has attained a cult following around the world. I generally recommend "[redacted]" as a good starting point to explore his work.”

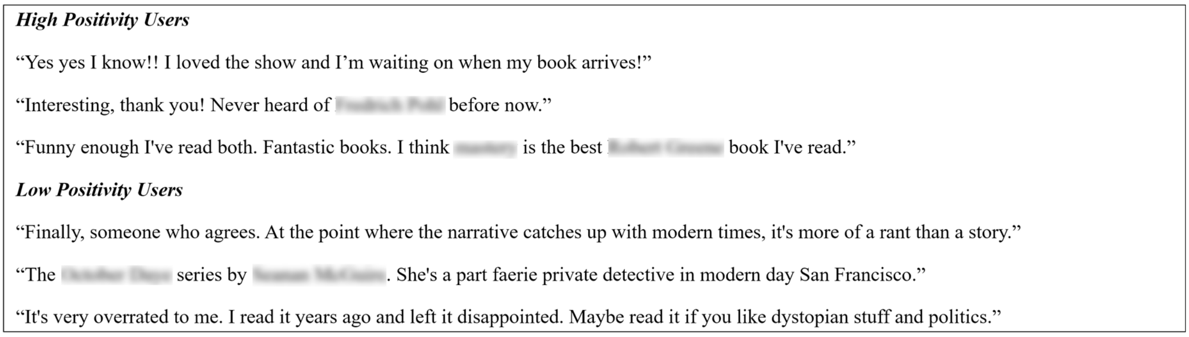
Low Contribution Users

“Surprised no one has mentioned it but [redacted] by [redacted]. One of my favorite books.”

“Ugh, yes, I love the series! I wish the third would come out already though!!!!!”

“I loved [redacted] and it made me search for books like it. Finished it in a day.”

Figure A1. Comment examples of users from high contribution and low contribution clusters.



High Positivity Users

"Yes yes I know!! I loved the show and I'm waiting on when my book arrives!"

"Interesting, thank you! Never heard of [redacted] before now."

"Funny enough I've read both. Fantastic books. I think [redacted] is the best [redacted] book I've read."

Low Positivity Users

"Finally, someone who agrees. At the point where the narrative catches up with modern times, it's more of a rant than a story."

"The [redacted] series by [redacted]. She's a part faerie private detective in modern day San Francisco."

"It's very overrated to me. I read it years ago and left it disappointed. Maybe read it if you like dystopian stuff and politics."

Figure A2. Comment examples of users from high positivity and low positivity clusters.

References

1. Russo Spena, T.; D'Auria, A.; Bifulco, F. Customer Insights and Consumer Profiling. In *Digital Transformation in the Cultural Heritage Sector*; Springer Nature: Cham, Switzerland, 2021; pp. 95–117. [CrossRef]
2. Smith, A. *Consumer Behaviour and Analytics*, 2nd ed.; Informa UK Limited: London, UK, 2023. [CrossRef]
3. Akar, E.; Mardikyan, S. User Roles and Contribution Patterns in Online Communities: A Managerial Perspective. *SAGE Open* **2018**, *8*, 2158244018794773. [CrossRef]
4. Bhattacharjee, D.R.; Pradhan, D.; Swani, K. Brand communities: A literature review and future research agendas using TCCM approach. *Int. J. Consum. Stud.* **2021**, *46*, 3–28. [CrossRef]
5. Veloutsou, C.; Black, I. Creating and managing participative brand communities: The roles members perform. *J. Bus. Res.* **2019**, *117*, 873–885. [CrossRef]
6. Lillqvist, E.; Moisander, J.K.; Firat, A.F. Consumers as legitimating agents: How consumer-citizens challenge marketer legitimacy on social media. *Int. J. Consum. Stud.* **2018**, *42*, 197–204. [CrossRef]
7. Reddit. How Community Recommendations Drive Collective Influence. 2023. Available online: https://connect.redditinc.com/hubfs/121662_Reddit%20Recommends%20Research%20Report_Superside_V4_V1.pdf (accessed on 3 March 2024).
8. Boyd, R.L. Psychological Text Analysis in the Digital Humanities. In *Data Analytics in Digital Humanities*; Springer International Publishing: Cham, Switzerland, 2017; pp. 161–189. [CrossRef]
9. Boyd, R.L.; Pennebaker, J.W. Language-based personality: A new approach to personality in a digital world. *Curr. Opin. Behav. Sci.* **2017**, *18*, 63–68. [CrossRef]
10. Lee, B.W.; Arockiaraj, B.F.; Jin, H. Linguistic Properties of Truthful Response. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 135–140. [CrossRef]
11. Boyd, R.L.; Pennebaker, J.W. Did Shakespeare Write *Double Falsehood*? Identifying Individuals by Creating Psychological Signatures With Text Analysis. *Psychol. Sci.* **2015**, *26*, 570–582. [CrossRef]
12. Gkikas, D.C.; Tzafilkou, K.; Theodoridis, P.K.; Garmpis, A.; Gkikas, M.C. How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in Facebook. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100067. [CrossRef]
13. Alzetta, C.; Dell'Orletta, F.; Miaschi, A.; Prat, E.; Venturi, G. Tell me how you write and I'll tell you what you read: A study on the writing style of book reviews. *J. Doc.* **2023**, *80*, 180–202. [CrossRef]
14. Dell'Orletta, F.; Montemagni, S.; Venturi, G. READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification. In Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, Edinburgh, UK, 30 July 2011; pp. 73–83. Available online: https://aclanthology.org/W11-2308 (accessed on 5 February 2024).
15. Forti, L.; Bolli, G.G.; Santarelli, F.; Santucci, V.; Spina, S. MALT-IT2: A new resource to measure text difficulty in light of CEFR levels for Italian L2 learning. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 7204–7211.
16. Biondi, G.; Franzoni, V.; Li, Y.; Milani, A.; Santucci, V. RITA: A Phraseological Dataset of CEFR Assignments and Exams for Italian as a Second Language. In Proceedings of the 2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Venice, Italy, 26–29 October 2023; pp. 425–430. [CrossRef]
17. Jian, L.; Xiang, H.; Le, G. English Text Readability Measurement Based on Convolutional Neural Network: A Hybrid Network Model. *Comput. Intell. Neurosci.* **2022**, *2022*, 6984586. [CrossRef]
18. Berggren, M.; Kaati, L.; Pelzer, B.; Stiff, H.; Lundmark, L.; Akrami, N. The generalizability of machine learning models of personality across two text domains. *Pers. Individ. Differ.* **2024**, *217*, 112465. [CrossRef]
19. Pan, S.; Ding, T. Social Media-based User Embedding: A literature review. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 6318–6324. [CrossRef]
20. Guimaraes, A.; Balalau, O.; Terolli, E.; Weikum, G. Analyzing the Traits and Anomalies of Political Discussions on Reddit. *Proc. Int. AAAI Conf. Web Soc. Media* **2019**, *13*, 205–213. [CrossRef]

21. Rivas, P.; Zimmermann, M. Empirical study of sentence embeddings for english sentences quality assessment. In Proceedings of the 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019, Las Vegas, NV, USA, 5–7 December 2019; pp. 331–336. [CrossRef]
22. Sepahpour-Fard, M.; Quayle, M.; Schuld, M.; Yasseri, T. Using word embeddings to analyse audience effects and individual differences in parenting Subreddits. *EPJ Data Sci.* **2023**, *12*, 38. [CrossRef]
23. Ahmad, H.; Asghar, M.Z.; Khan, A.S.; Habib, A. A Systematic Literature Review of Personality Trait Classification from Textual Content. *Open Comput. Sci.* **2020**, *10*, 175–193. [CrossRef]
24. Tegene, A.; Liu, Q.; Gan, Y.; Dai, T.; Leka, H.; Ayenew, M. Deep Learning and Embedding Based Latent Factor Model for Collaborative Recommender Systems. *Appl. Sci.* **2023**, *13*, 726. [CrossRef]
25. Schuld, M.; Durrheim, K.; Mafunda, M. Speaker landscapes: Machine learning opens a window on the everyday language of opinion. *Commun. Methods Meas.* **2023**, 1–17. [CrossRef]
26. Terreau, E.; Gourru, A.; Velcin, J. Writing Style Author Embedding Evaluation. In Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems, Stroudsburg, PA, USA, 10 November 2021; pp. 84–93. [CrossRef]
27. Curiskis, S.A.; Drake, B.; Osborn, T.R.; Kennedy, P.J. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Inf. Process. Manag.* **2019**, *57*, 102034. [CrossRef]
28. Bayrak, A.T. An application of Customer Embedding for Clustering. In Proceedings of the IEEE International Conference on Data Mining Workshops, ICDMW, Orlando, FL, USA, 28 November–1 December 2022; pp. 79–82. [CrossRef]
29. Cauteruccio, F.; Corradini, E.; Terracina, G.; Ursino, D.; Virgili, L. Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity. *J. Inf. Sci.* **2022**, *48*, 783–810. [CrossRef]
30. Arazzi, M.; Nicolazzo, S.; Nocera, A.; Zippo, M. The importance of the language for the evolution of online communities: An analysis based on Twitter and Reddit. *Expert Syst. Appl.* **2023**, *222*, 119847. [CrossRef]
31. Zhu, X.; de Melo, G. Sentence Analogies: Linguistic Regularities in Sentence Embeddings. In Proceedings of the 28th International Conference on Computational Linguistics, Stroudsburg, PA, USA, 8–13 December 2020; pp. 3389–3400, International Committee on Computational Linguistics. [CrossRef]
32. Simoulin, A. Sentence Embeddings and Their Relation with Sentence Structures. Ph.D. Thesis, Université Paris Cité, Paris, France, 2022.
33. Noguti, V. Post language and user engagement in online content communities. *Eur. J. Mark.* **2016**, *50*, 695–723. [CrossRef]
34. Santos, Z.R.; Cheung, C.M.K.; Coelho, P.S.; Rita, P. Consumer engagement in social media brand communities: A literature review. *Int. J. Inf. Manag.* **2021**, *63*, 102457. [CrossRef]
35. Zhang, Y.; Ridings, C.; Semenov, A. What to post? Understanding engagement cultivation in microblogging with big data-driven theory building. *Int. J. Inf. Manag.* **2022**, *71*, 102509. [CrossRef]
36. García-Rudolph, A.; Sanchez-Pinsach, D.; Frey, D.; Opisso, E.; Cisek, K.; Kelleher, J.D. Know an Emotion by the Company It Keeps: Word Embeddings from Reddit/Coronavirus. *Appl. Sci.* **2023**, *13*, 6713. [CrossRef]
37. Pennebaker, J.W.; Chung, C.K.; Ireland, M.; Gonzales, A.; Booth, R.J. *The Development and Psychometric Properties of LIWC2007*; University of Texas at Austin: Austin, TX, USA, 2007.
38. Yarkoni, T. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *J. Res. Pers.* **2010**, *44*, 363–373. [CrossRef]
39. Gjurković, M.; Šnajder, J. Reddit: A gold mine for personality prediction. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; pp. 87–97. [CrossRef]
40. Dover, Y.; Amichai-Hamburger, Y. Characteristics of online user-generated text predict the emotional intelligence of individuals. *Sci. Rep.* **2023**, *13*, 6778. [CrossRef]
41. Tavabi, L.; Tran, T.; Stefanov, K.; Borsari, B.; Woolley, J.D.; Scherer, S.; Soleymani, M. Analysis of Behavior Classification in Motivational Interviewing. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, Rio de Janeiro, Brazil, 8–13 July 2021; pp. 110–115. [CrossRef]
42. Biggiogera, J.; Boateng, G.; Hilpert, P.; Vowels, M.; Bodenmann, G.; Neysari, M.; Kowatsch, T. BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples' Conflict Interactions. In Proceedings of the ICMI '21 Companion: Companion Publication of the 2021 International Conference on Multimodal Interaction, New York, NY, USA, 18–22 October 2021; pp. 385–389. [CrossRef]
43. Nguyen, D.; Rosé, C.P. Language use as a reflection of socialization in online communities. In Proceedings of the Workshop on Languages in Social Media, Portland, Oregon, 23 June 2011; pp. 76–85.
44. Hay, J.; Doan, B.L.; Popineau, F.; Elhara, O.A. Representation learning of writing style. In Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), Online, 19 November 2020; pp. 232–243.
45. Huertas-Tato, J.; Martín, A.; Camacho, D. Understanding writing style in social media with a supervised contrastively pre-trained transformer. *Knowl. Based Syst.* **2024**, *296*, 111867. [CrossRef]
46. Strukova, S.; Ruipérez-Valiente, J.A.; Gómez Mármol, F. Computational approaches to detect experts in distributed online communities: A case study on Reddit. *Clust. Comput.* **2023**, *27*, 0123456789. [CrossRef]
47. Cork, A.; Everson, R.; Naserian, E.; Levine, M.; Koschate-Reis, M. Collective self-understanding: A linguistic style analysis of naturally occurring text data. *Behav. Res. Methods* **2022**, *55*, 4455–4477. [CrossRef]

48. Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; Blackburn, J. The pushshift reddit dataset. In Proceedings of the International AAAI Conference on Web and Social Media, Georgia, GA, USA, 8–11 June 2019; Volume 14, pp. 830–839.
49. Rani, S.; Ahmed, K.; Subramani, S. From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives. *Appl. Sci.* **2024**, *14*, 1547. [CrossRef]
50. Proferes, N.; Jones, N.; Gilbert, S.; Fiesler, C.; Zimmer, M. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Soc. Media + Soc.* **2021**, *7*, 20563051211019004. [CrossRef]
51. Bump, P. 24 Reddit Stats and Facts to Know in 2022. HubSpot. 2022. Available online: https://blog.hubspot.com/marketing/reddit-stats (accessed on 2 April 2024).
52. Hintz, E.A.; Betts, T. Reddit in communication research: Current status, future directions and best practices. *Ann. Int. Commun. Assoc.* **2022**, *46*, 116–133. [CrossRef]
53. Kilroy, D.; Healy, G.; Caton, S. Using Machine Learning to Improve Lead Times in the Identification of Emerging Customer Needs. *IEEE Access* **2022**, *10*, 37774–37795. [CrossRef]
54. Eberhard, L.; Popova, K.; Walk, S.; Helic, D. Computing recommendations from free-form text. *Expert Syst. Appl.* **2024**, *236*, 121268. [CrossRef]
55. Hutto, C.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proc. Int. AAAI Conf. Web Soc. Media* **2014**, *8*, 216–225. [CrossRef]
56. Lee, B.W.; Lee, J.H.J. LFTK: Handcrafted Features in Computational Linguistics. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Baltimore, MY, USA, 23–24 June 2014; pp. 1–19. [CrossRef]
57. Ruan, T.; Lv, Q. Public perception of electric vehicles on Reddit and Twitter: A cross-platform analysis. *Transp. Res. Interdiscip. Perspect.* **2023**, *21*, 100872. [CrossRef]
58. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 22–24 June 2014; pp. 1188–1196.
59. Aguilar, J.; Salazar, C.; Velasco, H.; Monsalve-Pulido, J.; Montoya, E. Comparison and Evaluation of Different Methods for the Feature Extraction from Educational Contents. *Computation* **2020**, *8*, 30. [CrossRef]
60. Budiarto, A.; Rahutomo, R.; Putra, H.N.; Cenggoro, T.W.; Kacamarga, M.F.; Pardamean, B. Unsupervised News Topic Modelling with Doc2Vec and Spherical Clustering. *Procedia Comput. Sci.* **2021**, *179*, 40–46. [CrossRef]
61. Karvelis, P.; Gavrilis, D.; Georgoulas, G.; Stylios, C. Topic recommendation using Doc2Vec. In Proceedings of the International Joint Conference on Neural Networks, Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–6. [CrossRef]
62. Wang, G.; Kwok, S.W.H. Using K-means clustering method with Doc2vec to understand the twitter users' opinions on COVID-19 vaccination. In Proceedings of the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 27–30 July 2021; pp. 1–4. [CrossRef]
63. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [CrossRef]
64. Iliescu, D.M.; Grand, R.; Qirko, S.; van der Goot, R. Much Gracias: Semi-supervised Code-switch Detection for Spanish-English: How far can we get? Computational Approaches to Linguistic Code-Switching. In Proceedings of the CALCS 2021—5th Workshop, Mexico City, Mexico, 11 June 2021; pp. 65–71. [CrossRef]
65. Adams, M.A.; Conway, T.L. Eta Squared. In *Encyclopedia of Quality of Life and Well-Being Research*; Michalos, A.C., Ed.; Springer: Dordrecht, The Netherlands, 2014. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.