Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 69-81 recibido 08-03-2024 revisado 29-04-2024 aceptado 16-05-2024

Comparison of Clustering Algorithms for Knowledge Discovery in Social Media Publications: A Case Study of Mental Health Analysis

Comparación de algoritmos de agrupamiento para el descubrimiento de conocimiento en publicaciones de redes sociales: un caso de estudio en salud mental

Manuel Couto,[1] **Javier Parapar**[2]
David E. Losada[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela
[2]Centro de Investigación en Tecnoloxías da Información e da Comunicación (CITIC),
Universidade da Coruña
[1]{manuel.couto.pintos, david.losada}@usc.es
[2]javier.parapar@udc.es

Abstract: In the age of social media, user-generated content is critical for detecting early signs of mental disorders. In this study, we use thematic clustering to analyze the content of the social media platform Reddit. Our primary goal is to use clustering techniques for comprehensive topic discovery, with a focus on identifying common themes among user groups suffering from mental illnesses such as depression, anorexia, gambling addiction, and self-harm. Our findings show that certain clusters are more cohesive, e.g., with a higher proportion of texts indicating depression. Furthermore, we discovered subreddits that are strongly linked to texts from the depressed user group. These findings shed light on how online interactions and subreddit themes may impact users' mental health, paving the way for future research and more targeted interventions in the field of online mental health.
Keywords: Mental Health, Social Networks, Clustering, Natural Language Processing.

Resumen: En la era de las redes sociales, el contenido generado por los usuarios es fundamental para detectar los primeros signos de trastornos mentales. En este estudio utilizamos el agrupamiento de publicaciones por tópicos para analizar el contenido de la plataforma Reddit. Nuestro objetivo primordial es utilizar técnicas de agrupamiento para descubrir temas centrales, con un enfoque en la identificación de temas comunes entre los grupos de usuarios que sufren enfermedades mentales como la depresión, la anorexia, la adicción a los juegos de azar y las autolesiones. Nuestros hallazgos muestran que ciertos clusters son más cohesivos, por ejemplo mostrando una mayor proporción de textos de personas con depresión. Además, hemos descubierto subreddits que estan fuertemente vinculados a textos escritos por usuarios deprimidos. Estos hallazgos arrojan luz sobre cómo las interacciones en línea y los temas que se tratan en los subreddits reflejan aspectos de salud mental, abriendo el camino para futuras investigaciones e intervenciones dirigidas a la prevención de trastornos.
Palabras clave: Salud Mental, Redes Sociales, Agrupamiento, Procesamiento de Lenguaje Natural.

ISSN 1135-5948 DOI 10.26342/2024-73-5

1 Introduction

In the era of social media, the rapid growth of online platforms has provided researchers with vast amounts of user-generated data. This wealth of information presents unique opportunities for studying and understanding human behaviour, particularly in the context of psychological profiling (Crestani, Losada, and Parapar, 2022). Clustering, as a fundamental data mining technique, plays a crucial role in the analysis and organisation of such data, enabling the identification of patterns and the discovery of valuable insights.

Clustering algorithms aim to group data points into distinct clusters based on their intrinsic characteristics. These unsupervised learning methods are powerful for uncovering hidden structures within large datasets. In psychological profiling of social media publications, clustering algorithms offer the potential to identify distinct user groups, including those exhibiting signs of specific mental disorders (Aragon et al., 2021; Shensa et al., 2018; Yazdavar et al., 2017).

Commonly employed clustering algorithms can be categorised into different classes. Partition-based algorithms, including K-means and Expectation-Maximization (EM), divide the data into non-overlapping clusters based on similarity metrics (Dempster, Laird, and Rubin, 1977; MacQueen, 1967). Hierarchical algorithms, such as Agglomerative and Divisive clustering, construct cluster hierarchies through merge and split operations (Day and Edelsbrunner, 1984). Density-based algorithms, including DBSCAN and OPTICS (Ester et al., 1996; Ankerst et al., 1999), group data points based on density-connected regions, effectively handling arbitrary-shaped clusters. Model-based algorithms, such as Gaussian Mixture Models (GMM) (Reynolds, 2009), assume certain probability distributions to describe the data and infer cluster assignments.

In this study, we compare and evaluate different clustering algorithms in the context of mental health and the analysis of user publications on social media. By leveraging a taxonomy of clustering algorithms, we can systematically assess their strengths, weaknesses, and suitability for the specific task of grouping users' posts, with a particular focus on texts related to various mental disorders.

The results of our study show that certain clusters share a strong thematic homogeneity (particularly topics focusing on romantic relationships or video games). On the other hand, we have seen that certain clusters are composed of a larger proportion of texts posted by depressed users.

In addition, we have been able to understand which communities (subreddits) are most correlated with the group of depressed users. This provides useful insights on which communities might have the largest proportion of depressed users or what kind of problems depressed people might be dealing with.

2 Related Work

The analysis of user-generated content on social media offers unique opportunities for psychological profiling, as it provides rich and extensive data about individuals' thoughts, emotions, and behaviour in an online environment (Crestani, Losada, and Parapar, 2022; Chancellor and De Choudhury, 2020; Parapar et al., 2022; Couto, Pérez, and Parapar, 2022). By leveraging clustering algorithms, researchers can identify distinct user groups and gain insights into the psychological characteristics of individuals (Clatworthy et al., 2005). However, conducting psychological profiling using social media data poses challenges due to the large volume of unstructured text and the need for accurate representation and analysis of user posts. The selection of the appropriate clustering algorithm depends on the nature of the data and the objectives of the study. In general-purpose clustering applications, partition-based algorithms are commonly used for their simplicity and efficiency, while density-based algorithms are effective in handling datasets with irregularly shaped clusters. Model-based algorithms assume specific probability distributions, making them suitable for certain applications such as topic modelling in psychological profiling (Fahad et al., 2014; Ezugwu et al., 2022).

Clustering algorithms have shown promise in identifying user groups exhibiting symptoms of various mental disorders, including depression, anxiety, eating disorders, addiction, and self-harm (Nguyen et al., 2022; Aragón, López-Monroy, and Montes-y Gómez, 2019; Aragon et al., 2021; Peres et al., 2021; Ghaharian et al., 2022; Crestani, Losada, and Parapar, 2022; Ríssola, Losada, and Crestani, 2021). By clustering users ba-

sed on their textual content, these studies have revealed distinct patterns and characteristics associated with different mental health concerns.

Evaluating the quality and effectiveness of clustering algorithms in psychological profiling studies requires appropriate evaluation metrics and validation techniques. Commonly used metrics include internal validation metrics such as Silhouette coefficient (Rousseeuw, 1987), Calinski-Harabasz Index (Caliński and Harabasz, 1974) or Davies-Boulding Index (Davies and Bouldin, 1979), which assess the compactness and separation of clusters. As external validation metrics, the Rand index (Rand, 1971), Normalize Mutual Information (Strehl and Ghosh, 2002) and Purity (Marutho et al., 2018) are commonly utilised to quantify the alignment between clustering outcomes and ground truth labels, thus capturing the agreement between them. Additionally, other validation techniques such as external indices (e.g., F-measure), internal indices (e.g., Dunn index), and visual inspection of cluster visualisations are employed to evaluate clustering solutions (Fahad et al., 2014; Emmons et al., 2016; Palacio-Niño and Berzal, 2019).

Studies applying clustering algorithms for psychological profiling on social media employ these metrics to assess the accuracy, reliability, and transferability of the obtained clustering results. For instance, Gao et al. (2023) provides a comprehensive review of methods and guidelines for employing clustering algorithms in the context of mental health. The authors emphasise the importance of internal and external validation metrics. Another example is the work by Ikeda et al. (2013), where hybrid methods are employed for user profiling on Twitter. In this study, clustering techniques are used to generate user groupings based on follower-followee relationships. Precision, recall, and F-measure were utilised as external validation metrics to evaluate the performance of the system. By employing appropriate evaluation measures, researchers can determine the effectiveness and applicability of different clustering algorithms in user profiling.

The present study aims to contribute to the advancement of psychological profiling techniques on social media platforms and improve the identification and support of individuals with mental health concerns. While other works focused on solving a classification problem, in this work we study the forms of representation and algorithms that behave best, with the aim of conducting an exploratory study. The primary objective of this study is to extract valuable insights from data. This knowledge enables us to gain a better understanding of the topics discussed by Reddit users experiencing depression. We explore the intricate correlations between Reddit communities and users displaying signs of depression. These findings are informative about mental health discussions in online communities and provide a valuable foundation for the development of effective tools for early detection of mental health issues.

3 Methodology

To obtain a diverse set of social media contents, we used two different **data collections**, the Webis-TLDR-17 collection (Völske et al., 2017) and the eRisk 2017 depression collection (Losada, Crestani, and Parapar, 2017).

Webis-TLDR-17. This collection consists of multiple posts from the social network Reddit, where each post has information about the subreddit (Reddit's subcommunity) where it was posted and a summary of the publication.

eRisk 2017 depression. This collection contains a thread of posts or comments from multiple Reddit users. Each user is annotated with a label indicating whether or not the user was diagnosed with depression (positive group or control group, respectively).

In this paper we thus work with texts that can be either posts or comments. We will use the generic term text or publication to refer to any individual post or individual comment. Any text is always linked to a user and a class. In the case of the Webis-TLDR-17 dataset, the class refers to the subreddit where the text was published. In the case of the eRisk 2017 depression dataset, the class refers to whether the text comes from a depressed user or a control user. We performed **sampling** on the Webis-TLDR-17 dataset as it contains millions of texts. We grouped all the content of the collection by subreddits and retained only those subreddits that had between 5,500 and 22,000 texts. This sampling allowed us to reduce the size of the dataset and focus on representative subreddits that have a substantial number of texts. By selec-

ting this specific range, we aimed to balance the availability of texts for each subreddit, avoiding subreddits with too few or too many entries. As a result, we ended up with 69 different subreddits.

The **preprocessing** stage plays a crucial role in cleaning and transforming the raw text, ensuring that it is suitable for subsequent analysis. We removed special characters with regular expressions, tokenised the texts into individual words or n-grams, eliminated stopwords, and normalised the text through techniques such as lemmatisation or stemming with the NLTK library, version 3.8.1 (Bird, Klein, and Loper, 2009).

Once the text data was cleaned, we employed various **vectorisation** techniques to represent the texts as numerical vectors suitable for clustering algorithms. We explored different alternatives, including:

Term Frequency (TF): The TF approach represents each document as a vector, with each dimension corresponding to a unique word in the corpus. Each numerical value represents the frequency of that word within the document. The final dimensionality of the TF vectors is equal to the vocabulary size (Croft, Metzler, and Strohman, 2010).

Term Frequency-Inverse Document Frequency (TF-IDF): The TF-IDF technique multiplies the term frequency by the inverse document frequency to represent the importance of words in a document. It assigns higher weights to words that are frequent within a document but rare across the entire corpus. The vector dimensionality is the same as that of the TF vectors (Croft, Metzler, and Strohman, 2010).

Bidirectional Encoder Representations from Transformers (BERT): It is a pre-trained language model that generates contextualised word embeddings. It captures the contextual meaning of words by considering their surrounding context within a sentence or document. BERT-based embeddings have been widely used in various natural language processing tasks, including clustering. The dimensionality of the $\text{BERT}_{\text{base}}$ embeddings used for this work is 768 dimensions (Devlin et al., 2018).

RoBERTa, which is another pre-trained language model that extends BERT's architecture. It incorporates additional pre-training techniques and achieves improved performance on various language understanding tasks. RoBERTa-based embeddings capture more nuanced contextual information and can enhance clustering results. The dimensionality of the $\text{RoBERTa}_{\text{base}}$ embeddings used for this work is 768 dimensions (Liu et al., 2019).

Generative Pre-trained Transformer 2 (GPT-2): GPT-2 is a powerful language model that generates coherent and contextually relevant text. It can be used to generate word embeddings that capture the semantic meaning of words and sentences. GPT-2-based embeddings have shown promising results in clustering tasks. The dimensionality of the GPT-2 embeddings used for this work is 768 dimensions (Radford et al., 2019).

TF and TF-IDF allow representing texts as vectors in a classic manner without any further computation step. For BERT and RoBERTa, we utilised mean pooling to create sentence embeddings (Devlin et al., 2018; Liu et al., 2019). Mean pooling calculates the average of all the word embeddings in a sentence, resulting in a fixed-length representation that captures the overall meaning of the sentence. This approach allows us to generate sentence embeddings that encapsulate the contextual information derived from the entire sequence. For GPT-2, we used the special token [CLS] as a summary of the sequence (Radford et al., 2019). The [CLS] token represents the entire sequence and carries information about the context and meaning of the text. By extracting the embedding of the [CLS] token, we obtain a condensed representation that captures the salient features and overall semantic meaning of the sequence.

By employing these vectorisation techniques (TF, TF-IDF, BERT, RoBERTa, and GPT-2), we aimed to capture different aspects of the textual data, including frequency-based importance, contextual understanding, and semantic meaning. These representations enable a more comprehensive analysis of the problem and facilitate effective clustering of the user-generated content.

In order to enhance the efficiency and effectiveness of our methodology, we incorporated the possibility of performing a **dimensionality reduction** step after vectorisation. This step aims to reduce the dimensionality of the text representations while preserving the most informative features.

We employed Singular Value Decomposition (SVD) as the chosen technique for di-

mensionality reduction. SVD decomposes the matrix of vectorised texts into three matrices representing the singular values, left singular vectors, and right singular vectors. By selecting a subset of the top-k singular values and corresponding singular vectors, we effectively reduced the dimensionality of the vectorised data.

The application of dimensionality reduction using SVD serves two primary purposes. Firstly, it significantly reduces the computational complexity of subsequent clustering algorithms, facilitating faster experimentation. Secondly, it can improve the performance of the clustering algorithms by eliminating noisy and irrelevant features, leading to more accurate and meaningful clusters (Ding and He, 2004; Kadhim, Cheah, and Ahamed, 2014).

3.1 Clustering Algorithms

In this section, we discuss the clustering algorithms employed to group the texts extracted from Reddit into meaningful clusters. Clustering is an unsupervised learning technique that aims to discover inherent patterns and structures in the data, allowing us to identify similarities and differences between the documents.

We explore several popular clustering algorithms known for their effectiveness in text analysis tasks. Each algorithm employs a unique approach to partition the data points into clusters based on their similarity. The choice of clustering algorithm depends on various factors, including the dataset characteristics, scalability, interpretability, and the desired clustering outcomes (Fahad et al., 2014; Ezugwu et al., 2022; Mahdi, Hosny, and Elhenawy, 2021).

The following clustering algorithms were selected for this study:

K-means is a widely used centroid-based clustering algorithm that aims to partition the data into a predefined number of clusters (K). It iteratively assigns data points to the nearest cluster centroid and updates the centroids until convergence. K-means is known for its simplicity, efficiency, and effectiveness in finding spherical-shaped clusters (Arthur and Vassilvitskii, 2006).

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that groups data points based on their density in the feature space. It identifies dense regions as clusters and considers low-density regions as noise or outliers. DBSCAN is effective in discovering clusters of arbitrary shapes and handling datasets with varying densities (Ester et al., 1996).

Gaussian Mixture Models (GMM) assume that the data points are generated from a mixture of Gaussian distributions. The GMM clustering algorithm fits a specified number of Gaussian components to the data, assigning data points to the most probable cluster. GMM is versatile, capturing clusters with different shapes and accommodating overlapping clusters (Reynolds, 2009).

By employing these diverse clustering algorithms, we aim to explore different approaches to organise the text data into clusters. Each algorithm brings unique characteristics and capabilities, enabling us to uncover distinct structures. While other algorithms were considered, such as Spectral clustering (Bolla, 2013) and Affinity propagation (Frey and Dueck, 2007), Agglomerative (Nielsen and Nielsen, 2016; Murtagh and Contreras, 2012), or Birch (Zhang, Ramakrishnan, and Livny, 1996), they were not included in this study due to their computational and memory limitations when working with large datasets (Mahdi, Hosny, and Elhenawy, 2021; Fahad et al., 2014).

3.2 Evaluation metrics

Internal validity metrics focus on evaluating the quality and coherence of the clusters based on the intrinsic characteristics of the data. The following internal validity metrics were utilised in our evaluation:

Silhouette Coefficient: The Silhouette coefficient (Rousseeuw, 1987) measures the average similarity of each data point to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating better-defined and more cohesive clusters.

Calinski-Harabasz Index (CH): The CH index (Caliński and Harabasz, 1974) computes the ratio of between-cluster dispersion to within-cluster dispersion. Higher values of this index indicate better-defined clusters with greater separation.

Davies-Bouldin Index (DB): DB index (Davies and Bouldin, 1979) measures the average similarity between clusters and provides a measure of the cluster separation. Lower

values indicate better-defined and more separated clusters.

External validity metrics assess the clustering results in relation to a ground truth with known class labels. These metrics measure the agreement between the clustering assignments and the true labels. The following external validity metrics were employed in our evaluation:

Adjusted Rand Index (ARI): The ARI (Hubert and Arabie, 1985) quantifies the similarity between the clustering assignments and the true labels. It considers all pairs of samples and evaluates the agreement between them, providing a value between -1 and 1. A higher value indicates better agreement.

Normalized Mutual Information (NMI): The NMI (Strehl and Ghosh, 2002) measures the amount of mutual information shared between the clustering assignments and the true labels, taking into account the class distribution. It ranges from 0 to 1, with higher values indicating better agreement.

3.3 Experimentation setup

The experimentation stage was divided into two phases. In the first phase, our objective was to determine the best combinations of vectorisers, clustering algorithms, and hyperparameters. Therefore, the space for experimentation was vast. We have optimized the number of clusters in all the models (testing the range between 2 and 10, in steps of 1, and the range between 10 and 100, in steps of 10). In addition, we tested two tolerance parameter configurations (values of 10^{-3} and 10^{-2}) in both GMM and k-means. We have also optimized certain model-specific parameters, such as the hyperparameters "algorithm" in kmeans, the "covariance type" in gmm, or the "eps" and "neightbours" of DBSCAN. However, performing a full exploration of hyperparameters would require a significant amount of time. To address this challenge, we adopted a non-exhaustive search approach to find optimal hyperparameter configurations. We utilized the RepeatedStratifiedKFold[1] algorithm and RandomizedSearchCV[2] technique, guided by the external validation metrics ARI and NMI. This allowed us to efficiently explore the hyperparameter space and identify promising configurations without exhaustively evaluating every possible combination. The search was performed on the Webis-TLDR-17 dataset, where the labels correspond to the subreddits from which the texts were extracted. Once we identified the best clustering configuration based on the validation metrics, we proceeded to a second phase, in which we leveraged the optimal clustering results obtained from the previous phase to conduct an exploratory analysis of the data.

[1] `https://scikit-learn.org/stable/modules/cross_validation.html`.

[2] `https://scikit-learn.org/stable/modules/grid_search.html`.

4 Clustering results

We begin by providing a summary of the best combination of clustering algorithm and vectorial representation. Table 1 reports the performance metrics obtained for the best configurations. It can be noted that the vectorial representation that has achieved the best results in terms of internal validation is TF. This could be attributed to a potential bias towards text length, as the TF approach emphasizes the importance of individual terms within the documents without any kind of normalisation. The TF feature values grow with no bound, while TF-IDF, BERT, RoBERTa, and GPT-2 have their vectorial representation values normalised (because of the TF-IDF weighting scheme or the layer normalisation technique in the transformer architecture).

On the other hand, when considering external validation metrics, it is observed that RoBERTa performs the best among the vectorial representations. RoBERTa, being a transformer-based language model, is capable of capturing more nuanced semantic information, which likely contributes to its superior performance in capturing the underlying patterns in the data.

Additionally, in terms of external validation, it is generally observed that GMM and K-means clustering outperform DBSCAN. This suggests that GMM and K-means are more effective in capturing the structure and patterns of the user-generated content compared to DBSCAN.

After conducting a non-exhaustive search for the best configurations, we have found that the combination yielding the highest ARI and NMI score is the GMM model with RoBERTa's representation. The optimal hyperparameters for this configuration were as

		External		Internal		
Algorithm	Vectorisation	ARI	NMI	Silhouette	CH	DB
GMM	TF N:3	0.096	0.092	0.430	236614.452	0.806
K-means	TF N:3	0.053	0.087	0.392	**258067.923**	0.8671
DBSCAN	TF N:3	0.008	0.002	**0.631**	233440.350	**0.579**
GMM	TF-IDF N:3	0.151	0.172	0.281	118091.257	1.129
K-means	TF-IDF N:3	0.126	0.164	0.311	132990.815	1.049
DBSCAN	TF-IDF N:3	-0.009	0.025	0.092	9685.614	2.341
GMM	BERT N:3	0.094	0.123	0.277	126807.031	1.072
K-means	BERT N:3	0.102	0.108	0.361	148423.361	0.965
DBSCAN	BERT N:3	0.001	0.001	-0.153	219.810	117.855
GMM	RoBERTa N:3	**0.180**	**0.200**	0.287	141981.974	1.211
K-means	RoBERTa N:3	0.169	0.183	0.293	149548.197	1.104
DBSCAN	RoBERTa N:3	0.078	0.147	0.200	99696.919	1.513
GMM	GPT-2 N:3	0.010	0.015	0.252	156534.401	1.129
K-means	GPT-2 N:3	0.010	0.015	0.256	226903.435	0.983
DBSCAN	GPT-2 N:3	-0.001	0.000	-0.153	187.103	336.351

Tabla 1: Best Hyperparameter Configurations and Performance Metrics (Webis-TLDR-17).

follows: Covariance type: "tied", Initialization parameters: "kmeans", Number of components: 5, Tolerance: 0.001.

Next, we proceed with an analysis phase of this clustering approach. We clustered the entire Webis-TLDR-17 corpus using this configuration and we counted the number of texts from each subreddit. By visualising this information for each cluster, we can gain insights into the distribution of subreddits across the identified clusters. By examining these trends, we can identify dominant or underrepresented subreddits within specific clusters, which could indicate shared themes or topics among the groups.

In order to accurately visualise the distribution of subreddits across different clusters, we have generated bar graphs or histograms. These histograms, shown in Figure 1, represent the most prominent subreddits in a every cluster. This visualisation provides interesting insights. For instance, it is evident that Cluster 1 and Cluster 4 are clusters with a predominant theme.

Cluster 1 is mainly composed of texts from subreddits such as relationships, relationship_advice, sex, and dating_advice, indicating a clear thematic homogeneity. Note also that the relationships subreddit has also a strong presence in clusters 2 and 3 but these other clusters show a more varied set of subreddits.

Cluster 4 shows a prominent subreddit leagueoflegends and other relevant subreddits related to various video games. This cluster has a clear thematic focus on video gaming. The leagueoflegends subreddit and other video game-related subreddits are also significantly present in other clusters, specifically clusters 0 and 3.

Clusters 0, 2, and 3 consist of subreddits covering a wider range of topics. For instance, we can see that cluster 0 includes texts from subreddits such as explainlikeimfive, atheism, Fitness, askscience, leagueoflegends, DnD, trees and so forth. These subreddits are not associated to a single theme but, instead, represent communities oriented to asking questions and engaging in discussions about various subjects. In the case of Cluster 2, we can see prominent subreddits such as Personalfinance, politics, relationships, explainlikeimfive, adviceanimals, worldnews or legaladvice. This group does not represent a focused set of discussions but it seems to include more formal subjects such as politics, finance, and even divorce-related issues. Lastly, Cluster 3 is predominantly composed of subreddits such as tifu, relationships, trees, leagueoflegends, fitness and adviceanimals. Once again, there is no specific common topic, and this group includes texts related to complaints or grievances about relationships, plant care, animal-related discussions, or following a training plan.

While certain clusters exhibit a strong thematic coherence, other clusters comprise texts from highly different sources. This suggests the existence of cross-discussions or shared interests among different clusters, highlighting the diverse nature of user-generated content within the Webis-TLDR-17 dataset.

Additionally, we analysed the pattern of assignment of texts from 69 subreddits to the five clusters. This analysis, presented in Figure 2, allows us to understand, for example, whether a given subreddit is concentrated on a single cluster or, instead, dispersed

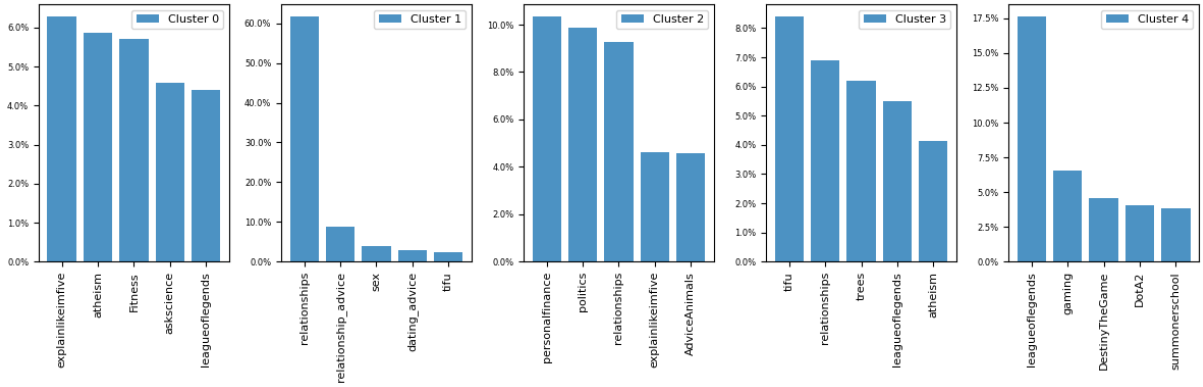


Figura 1: The subreddits with the highest proportion of texts in each cluster.

across several clusters. By understanding the distribution of each subreddit's texts within clusters, we gain insights into the subreddit's content cohesion and heterogeneity.

We can observe that certain subreddits such as Adviceanimals, IAmA, changemyview, or Worldnews are dispersed across multiple clusters. On the other hand, subreddits like DestinyTheGame, askscience, dating_advice, relationships and personalfinance are highly concentrated in specific clusters.

Some interesting observations can be made from the correlation matrix. For instance, there are high correlation values between the subreddit electronic_cigarette and various video game-related subreddits. This could be indicative of a potential overlap or shared interests among individuals who engage in both activities. There could also exist a subset of users who are active participants in discussions related to both electronic cigarettes and video games. Additionally, there is a notable correlation between the subreddit for depression and subreddits such as AskMen, AskWomen, TwoXChromosomes, self and sex. This suggests a connection between mental health discussions and topics related to relationships, gender, and self-expression. In fact, individuals seeking support or information about depression can be prone to engage in conversations about these related subjects.

4.1 eRisk collection

The clustering organisation described above was further exploited to analyse the publications from the eRisk 2017 depression dataset. To that end, we report here about the assignment of writings posted by different categories of eRisk users (depressed vs non-depressed) into the previously obtained clusters. By imputing each eRisk text to its closest cluster we can gain further insights into the posting patterns of users experiencing depression and try to understand how this reflects on different Reddit communities. This imputation process enables us to associate the depression-related texts with specific clusters, facilitating a comprehensive analysis of the data. The results of the imputation can be observed in Figure 3. Clusters 1 and 3 exhibit a higher proportion of texts from the depressed group compared to the control group. Note that these two clusters (see Figure 1) have a substantial portion of texts discussing personal experiences (e.g., relationships).

Next, we focus on the comparison between the depressed group and different subreddits. Figure 4 shows the correlation between the depressed group and several subreddits. These correlations were estimated by comparing the distribution across clusters of the depressed publications and the distribution across clusters of the subreddit's publications. Then we ordered in descending order the correlations. Notably, subreddits such as trees, NoFap, Drugs, ADHD, tifu, and talesfromtechsupport yield stronger correlations with the depressed group.

5 Discussion and Conclusions

After conducting this exploratory study, we have obtained relevant information from both datasets. Firstly, from the perspective of the formed clusters, we gained insights into the thematic content of each cluster. Secondly, from the perspective of the subreddits, we acquired statistical information about each subreddit, enabling us to compare them based on the distributions of their texts across dif-

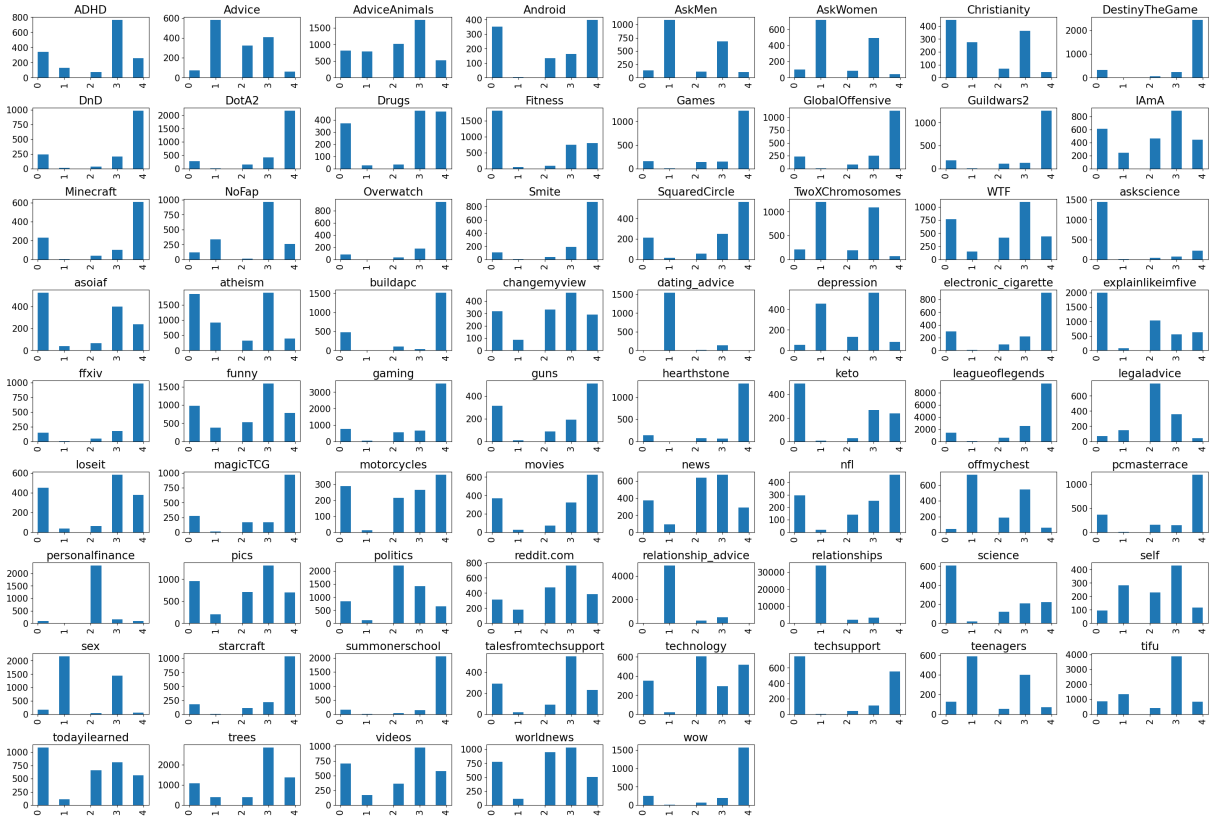


Figura 2: Histograms representing the distribution of the subreddit's texts in the clusters.

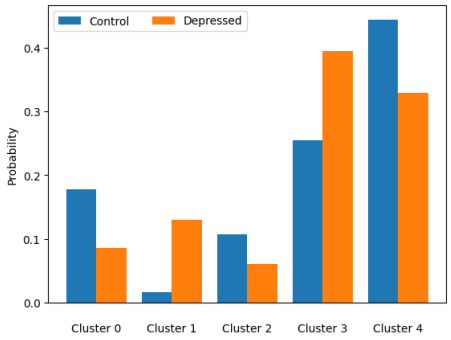


Figura 3: Distribution of probabilities of the texts from a certain user group (depressed/control) across the clusters.

ferent clusters.

As a result of this analysis, several significant findings have emerged. Clusters 1 and 4 are somehow focused clusters that share texts from various subreddits, focusing on the themes of relationships and video games, respectively.

Furthermore, after imputing the depression texts into the previously constructed clusters, we have obtained additional relevant information. From the perspective of the clusters, we observed that clusters 1 and 3 have a higher proportion of texts from depressed users compared to control users. This finding could potentially help to detect at-risk users. For instance, the activity of users within certain communities –not necessarily related to mental health– could be informative and, thus, act as supporting evidence or define new predictive features.

Texts from the depressed group exhibit a strong correlation with subreddits like trees, NoFap, Drugs, and ADHD. Some relevant words from these subreddits are shown in Figure 5 in the form of a wordcloud. Interestingly, the subreddit yielding the highest correlation is trees which, at first glance, may not seem directly related to depression. This subreddit is mainly focused on cannabis consumption. Smoking marijuana is closely associated with a variety of mental diseases, including depression. The Drugs subreddit is a similar case. Drug use is the primary topic of discussion in this forum; but substance abuse has been often linked to the development of mental disorders. The other subreddits are slightly different. NoFap is a pornography and sex addiction forum. On the other hand, ADHD, is a community where

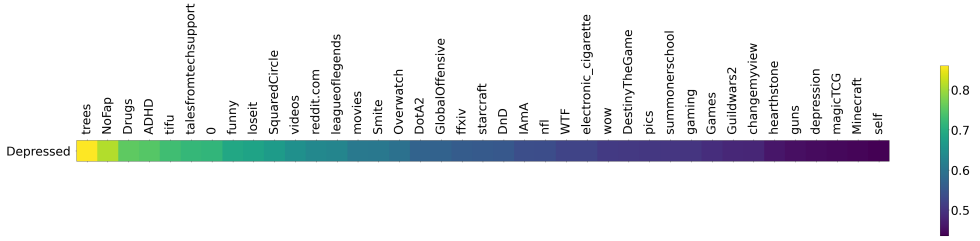
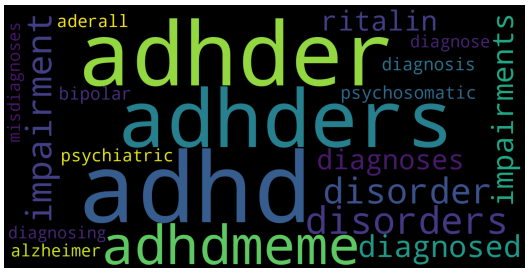
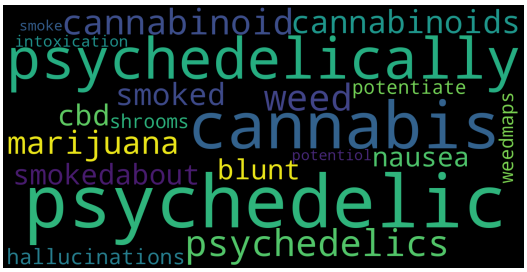


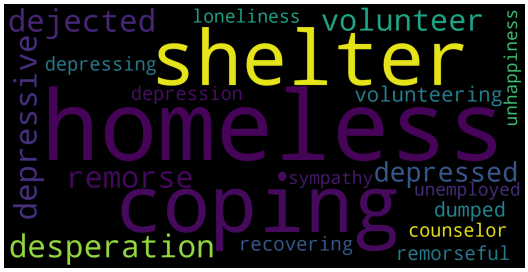
Figura 4: Most Correlated subreddits to the depressed set.



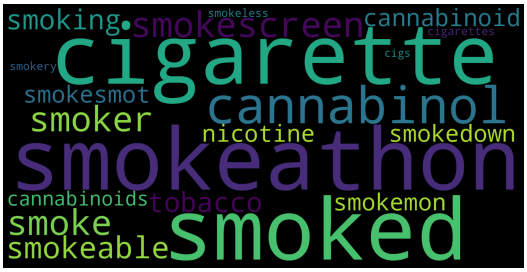
(a) Word cloud from the ADHD subreddit.



(b) Word cloud from the Drugs subreddit.



(c) Word cloud from the NoFap subreddit.



(d) Word cloud from the Trees subreddit.

Figura 5: Word cloud of the main topics from the most relevant subreddits.

people with ADHD can share their stories, struggles, and non-medication solutions. All of these forums are about problems, addictions, and people sharing negative experiences, life struggles, or something about themselves that makes them unhappy.

Acknowledgements

This work was supported by project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU). The first and third author thank the financial support supplied by the Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidade (ED431G 2023/04, ED431C 2022/19) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System. David E. Losada also thanks the financial support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego) and project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund).

The second author thanks the financial support supplied by the the CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, funded by the Xunta de Galicia and the EU through the ERDF Galicia 2021-27 operational program (ref. ED431G 2023/01) and project PID2022-137061OB-C21 (MCIN/AEI/ 10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by “ERDF A way of making Europe”, by the “European Union”).

References

Ankerst, M., M. M. Breunig, H.-P. Kriegel, and J. Sander. 1999. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

Aragon, M. E., A. P. Lopez-Monroy, L.-C. G. Gonzalez-Gurrola, and M. Montes. 2021. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *IEEE Transactions on Affective Computing*.

Aragón, M. E., A. P. López-Monroy, and M. Montes-y Gómez. 2019. Inaoe-cimat at erisk 2019: Detecting signs of anorexia using fine-grained emotions. In *CLEF (Working Notes)*.

Arthur, D. and S. Vassilvitskii. 2006. k-means++: The advantages of careful seeding. Technical report, Stanford.

Bird, S., E. Klein, and E. Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bolla, M. 2013. *Spectral clustering and biclustering: Learning large graphs and contingency tables*. John Wiley & Sons.

Caliński, T. and J. Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Chancellor, S. and M. De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Clatworthy, J., D. Buick, M. Hankins, J. Weinman, and R. Horne. 2005. The use and reporting of cluster analysis in health psychology: A review. *British journal of health psychology*, 10(3):329–358.

Couto, M., A. Pérez, and J. Parapar. 2022. Temporal word embeddings for early detection of signs of depression. In *Proceedings of the CIRCLE (Joint Conference of The Information Retrieval Communities in Europe)*.

Crestani, F., D. E. Losada, and J. Parapar. 2022. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*, volume 1018. Springer Nature.

Croft, W. B., D. Metzler, and T. Strohman. 2010. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading.

Davies, D. L. and D. W. Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Day, W. H. and H. Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, C. and X. He. 2004. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29.

Emmons, S., S. Kobourov, M. Gallant, and K. Börner. 2016. Analysis of network clustering algorithms and cluster quality metrics at scale. *PloS one*, 11(7):e0159161.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Ezugwu, A. E., A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu. 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743.

Fahad, A., N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279.

Frey, B. J. and D. Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Gao, C. X., D. Dwyer, Y. Zhu, C. L. Smith, L. Du, K. M. Filia, J. Bayer, J. M. Menssink, T. Wang, C. Bergmeir, et al. 2023. An overview of clustering methods with guidelines for application in mental health research. *Psychiatry Research*, page 115265.

Ghaharian, K., B. Abarbanel, D. Phung, P. Puranik, S. Kraus, A. Feldman, and B. Bernhard. 2022. Applications of data science for responsible gambling: a scoping review. *International Gambling Studies*, pages 1–24.

Hubert, L. and P. Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.

Ikeda, K., G. Hattori, C. Ono, H. Asoh, and T. Higashino. 2013. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, 51:35–47.

Kadhim, A. I., Y.-N. Cheah, and N. H. Ahamed. 2014. Text document preprocessing and dimension reduction techniques for text document clustering. In *2014 4th international conference on artificial intelligence with applications in engineering and technology*, pages 69–73. IEEE.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Losada, D. E., F. Crestani, and J. Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360. Springer.

MacQueen, J. 1967. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA.

Mahdi, M. A., K. M. Hosny, and I. Elhenawy. 2021. Scalable clustering algorithms for big data: A review. *IEEE Access*, 9:80015–80027.

Marutho, D., S. H. Handaka, E. Wijaya, et al. 2018. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 international seminar on application for technology of information and communication*, pages 533–538. IEEE.

Murtagh, F. and P. Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.

Nguyen, T., A. Yates, A. Zirikly, B. Desmet, and A. Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. *arXiv preprint arXiv:2204.10432*.

Nielsen, F. and F. Nielsen. 2016. Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, pages 195–211.

Palacio-Niño, J.-O. and F. Berzal. 2019. Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*.

Parapar, J., P. Martín-Rodilla, D. E. Losada, and F. Crestani. 2022. erisk 2022: pathological gambling, depression, and eating disorder challenges. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, pages 436–442. Springer.

Peres, F., E. Fallacara, L. Manzoni, M. Castelli, A. Popovič, M. Rodrigues, and P. Estevens. 2021. Time series clustering of online gambling activities for addicted users' detection. *Applied Sciences*, 11(5):2397.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Reynolds, D. A. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Ríssola, E. A., D. E. Losada, and F. Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2), mar.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Shensa, A., J. E. Sidani, M. A. Dew, C. G. Escobar-Viera, and B. A. Primack. 2018. Social media use and depression and anxiety symptoms: A cluster analysis. *American journal of health behavior*, 42(2):116–128.

Strehl, A. and J. Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

Völske, M., M. Potthast, S. Syed, and B. Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Yazdavar, A. H., H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198.

Zhang, T., R. Ramakrishnan, and M. Livny. 1996. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114.