Measuring the Effectiveness of LDA-Based Clustering for Social Media Data

AYSHA KHAN, RASHID ALI
Department of Computer Engineering ZHCET, Aligarh Muslim University Aligarh, INDIA

Abstract: Social media has come out as a great platform for users to communicate and share their opinions, photos, and videos that contemplate their moods, feelings, and emotions. This wide variety of data provides multiple possibilities for exploring social media data to investigate feelings and sentiments based on their moods and attitudes. With the enormous increase in mental health disorders among individuals, there is a massive loss in productivity and quality of life. Social media platforms like Reddit are used to seek medical advice on mental health issues. The structure and the content on various subreddits can be employed to interpret and connect the posts for mental health diagnostic problems. In this work, we have focused on seven mental health disorders, namely Anxiety, Depression, Bipolar, Autism, Borderline personality disorder, Schizophrenia, and mental health, which are actually subreddits posted by users on the Reddit social media platform. In this work, we have measured the effectiveness of topic modeling using Latent Dirichlet Allocation on these social media posts to identify the most used words and discover the hidden topics in their posts and also analyzed the results on evaluation metrics based on perplexity and coherence scores on unigrams, bigrams, and trigrams.

Keywords: Reddit, latent dirichlet allocation, topic modeling, mental health disorders, bigrams, trigrams, social media mining

Received: March 14, 2022. Revised: October 19, 2022. Accepted: November 24, 2022. Published: December 31, 2022.

1. Introduction

The proliferation of online social media has recreated the way people interact with each other. Online social networking sites like Facebook, Twitter, Reddit, and others provide an excellent platform of communication for users willing to convey their feelings, emotions, and sentiments regarding multiple topics or issues. From the user's perspective, online social media gives users to communicate and contribute to any case freely, and from the researchers' perspective, it provides deep insight about what could be the mental state of a user while they have interacted with any topic. To get more insights regarding these topics, machine learning techniques can be quite helpful for identifying the unique hidden features in this online communication and revealing the individual's corresponding mental state. In a survey by WHO, India, China, and the USA came out to be the terrible sufferers of anxiety, schizophrenia, and bipolar disorder [20].

Although effective treatment for mental diseases is available, only 10% of people take this treatment, and the majority is unable to approach this treatment because of the stigma associated with mental health. Other factors associated are inaccurate assessment, untrained healthcare professionals, and lack of education and resources. With the easy accessibility and enormous amount of data available, users tend to use the Internet to share and gather health-related information. In 2009, Pew Research Center published a report showing that 61% of adult internet users search for medical information online [3]. When any user or their close one is encountered with any medical condition, they start searching for it online by comparing the information available on multiple sources online [2]. Users' most searched topics are medication, side effects, symptoms, and healthcare professionals [9].

Looking from the patient's perspective, information that is available for users online serves two purposes. In the first scenario, users generally look for symptoms and treatments for mental health disorders and then diagnose themselves based on the information available. These solutions are convenient for users with little or no access to healthcare. On the other hand, users can have a deeper understanding of their medical condition while interacting with online medical health communities. But not all online communities cater all the information about the disorders, so sometimes people misdiagnose their health disorders. Lykke et al. showed that the queries used by experts and the naïve users for searching for medical information are remarkably different, which can make a user end up with the wrong diagnosis information [10]. When misinformation reaches a user, they can become more depressed and can end up in depression [7]

Reddit is an online social media platform used to seek and give healthcare advice on various topics and issues. Reddit allows users to freely and honestly share their concerns as their identity is hidden, and there is no stigma on social media regarding this as can be seen in fig:1 where a social media user on Reddit is expressing himself/herself by asking for an opinion/thought on his/her post in the mental health subreddit.

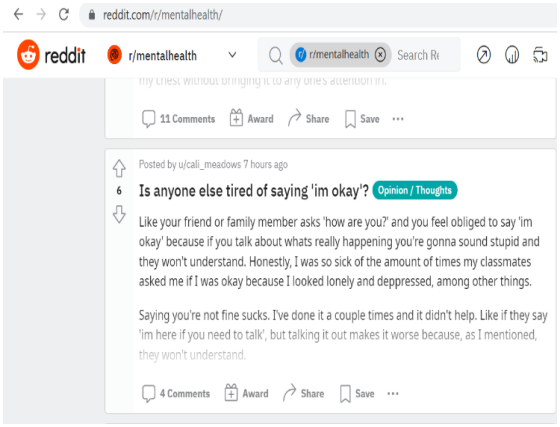

Fig: 1 Snapshot of a subreddit post by a user on the Reddit Platform.

When patients interact with practitioners about their mental health issues, the content shared having the questionnaire, and the interview is biased as they are organized in a clinical environment [12]. For more additional insights, the social media information of the user can also be linked with the above two traditional techniques. This social media content will help extract meaningful information about the attitude and awareness of the patient regarding their condition and what kind of treatment they need. However, accessing and using such social media information generally requires the users' consent to maintain their confidentiality. In this way, healthcare can be merged with social media, and health practitioners can aggregate this data to give personalized treatment to the users. This platform is persistent for conversations related to mental health. Its subreddits are discussion forums committed to specific discussions created by the users. Some of the famous subreddits for mental health are "r/mentalhealth" and "r/mentalillness" with more than 100k members each.

However, using such information requires de-identification and the consent of patients following the patient confidentiality requirements. In this case, a patient can provide their handle on Reddit and permission for the specified use, and their content can be collected and analyzed to predict the presence and progression of mental health conditions. Thus, a mental health professional (MHP) can aggregate various signals and personalized insights from diverse sources, including patient Electronic Health Records (EHR), questionnaires, interviews, and social media.

2. Motivations and Contributions of Paper

The inspiration behind this work is the immense popularity of online social media among every age group. This gives rise to a tremendous amount of data available on such platforms. This data can give us insights into various aspects of people's lives. To the best of our knowledge, we are the first to use topic modeling for mental health-related posts on a social media platform. Most of the papers we came across focus on the classification of social media posts. This paper proposes an LDA-based topic modeling technique using subreddits from the Reddit social media platform with seven classes. This work generally focuses on topic modeling with discovering the hidden patterns in the insightful data of the users. After performing various analysis, LDA is evaluated with multiple evaluation metrics.

The remaining paper is organized as follows: Section 2 describes the related work, followed by section 3 of the Preliminaries needed to understand our system. Section 4 describes the methodology, followed by section 5 with the conclusion.

3. Related Work

3.1 Using Social Media for Text Analysis

Twitter data was leveraged to predict the depressed users' intensity level from no to severe depression. They labeled the data weekly in a self-supervised way. The features were extracted on different levels ranging from emotional, topical, behavioral, user-level, and depression-related n-gram features for representing each user, and these features were then fed into a Long short-term memory (LSTM) network. With the experiments, it was discovered that depressed users tend to use words like stress and sad. They have a pattern of posting late at nights using personal pronouns. They compared their model with SVM, DNN, and GRU and claimed that their model performed the best [18].

Minjoo et al. analyzed subreddit communities for bipolar and depressive disorders to study how users share their illnesses and seek advice. A semantic network analysis was performed on this data and showed that users of both diseases had sleep disorders and financial problems. It was also discovered that users with bipolar disorder were more curious about medication, while users with depression had more interest in suicide. LIWC dictionary was used, and mean scores were calculated on four indicators: analytical thinking, clout, authenticity, and emotional tone. [15].

Han et al. worked on the Big-Five model to identify personality traits. The authors developed a Big-Five questionnaire system for 60-item personality inventories with the IDs of the users. Initially, personality-related keywords are obtained from the user's microblog, then word embeddings are applied and fed into a clustering algorithm. These words forming into these clusters are given different classes defining a semantic concept [6]. Park et al. [13] examined the language used on Twitter to see if it indicated depressive moods, attitudes, and behaviors. They also performed face-to-face interviews to identify the interdependence between these interviews and Twitter data.

Tsugawaet et al. [17] leverage tweets from Twitter to investigate the posting behavior of Japanese-speaking users to predict depression among them. Xue et al. [19][20] explored social media data for stress detection from tweets of teenagers by first detecting their psychological pressures and then assisting them in their stress through microblogs. Trotzeket et al. [16] gave a deep learning-based framework for proactively detecting depression from social media data. They used different word embeddings for their CNN model and compared it with user-level-linguistic metadata-based classification. Shen et al. [4] built a dataset on depression having depression and non-depression labels on Twitter by extracting six depression-related features covering clinical depression and online behavior on social media.

Zhancheng et al. developed a multi-label personality detection model using BERT on the MTBI and the Big-Five dataset by combining the semantic and emotional features on CNN, LSTM, and GRU. They evaluated their models on different datasets [21]. Majumdar et al. used word embeddings to analyse the semantics of the 2457 essays using convolutional neural networks [11]. In the past years, a new technique for word embedding has been developed i.e. transformers. Keh et al. used Bert embeddings for a personality detection model from social media data on Twitter and found improved accuracy in their work [8].

4. Preliminaries

4.1 Mental Health Disorders

A mental health disorder is characterized by a wide variety of mental illnesses that can take hold of an individual's mood, thinking, and behavior. In today's world, most people are involved with mental health, and it becomes a matter of concern when it persists continuously, causing untimely

stress and hampering the ability to function. Mental health can affect any person of any age. The symptoms of mental illness can be treated by combining medication and therapy.
According to a survey conducted by National Alliance on Mental Illness, every 1 in 5 adult experience mental illness every year in the U.S. In another study by the National Alliance on Mental Illness [23], 50% of mental illness begin by the age of 14, and 75% starts by the age of 24. In this research, it was found that mental health disorder is not the result of any single incident, it is the outcome of multiple factors like genetics, history of psychological or physical abuse, brain injury, consuming alcohol and recreational drugs, environment, lifestyle, biochemical processes, chemical imbalances in the brain are the factors associated with it [24].
The different mental health disorders are Anxiety disorders, Bipolar disorder, Borderline Personality disorder, Depression, Obsessive-Compulsive disorder, Post Traumatic Stress Disorder, Psychosis, Schizoaffective disorder, Schizophrenia etc.

4.2 N-Grams

N-Grams are defined as a contiguous sequence of n items that are initiated from a sample of text. This n can be any number 1,2,3 and so on.

- Unigrams: When the value of n is 1, then it is a unigram. It is the most commonly used approach and generally presents the words as single tokens. This method doesn't take into consideration any meaning of the words.
- Bigrams: When the value of n is 2, then it is a bigram. In this approach, two tokens are presented from the dataset.
- Trigrams: When the value of n is 3, then it is a trigram. In this method, three tokens are presented simultaneously for the document.

4.3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a language model for topic modeling based on clustering. It is an unsupervised technique for analyzing text. It is based on probabilistic distribution, which tells us the topics hidden in the documents. LDA gives two sets of probability distributions as a result. Firstly, for each document, it gives a topic set of distribution. Secondly, for each topic, it gives a word set of distributions [1] as can be seen in fig: 2.

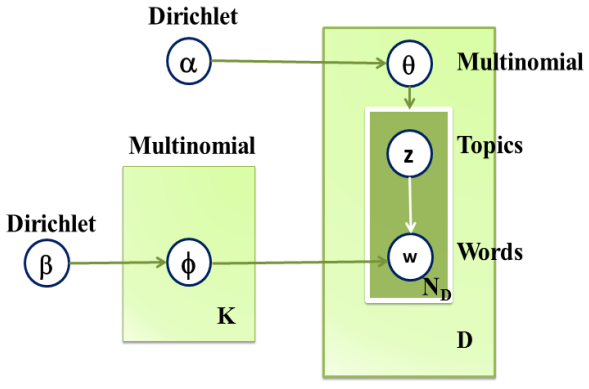

Fig: 2 Architecture of Latent Dirichlet Allocation

Here, the parameters of LDA are described as follows in fig:2:

D: the number of documents.
N_d: no. of words in a given document.
β: dirichlet prior on the per-document topic distribution.
α: dirichlet prior on the per-topic word distribution.
θ_i: topic distribution for document i.
φ_k: word distribution for topic k.
z_{ij}: topic for the j_{th} word in document i.
w_{ij}: specific word.

There are three essential hyperparameters for LDA:

- Document-topic density (α): It tells us the exact number of topics expected from the documents. If the value is smaller, then fewer topics will be given. If the value is greater, then more topics will be given.
- Topic-word density (β): It tells us about the word distribution in the topics. If it is smaller, then there are fewer words in the topic. If the value is greater, then more words are in the topic.
- Number of topics (K): This value tells us about the number of topics you want from the documents. This value depends on the domain in which you are working. Testing the different values of k can give better insights into the compositions.

4.4 Evaluation Metrics

For evaluating our topic modeling model LDA, we use evaluation metrics like the Perplexity and the Coherence score. It helps us understand how explainable the topics are to humans. Topics are basically represented as the top N words with the highest probability of being associated with the particular topic. In other words, it is the measure of the similarity of these words with each other.

1) Perplexity:

It is the metric for evaluating how successfully the trained model predicts the new data. It is defined as the decreasing function of the likelihood of new documents. Perplexity decreases with the increase in the likelihood of the words that are appearing together. A good topic is identified with a low perplexity score [24]. A low perplexity score says that it has predicted the words correctly in the documents. However, perplexity and human-judgment can differ in the same results and are often not correlated.

2) Coherence score:

It is defined as the measure of calculating the degree of semantic similarity between the high-scoring words in the topic. In this work, we have used the cv coherence score which generates content word vectors using their cooccurrences. For calculating the score, it uses cosine similarity and normalized pointwise mutual information (NPMI). It is one of the most popular and default metric techniques in the genism pipeline of python. It is denoted as c_v in genism [25]

4.5 Problem Definition

Mental health disorders impact people's lives in a way that is not easily identifiable. Like physical disorders, mental health disorders are not visible from the outside, and thus it becomes difficult to identify them quickly. But with the rise of social media, it has become easier to look at such disorders. In this

work, we have tried to work on social media posts of users on mental health disorders among people on the popular social media platform Reddit. We obtained the subreddits for Anxiety, Depression, Mental Health, Bipolar Disorder, Borderline Personality Disorder, Autism, and Schizophrenia [5]. The problem here is to identify the most discussed topics among the posts of these social media users on these particular subreddits. The main intention of using Latent Dirichlet Allocation (LDA) is that it is the most widely used topic model which helps in identifying the hidden topics in the text documents.

5. Methodology

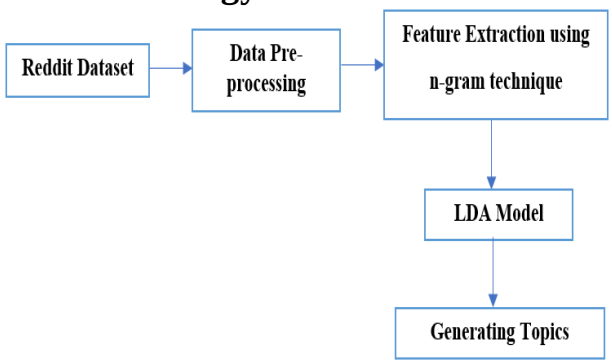

Fig: 3 System architecture of our work

For our work, we gathered data from the Reddit platform consisting of subreddits posted by various users. Initially, the dataset consisted of unstructured text and unwanted noise so this data was pre-processed. After that, features were extracted from the corpus. We have tried to build LDA on unigrams, bigrams, and trigrams. Then, these features were fed into the LDA-based model and are evaluated on different evaluation metrics, and finally the topics are generated as can be observed in fig: 3.

6. Experiments and Results

6.1 Dataset

In this work, we have used the Reddit dataset consisting of subreddit posts of users with seven mental health conditions. These mental health disorders are Anxiety, Schizophrenia, Bipolar disorder, Depression, Borderline Personality disorder (BPD), Autism, and Mental health [5] with a total of 227122 posts as shown in table 1. As can be seen in table 1, the total number of posts corresponding to each subreddit is presented.

Table 1: Subreddit with the total number of posts

Subreddit	Number of posts
Anxiety	25881
Autism	3572
Depression	129308
Bipolar	20759
Borderline Personality Disorder	19134
Schizophrenia	8769
Mental Health	19699
Total posts	227122

Table 2: Sample of a subreddit from Anxiety Disorder

Text	Subreddit
I have struggled with social anxiety from childhood and the main advice from friends, self-help books and professionals is to expose myself to those environments but I'm not better off!	Anxiety

Reddit allows users to post their content in different subreddits available. From Table 2, it can be observed that the user has posted about his/her anxiety in the Anxiety Subreddit. For our LDA, model we have not used the class of the particular subreddit, we only worked on the text available in the documents.

Initially, the dataset obtained was containing some erroneous characters. We saved the dataset in CSV files and used utf-8 encoding to remove that. Fig 4 shows us the word cloud of the words most prominent in our mental health dataset. We can see terms like people, know, feel, friend, want, life, depression, anxiety, and problem.

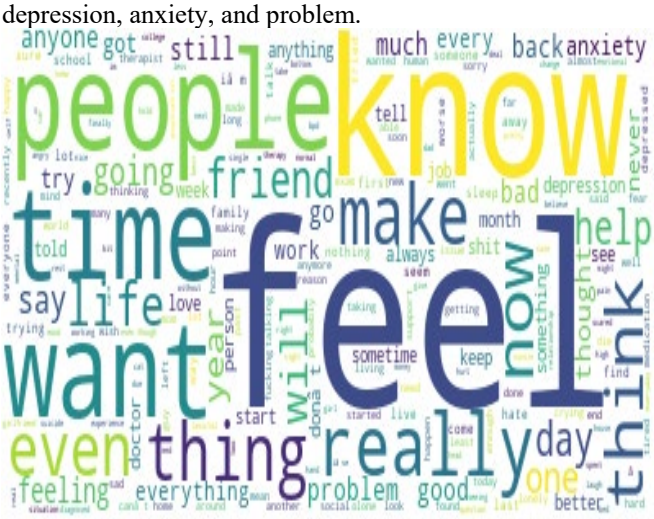

Fig: 4 Word cloud of Reddit dataset

6.2 Data Pre-processing

In our work, pre-processing has been performed in multiple steps. First, the text is cleaned for unnecessary special characters. Then punctuations are removed, and then finally the dataset is converted into the lower text. In the next step, the stopwords are removed as shown in table 3. In the next step, the data is tokenized, and then unigrams, bigrams, and trigrams are created. In the final step, the data is lemmatized separately for each unigram, bigram and trigram. For this pre-processing, the python libraries used are spacy, genism, and NLTK.

Table 3: Example of data pre-processing on the subreddit

After-preprocessing
i have struggled with social anxiety main advice from friends self help books professionals expose myself those environments i am not better off

6.3 Feature Extraction

High dimensional feature detection and extraction can perform better as it is one of the most significant steps of text mining [13]. The features that are extracted here are the unigrams, bigrams, and trigrams. In table 4, there is a glimpse

of how these are generated for our sample text shown in table 4.

Table: 4 samples of unigram, bigram, and trigram for our text

Text	i have struggled with social anxiety main advice from friends self help books professionals expose myself those environments i am not better off
Unigrams	'i', 'have', 'struggled', 'with', 'social', 'anxiety', 'main', 'advice', 'from', 'friends', 'self', 'help', 'books', 'professionals', 'expose', 'myself', 'those', 'environments', 'i', 'am', 'better', 'off'
Bigrams	'i have', 'struggled with', 'social anxiety', 'main advice', 'from friends', 'self help', 'professionals expose', 'myself those', 'environments i', 'am better', 'off'
Trigrams	'i have struggled', 'with social anxiety', 'main advice from', 'friends self help', 'professionals expose myself', 'those environments i', 'am better off'

6.4 LDA Results

Once the n-grams are generated, then a corpus is created for each of the n-grams structure using tf-idf which is then fed into the LDA model. The LDA model was run for k=7 number of topics, 10 number of passes with a chunk size of 100, a random state of 200, and default values for alpha and beta hyperparameters. The model gave top words for each topic in unigrams, bigrams, and trigrams which can be seen in table 5, table 6, table 7 respectively. In table 6, in topic 1, top words like get, life, depression, think, back and stress is prominent, which can be concluded to align with the depression class.

Table: 5 Top five words from seven topics for unigram

Topics	Top Words
Topic 0	Feel, never, more, month, live
Topic 1	Get, think, start, life, depression
Topic 2	Want, now, see, work, year
Topic 3	Time, bad, day, sleep, anxiety
Topic 4	Tell, say, give, find, sad
Topic 5	Go, make, try, people, leave
Topic 6	Help, thing, close, die, reason

Table 6 : Top five words from seven topics for bigram

Topics	Top Words
Topic 0	Attack, panic, exposure, stone, diagnosis
Topic 1	Get, life, depression, think, stress
Topic 2	Time, bad, friend, anxiety, sleep,
Topic 3	Autistic, therapy, learn, problems, birth
Topic 4	Therapista, environment, mood, life, bipolar
Topic 5	Struggle, throwaway, bpd, cry, work
Topic 6	Mental, vomit, negative, time, lose, need

Table: 7 Top five words from seven topics for trigrams

Topics	Top Words
Topic 0	Attack, panic, conversation, anxiety, time
Topic 1	Make, wake, medicine, social, bipolar
Topic 2	Autism, one, birth, issues, therapy
Topic 3	Feel, like, life, time, depression
Topic 4	Mood, life, environment, bipolar, sleep
Topic 5	Struggle, throwaway, hard,lose, mental
Topic 6	Paranoid, step, little, last, something, attack

6.5 Evaluation Metrics

Table: 8 Evaluation metrics for n-grams

No. of topics	Perplexity	Coherence Score
Unigrams	-8.17	0.25
Bigrams	-7.45	0.78
Trigrams	-7.06	0.72

For evaluating the topic models, perplexity and coherence scores are used. Table 4 shows us the evaluation metrics of the LDA model for number of topics (k = 7). Here, we have generated perplexity and coherence scores for unigrams, bigrams, and trigrams. As we mentioned earlier, negative perplexity is better, so for our model, it is best for unigrams i.e. -8.176. It can also be observed from table 8, that the perplexity is not much varied for unigrams, bigrams, and trigrams which clearly implies that perplexity is not a good measure for the n-gram features as they are almost close to each other. Whereas, the higher the value of the coherence score, the better it is. Here, it is best for bigrams i.e. 0.78. As unigrams are just a tokenized version of our text, so all the meanings are generally lost and because of that we got a very low coherence score of 0.25 for unigrams. Coherence score for bigrams and trigrams is very close to each other which implies that coherence score is a good measure.

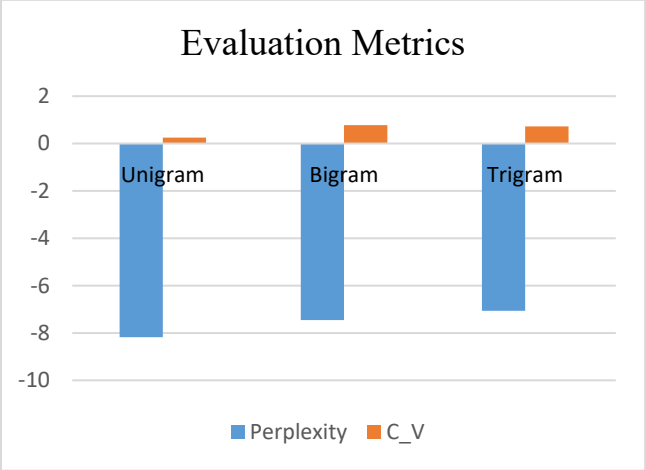

Fig: 4 Evaluation metrics for unigram, bigrams, trigrams

Here, in fig 4, perplexity and coherence scores for unigrams, bigrams, and trigrams are presented in graphical format. As, perplexity is generally negative it can be seen in a downwards

direction. The coherence score for unigram, bigram, and trigram is positive so it is upwards.

7. Conclusion

Mental health is impacting individuals worldwide. Finding a solution to combat it before it lets a person commit suicide is of utmost importance as it is still considered taboo to talk about it and consult a therapist for such problems. With the advent of social media, it is easy for people to share their issues with others, be it friends, healthcare professionals, experts etc. In this work, we proposed a novel topic modeling technique to find the inherent groups in individuals' mental health disorder datasets. The motive of the study is to identify the word usage of users on online social media platforms to understand how individuals are perceiving and sharing their experiences about mental health disorders. We applied the topic modeling technique i.e. Latent Dirichlet Allocation to understand these disorders. Therefore, using NLP techniques combined with social media texts can help researchers get better insights into the problems of individuals who cannot share them with their healthcare professionals.

8. Limitations and Future Work

This work has many limitations which we can remove in the future. Some of these are:

- When evaluating in terms of coherence score, we have only considered the CV coherence score. Although, there are more versions of the coherence score available.
- In this work, we have only presented the top words present in each topic only. We have not analyzed which of the subreddits are falling into which class.
- We have compared this work with different n-gram models using LDA.

In the future, we can also focus on improving this work. Some of them are:

- Talking about coherence score, we have only worked on one coherence score i.e. cv. There are other variants of coherence scores also available on which we are working.
- We have only focused on the Reddit platform. This work can also be performed on other social media platforms like Twitter, where users are vocal about their issues and problems, and try to implement it with deep learning models, and classify the different subreddits in the given seven classes.
- We can also compare this work with other topic modeling models Latent Semantic Indexing (LSI).

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation.", J. Mach. Learn. Res., vol. 3, pp. 993-1022, 2003.

[2] Fiksdal, A. S., Kumbamu, A., Jadhav, A. S., Cocos, C., Nelsen, L. A., Pathak, J., & McCormick, J. B., "Evaluating the process of online health information searching: A qualitative approach to exploring consumer perspectives",Journal of Medical Internet Research, vol.16, no. 10, 2014.

[3] Fox, S., "The Social Life of Health Information", Retrieved from http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information, 2014.

[4] G. Shenet al., "Depression detection via harvesting social media:A multimodal dictionary learning solution," inProc. 27th Int. Joint Conf.Artif. Intell., pp. 3838–3844, Aug 2017.

[5] Gkotsis, G., Oellrich, A., Velupillai, S. et al. Characterisation of mental health conditions in social media using Informed Deep Learning. Sci Rep 7, 45141, 2017.

[6] Han, S., Huang, H., & Tang, Y., "Knowledge of words: An interpretable approach for personality recognition from social media. Knowledge-Based Systems", Article 105550

[7] Johnson, J. D, "Health-related information seeking: Is it worth it", Information Processing & Management, vol.50, no.5, pp. 708–717, 2014

[8] Keh, S.S., .& Cheng, I.T., "Myers-Briggs personality classification and personality-specific language generation using pre-trained language models", arXiv preprint arXiv.06333, 2019.

[9] Lawhon, L., "Patients Rely on Facebook and Condition-Specific Web Sites to Share Information" , Get Support, and Start Discussions with Healthcare Providers. Retrievedfromhttps://health-union.com/news/online-health-experience-survey/, 2016.

[10] Lykke, M., Price, S., & Delcambre, L, "How doctors search: A study of query behaviour and the impact on search results", Information Processing & Management, vol. 48,no.6, pp.1151–1170, 2012

[11] Majumder, N., Poria, S., Gelbukh, A., & Cambria, E, "Deep learning-based document modeling for personality detection from text", IEEE Intelligent Systems, vol. 32, no.2, pp.74–79, 2017.

[12] Matthew R Jamnik and David J Lane., "The Use of Reddit as an InexpensiveSource for High-Quality Data", Practical Assessment, Research & Evaluation, 2017

[13] M. Park, C. Cha, and M. Cha, "Depressive moods of users portrayedin twitter," inProc. ACM SIGKDD Workshop Healthcare Informat. (HI-KDD), pp. 1–8, 2012

[14] Minjoo Yoo, Sangwon Lee, Taehyun Ha, "Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit", Information Processing & Management,Vol. 56, no.4,2019,Pages 1565-1575,

[15] Monalisha Ghosh; Goutam Sanyal, "Analysing sentiments based on multi-feature combination with supervised learning", International Journal of Knowledge Engineering and Data mining, vol.11, no.4, pp. 391-416, 2019

[16] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networksand linguistic metadata for early detection of depression indicationsin text sequences,"IEEE Trans. Knowl. Data Eng., vol. 32, no. 3,pp. 588–601, Mar. 2020

[17] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki,"Recognizing depression from Twitter activity," inProc. 33rd Annu.ACM Conf. Hum. Factors Comput. Syst., pp. 3187–3196, 2015.

[18] S. Ghosh and T. Anwar, "Depression Intensity Estimation via Social Media: A Deep Learning Approach," in IEEE Transactions on Computational Social Systems, vol. 8, no. 6, pp. 1465-1474, Dec. 2021.

[19] Y. Xue, Q. Li, L. Feng, G. D. Clifford, and D. A. Clifton, "Towardsa micro-blog platform for sensing and easing adolescent psychologicalpressures," inProc. ACM Conf. Pervas. Ubiquitous Comput. AdjunctPublication, pp. 215–218, Sept 2013.

[20] Y. Xue, Q. Li, L. Jin, L. Feng, D. A. Clifton, and G. D. Clifford,"Detecting adolescent psychological pressures from micro-blog," inProc. Int. Conf. Health Inf. Sci.Melbourne, VIC, Australia: Springer, pp. 83–94, 2014.

[21] Zhancheng Ren, Qiang Shen, Xiaolei Diao, Hao Xu, "A sentiment-aware deep learning approach for personality detection from text", Information Processing & Management, vol.58, no.3, 2021.

[22] India is the Most Depressed Country in the World, https://www.indiatoday.in/education-today/gk-current-affairs/story/india-is-the-most-depressed-country-in-the-world-mental-health-day/-2018-1360096-2018-10-1, 2018 ,Accessed on May 8, 2022

[23] Mental Health Illnesses, https://www.nami.org/About-Mental-Illness/Mental-Health-Conditions, Accessed on June 6. 2022

[24] Medline plus, https://medlineplus.gov/mentaldisorders.html, Accessed on May 3, 2022

[25] Evaluate Topic Models: Latent Dirichlet Allocation (LDA), 2019, https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0, Accessed on June 6, 2022