Unsupervised Topic Modeling with LDA for Textbook Content Comprehension, a qualitative survey

Md Iftekharul Mobin[1,2*], Second Author[2,3†] and Third Author[1,2†]

[1*]Department, Organization, Street, City, 100190, State, Country.
[2]Department, Organization, Street, City, 10587, State, Country.
[3]Department, Organization, Street, City, 610101, State, Country.

*Corresponding author(s). E-mail(s): iftekhar.mobin@gmail.com;
Contributing authors: iiauthor@gmail.com; iiiauthor@gmail.com;
†These authors contributed equally to this work.

Abstract

In this research LDA driven exploratory analysis is shown to depict various keywords related to subtle topic in context. It identifies latent topics within textbook lessons, uncovers coherent themes from textual data, aims to improve the curriculum provided English textbook content synthesis and acquisition skill of learners. It is anticipated extracted topics enable readers to comprehend the curriculum material more efficiently. Extensive analysis is conducted to visualize, high impact keywords, co-occurrence patterns and correlation between extracted topics. A prototype mobile app is developed which incorporates topic modeling extracted keywords. Furthermore, qualitative research survey is undertaken to evaluate its effectiveness on end-users (course instructors of Bangladesh's higher secondary school). The challenges, future potential of LDA extracted content integrated mobile app into the learning process is explored. After collecting feedback, word clouds were used to analyze the participants' recommended terms, and the LIWC approach is used to estimate overall sentiment. LIWC score showed positive sentiment and survey process enticed the participants, demonstrates learners eager to use NLP technology driven topic modeling approach in teaching and learning, and there are tremendous opportunities.

Keywords: Natural Language Processing (NLP), Topic Modeling, Latent Dirichlet Allocation (LDA), Exploratory Analysis, Textbook Learning, Coherence

1 Introduction

In Bangladesh there is lacking in effective acquisition, synthesis skill of English language from National Curriculum and Textbook Board (NCTB) curriculum provided textbook [1]–[4]. Specially in the rural area in National Board examinations like SSC, HSC most of the student get poor marks in English subject. It is anticipated students have lacking in understanding context. Topic modeling can play a significant role for context understanding for curriculum provided English textbook. Topic modeling is machine learning technique of Natural Language Processing field, offers a promising unsupervised approach to identify latent topics within provided documents. It can help identify the main themes, concepts, and topics within the textbook's content, enabling instructors to tailor the learning experience to individual students. LDA is one the prominent algorithm which can be used for topic modeling. It provides coherent topics, dominant keywords, latent combination of features that characterizes similarities between topics. In this research LDA extracted keywords are rearranged and incorporated into a mobile app to observe user experience. Some renowned mobile apps are Duolingo, Busuu, Babel, Voxy etc [25]. Across all English learning applications via digital media, 55% have activities for vocabulary learning and other exercises are about 41% [18], [19] includes quizzes, exercises, and game for enhancing learners' comprehension and self-checks [17]. One caveat is of these apps are these are not based on Curriculum Board provided Textbook for learning English hence, could not able to attract a large number of pupils in Bangladesh who are mostly depended on NCBI textbook. To grasp the English language knowledge from curriculum provided textbook a novel approach LDA based unsupervised Topic modeling using textbook corpus is adopted and exploratory analysis is demonstrated in this study. Our anticipation is through this way student can able to interpret meaningful information facilitates students to understand the correlated topics and important keywords related to that topics leads to understand the subtle meaning of the textbook context.

1.1 Research Overview

To detect underlying themes or topics keywords within a Textbook corpus an unsupervised probabilistic topic modeling technique Latent Dirichlet Allocation (LDA) is used in this research. Topic modeling of LDA doesn't directly account for student engagement such as learning tasks in mobile apps. Hence a prototype app is developed and qualitative survey is conducted to observe the instructor's sentiment impact. Qualitative survey research is undertaken to

evaluate the effectiveness of unsupervised topic modeling LDA Bangladesh's National Curriculum Textbook Board (NCTB) provided English Textbook for Higher secondary school education. This study seeks to ascertain if students can learn English better if a mobile app is introduced which includes NLP's LDA driven topic modeling applied extracted keywords and analysis. A prototype mobile app is developed to incorporate the topic modeling extracted keywords into the app. This article presents the key findings and insights from the survey, shedding light on the prospective of learners especially instructors. In the survey questions, it was indicated whether the students, teachers/instructors, and government organizations would find it acceptable and appreciated if textbook information were made available through a mobile app and presented in interactive format. To demonstrate the mobile app idea during the interrogation survey session a prototype is also prepared. Participants were asked for suggestions on how to make the app better and about any shortcomings. After collecting feedback, word clouds were used to analyze the frequency of the participants' recommended terms, and the LIWC approach was used to estimate overall sentiment. The survey's findings show that teachers are eager to use NLP provided extracted keywords technology in teaching and learning, and there are tremendous opportunities.

1.2 Topic Modeling

Different techniques have been developed to perform topic modeling in the unsupervised topic modeling domain of Natural Language Processing (NLP), having their own strengths and limitations. Apart from LDA, Mallet LDA, Structural Topic Model (STM), Hierarchical Dirichlet Process (HDP), Non-Negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA) etc are also prevailing and can be considered for comparative research study.

1.2.1 Topic models comparative analysis

While some variations of LDA, Mallet LDA is considered for large corpus processing and analysis. It focuses on scalability, If large corpus needs to analyze, Mallet LDA might be more suitable. LDA in general can still be efficiently applied to moderately sized corpora. Analyzing topics within the context of metadata, STM could be a better fit. Hierarchical Dirichlet Process (HDP) can be useful when we cannot guess the number of topics in advance. However, as a baseline model LDA is often considered one of the most prominent choices. In this study Textbook corpus is divided into lessons which is a mixture of topics and using LDA expecting to determine which word in the lesson belong to Lesson's topics. LDA produces interpretative results for exploratory topic analysis. The identified topics are represented as distributions over words, making it easy to assign meaningful labels to topics. Provided by most of the libraries and tools, making it easy to implement and can be integrated into existing workflows. Hence, LDA serves as a solid baseline for topic modeling tasks.

2 Latent Dirichlet Allocation (LDA)

LDA model considers documents are mixes of topic, and each topic is a distribution over words. The objective is to derive the hidden topic assignments and the topic-word distributions that most effectively describe the observed documents. The goal of LDA is to uncover these latent topics from a collection of documents without needing any prior labeling or categorization of the content. An expression for the joint distribution of the LDA model is described below:

$$P(\theta_d, z, w \| \alpha, \beta) = P(\theta_d \| \alpha) \prod_{n=1}^{N} P(z_{d,n} \| \theta_d) P(w_{d,n} \| z_{d,n}, \beta) \tag{1}$$

Where $w_{d,n}$ the n_{th} word in document d, $z_{d,n}$ the topic assigned to the n_{th} word in document d, α, β are the Dirichlet LDA model parameters. controls per-document topic distribution, and per topic word distribution. θ_d represent the topic distribution. $P(\theta_d \| \alpha)$ Dirichlet distribution representing the document-topic distribution, $P(z_{d,n} \| \theta_d)$ is the word topic assignment for the n_{th} word in document d, $P(w_{d,n} \| z_{d,n}, \beta)$ is the distribution representing the observed word given a topic. We have chosen LDA for baseline statistical topic modeling tool. However, how many topics are ideal it is needed to determine and also topic modeling quality needs to measure.

2.1 Optimal Topics with Coherence

Coherence score measure how coherent or interpret the words in that topic and estimates number of topic clusters. Coherence score assess the quality of the topics produced by LDA and ensures that the topics generated are statistically significant. Coherence C_{topic} can be expressed as follows

$$C_{topic} = \sum_{i=1}^{N} (\frac{1}{N(N-1)} \sum_{j=1}^{i} PMI(w_i, w_j) \tag{2}$$

Where, $PMI\,(w_i, w_j)$ represent pointwise mutual information statistical association between two words occurring together. PMI score indicates that the two words are more closely related within a topic. $PMI\,(w_i, w_j)$ can expressed as

$$PMI\,(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \tag{3}$$

where $P\,(w_i, w_j)$ is joint probability of occurrence of words w_i and w_j. To calculate the coherence score genism library provides range of options such as $u_{mass}, c_v, c_{uci}, c_{npmi}$. u_{mass} and c_vThese two methods are most popular. For given topic with words $\{w_1, w_2, w_3, \ldots.., w_n\}$ a fixed context window size is provided (default size 10 words) then coherence score is calculated using an equation $\sum_{j=1}^{i} PMI\,(w_i, w_j)$ which provides negative coherence score. c_v can

be expressed as

$$c_v = \frac{1}{N(N-1)} \sum_{j=1}^{i} similarity\,(w_i, w_j) \tag{4}$$

in which $similarity\,(w_i, w_j)$ represent the pairwise similarity between terms based on $PMI\,(w_i, w_j)$ scores. c_v provides a positive coherence score. Higher coherence values indicate that the topics are more coherent and representative of meaningful themes within the text data. Coherence score 0.5 are fairly good [43].

3 Literature Review

LDA based topic modeling has been used for semantic search, ontology exploration, classification, dominant keywords searching in many research studies. For curriculum based textbook study it could be an option is not revealed from rigorous search in online repositories, which we addressed in this research study. This research model is infusing Textbook features into LDA based topic modeling to discover a set of topic-words from the provided document for Textbook comprehension and understanding context. Similar approaches and implications are described below:

3.1 LDA for Textbook Content

According to Krishna Raj [41] the human brain has a propensity to forget a number of facts regarding the events in book. The LDA model can able to scan through large quantities of text in the book and extract intriguing key concepts terms. Thereby, a book can be adapted by learners with the help of the LDA model, greatly refine the learning process. Similar approch proposed by Rani For Hindi books and stories, a topic modeling text summarizing approach proposed in 2020 [42]. By incorporating linguistic features into LDA-based topic modeling, the suggested model is can find a set of topic-words from the given material to understand context. Educational content-based topic modeling for an Intelligent system to develop a tutoring system is proposed in [34] by researcher Stefan Slater in 2017. To understand the linguistic content of mathematical problems, a personalized learning system is suggested that makes use of correlated topic modeling of natural language processing, an approach which can extract important keywords. A variety of significant and useful contents are explored in the context of addressing mathematical difficulties. They demonstrated that topic modeling is a useful method for a personalized learning systems. For key terms detection within articles LDA based topic modeling has been used in many research studies in which dominant keywords reveals future research trends or most prominent topics. Investigation of Julio Guerra in 2013 demonstrated how the LDA model can be utilized for online content linking for any major subject such as: elementary algebra. It can also be used

to simplify context comprehension and content modeling for collections of reference books on the same topic [33]. They concluded that the recommendation provided by LDA topic modeling for online educational systems is promising.

3.2 LDA for Dominant Keywords Determination

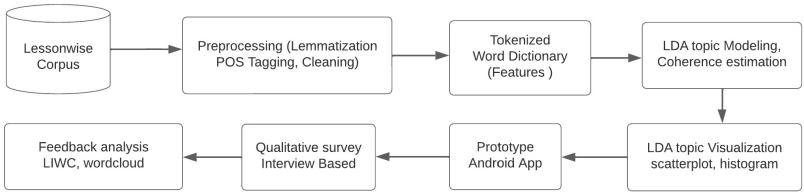
In 2019 Wafa Shafqat [38] proposed an architecture model for better understanding of crowdfunding comments posted by the investors to understand their motive to classify whether comments are scam or legitimate comments. Deep neural network Language modeling either LSTM or RNN encoded embedding vectors are fed into a LDA based topic modeling model to understand the context of discussion trends. Compared to simple Neural Networks (NNs) and non LDA based approach this techniques performs better understanding crowdfunding comments. The capability of LDA-based topic modeling to detect the reseach trend of Bengali news published in web was analyzed by Kazi Masudul Alam in 2020 [35]. Their research demonstrates that using an appropriate corpus and labeled LDA is an effective combination model for predicting news topics efficiently. LDA Labels key terms makes articles easier to read. Another research article of J. Lee published in 2022 conducted experiment on research trends of "COVID-19 and sports [36]. It used LDA and explored latent knowledge connectivity dimension and structures in the articles. Rahul Gupta [37] uses the application of LDA in 2022 to analyze research patterns of "Applied Intelligence" in 3269 articles published between 1991 to 2021. In this research BoW and TF-IDF embedding are used for LDA based topic modeling.

3.3 LDA based Sentiment Analysis

Sentiment analysis has been a key research area of NLP based research domain, where LDA has been applied to determine significant features and those features contributes to segregate sentiments and provide recommendations. LDA based topic modeling has been used in sentiment analysis task. In 2021 Y. Cho published research study of LDA-based topic modeling for sentiment analysis using topic/document/sentence (TDS) model [39]. This article proposed TDS novel approach that combines LDA-based topic modeling for sentiment analysis within documents. H. Wang and et. al examined Chinese people's public perception about Omicron variants on social media Sina Weibo. Social Media's 121,632 omicron related data post were analyzed using LDA-based topic modeling and sentiment analysis [40]. From topic analysis they realized omicron's impact, infection situation, pandemic prevention and control geographically. Hence, it is actually revealed LDA based topic modeling can be used for understanding subtle topics and exploring various facts. Hence, from the above literature review we can infer LDA could help analyze the content of the textbook and identify the main topics covered. This information could then be used to enhance the learning experience.

4 Methodology

The methodology involves pre-processing textual data, training the LDA model on the pre-processed text, and subsequently interpreting and visualizing the generated topics. Data is collected from NCBI's English Textbook for class 9 of higher secondary school. Data mining approach is applied to segregate into subsequent lessons. Then textual data is pre-processed to remove noise, text standardizing, followed by the application of LDA to identify underlying topics, then coherence measurements are applied. Research overview is depicted as follows. ,



LDA algorithm automatically discovers latent topics within the documents based on word co-occurrences. Each topic will be represented by a set of words. Interpret these words to understand the main concepts associated with each topic. Extensive exploratory analysis is conducted to visualize the topic modeling outputs. Analyze the topics generated by the model require manual review and adjustment to ensure the topics make sense.

5 Data Processing and Feature Extraction

a. First NLP's data processing or data mining techniques are applied for meaningful token or feature extraction. Text is converted to Lowercased and Normalized to ensure consistent pre-processing.
   i. **Data cleaning**: unwanted characters, punctuation and special character removed and stop words (such as "and," "the," "is," etc) are removed. Spacy library's English word model and NLTK's stopwords list are used together. Also, words less than two characters are removed such as: I, Hi, Oh etc. Hence, Noise is removed and irrelevant characters, symbols, or data artifacts that have been introduced during data collection or scraping from pdf file to text file generation are separated. Hence, we found a cleaned corpus.
   ii. **Lemmatization**: Root words are collected words to their dictionary form (lemma) is extracted using NLTK's WordNetLemmatizer package. Stemming Reduce words to their base or root form is not used since sometimes it changes the actual words.
   iii. **Part-of-Speech Tagging**: Spacy's English model 'en_core_web_sm' is used to extract interested words (such as noun, verb, adjective) and excluded (CCONJ, AUX, DET, INTJ, PART etc which are Coordinating Conjunction, Auxiliary, Determinator, Interjection, Particle etc) thereby

token is collected for only which are not punctuation, conjunction, symbol etc.

6 Dataset

Text based Suicide and depression detection classification task uses NLP techniques and methodologies which are highly dependent on the quality of dataset, accurate annotation of labels and size of samples. Text based samples are mostly collected from Twitter, Reddit [33], facebook, weibo etc websites donated by various institutions or researchers. Several researchers contributed publicly available dataset [27]. In 2021 the Computational Linguistics and Clinical Psychology CLPsych 2021 workshop organized a Task challenge for detecting suicidal risk [18]. It facilitated participants providing sensitive authentic dataset on the problem of predicting suicide risk from social media Twitter. The dataset for the task includes information who attempted suicide or succeeded along with some control who have not. After collecting dataset from social sites, proper labeling is crucial for training machine learning classifier models. In[17] research study used Twitter post collection API for collecting Tweets and collected Tweets of size 2509 were obtained, of which 216 post were found relevant by 3 Expert psychologists evaluators. Furthermore, using LIWC, dictionary of the Linguistic Inquiry and Word Count [24], which is a linguistic feature analysis software that calculates the degree of positive and negative emotions across a wide spectrum of texts, Tweets were evaluated and results are statistically presented. In 2018 shing et. al [31], and in 2019 Gaur et. al [9] consulted with the professional practitioner psychiatrist to annnotate the dataset and segregated into several categories. [9] contains gold standard dataset of 500 redditors prepared from 2181 redditors post and validated by four practicing psychiatrists following the guidelines outlined in Columbia.Suicide Severity Rating Scale (C-SSRS).

7 Exploratory analysis of Textbook content

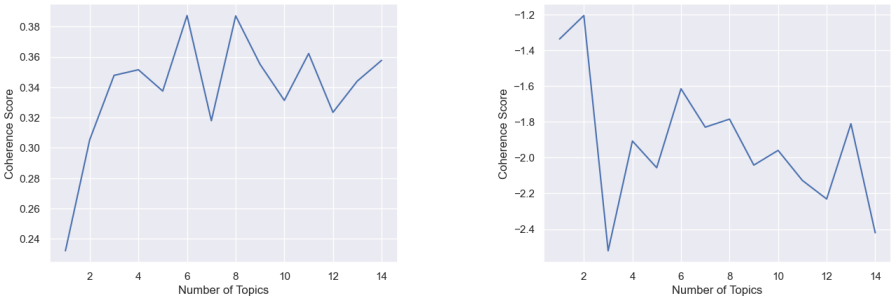
Here whole book is segregated into Lessons and we wanted to explore the important topics within the content. Similar topic words remain together. Therefore, assumptions are, it helps students to understands the words, sentences and context of the book. Ideal number of topics are determined using coherence score.

7.1 Coherence for LDA model

To measure coherence in the context of LDA, following steps are followed:

1. Cleaned document samples are prepared using python's NLP data mining techniques explain in detailed in data reprocessing section. Prepared $T(d_i)$ set of tokens in Documents d_i for i^{th} document samples in corpus D.

2. Doc to BOW corpus dictionary is prepared with Doc2Bow vector. This vector x_d can be represented as where $n(w_i, d)$ denotes the count of words w_i for the document d.
3. Trained LDA Model: During the training phase gensim's MulticoreLDA model with four CPU worker thread is set. Doc2Bow dictionary is applied along with 20 iterations is invoked. The rest of the parameters for LDA model training was default parameter settings of gensim library.
4. Calculate Coherence: To Calculate the coherence score for each LDA model for n number of topics step 3 is iterated for $n = 15$ times.
5. Iteration result coherence score for n number of topics are saved in a list and plotted using seaborn.



(a) Coherence for c_v approach (b) Coherence for u_{mass} approach

Fig. 1: Coherence score to estimate optimal number of Topics

From the chart we can see that six topics are dominant in our provided corpus. The chart shown at the left shows the coherence score for u_{mass} and the right chart represents the score for c_v for multiple iterations. Using 6 topics we can see the output of corresponding topic and top 10 words in a topic.

Topic 01: ['energy' 0.060, 'source' 0.029, 'renewable' 0.018, 'water' 0.016, 'use' 0.013, 'gas' 0.013, 'produce' 0.013, 'green' 0.013, 'warm' 0.013, 'cause' 0.012,]

Topic 02: ['pastime' 0.024, 'computer' 0.024, 'social' 0.023, 'user' 0.022, 'network' 0.020, 'student' 0.019, 'class' 0.017, 'change' 0.016, 'book' 0.015, 'survey' 0.013,]

Topic 03: ['mother' 0.083, 'buy' 0.021, 'love' 0.018, 'child' 0.014, 'worker' 0.014, 'begin' 0.013, 'cultural' 0.012, 'observe' 0.012, 'thing' 0.012, 'language' 0.011,]

Topic 04: ['life' 0.016, 'Bangladesh' 0.016, 'family' 0.015, 'home' 0.014, 'root' 0.014, 'language' 0.014, 'country' 0.013, 'Pakistan' 0.010, 'war' 0.010, 'man' 0.009,]

Topic 05: [‘country’ 0.031, ‘river’ 0.022, ‘India’ 0.022, ‘land’ 0.021, ‘boat’ 0.015, ‘small’ 0.015, ‘population’ 0.015, ‘lake’ 0.013, ‘group’ 0.012, ‘house’ 0.011,]

Topic 06: [‘job’ 0.064, ‘English’ 0.023, ‘learn’ 0.021, ‘teacher’ 0.017, ‘use’ 0.016, ‘dream’ 0.016, ‘think’ 0.016, ‘thing’ 0.015, ‘school’ 0.014, ‘education’ 0.013]

7.2 Count vs Relative Importance measurement

Word frequency $n(w_j, d_i)$ in each document D is measured as below which identifies the most frequent words within each document and across the entire corpus.

$$D = \sum_{d_i \in D} \begin{cases} 1, n\left(w_i, d_i\right) > 0 \\ 0, n\left(w_i, d_i\right) = 0 \end{cases}$$

we can visualize relative importance of any keywords in terms of frequency and plotted inclined with LDA provided weights.

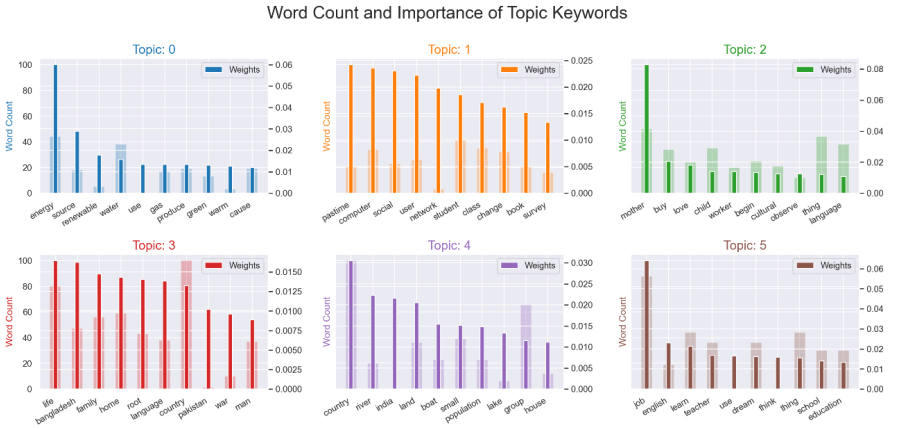


Fig. 2: Word frequency and its relative importance

8 Dominant topic and contribution

In LDA models, each document is composed of multiple topics. But typically, some specific topics are dominant. The following experiment extracts this dominant topic for each sentence and shows the relative weight of the topic and the keywords. It estimated which document belongs predominantly to which topic. How frequently the words have appeared in the documents and the weights of

each keyword in the same chart, words that occur in multiple topics and the ones whose relative frequency is more than the weight.

8.1 Topic-Term Matrix Visualization and Inter-Topic Distance Map

Visualizing the topics and their relationships in a topic model Python library PyLDAvis is used provides an interactive web-based interface to explore and analyze the LDA results of topic modeling. PyLDAvis itself abstracts away much of the underlying mathematical complexity and provides a user-friendly way to generate visualizations and interactively explore topics and their relationships. Key components distance among topics and salient terms are explained below:

8.1.1 Inter-Topic Distance Map

Distance among topics refers to the measurement of similarity between topics in a high dimensional space matrix provided by the LDA model. PyLDAvis library is used to conserve dimensionality reduction using PCA and for calculating distance between topics metric like Euclidean distance or Cosine Similarity . Topic-topic distribution matrix $Q\left(t_1, t_2\right)$ for topic t_1 and topic t_2, distance Dbetween t_1, t_2 can be represented as

8.1.2 Salient Terms or dominant keywords

Salient Terms in a topic are words W that are most strongly associated with specific topic. The mathematical expression for finding salient terms w for a topic t involves, extraction of top $n$words that poses the highest probability scores for topic t in the topic-term matrix $P[t, w]$.

Top 30 most salient terms are showed at right in the bar chart histogram and left figure shows inter topic distance, their size etc. PCA dimensionality reduction technique is applied here to embed the LDA result into a 2D plain scale. Projected the data in lower-dimensional subspace by computing eigenvalues reduced the circle overlapping. Topics that are closer together in the map are more similar in terms of the distribution of words.

9 Englisher Mobile App

A mobile application (Englisher) is being developed with content from the NCTB's English Textbook for class 9. The extracted keywords are organized into lessons and furthermore quiz is introduced as an exercise. Each sentence's and word's Bengali meaning is provided in accordance with the lesson. Students can take quizzes, and their results are recorded in the history so that history can be reviewed and performance can be improved by more practice in the future.

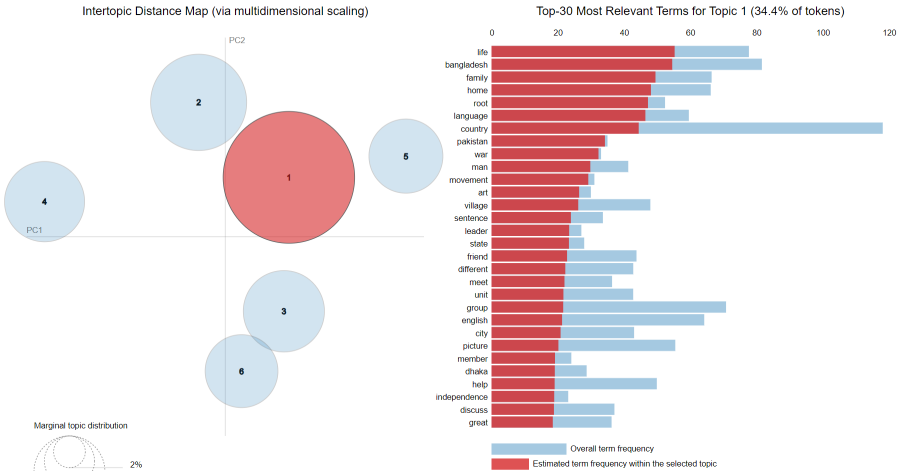


Fig. 3: Topic model co-occurrence visualization with dominant keywords'

10 Qualitative survey

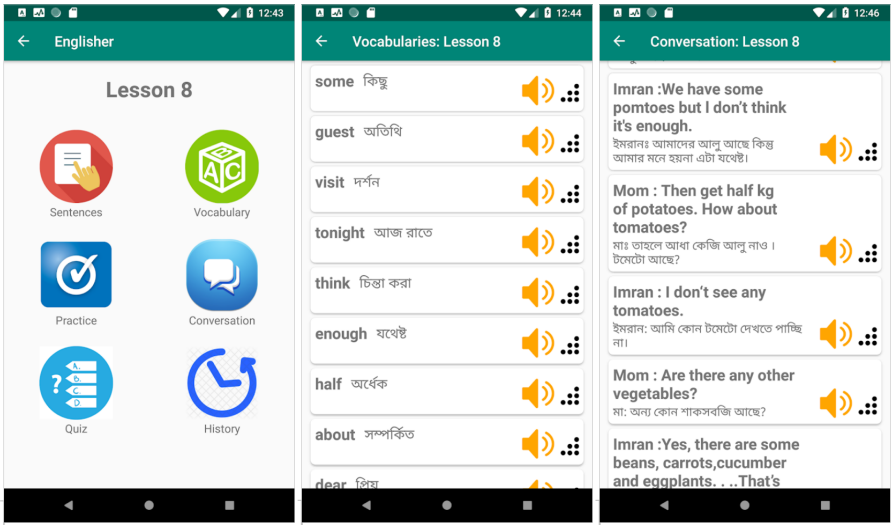
In the survey questions, it was indicated whether the students, teachers/instructors, and government organizations would find it acceptable and appreciated if textbook information were made available through a mobile app and presented in interactive format. To demonstrate the mobile app idea during the interrogation survey session prototype app Englisher is prepared. Participants were asked for suggestions on how to make the app better and specify shortcomings. Presumably It provides an insight of teacher's emotion about inclusion of mobile technology in higher secondary English education system.

10.1 Survey Planning

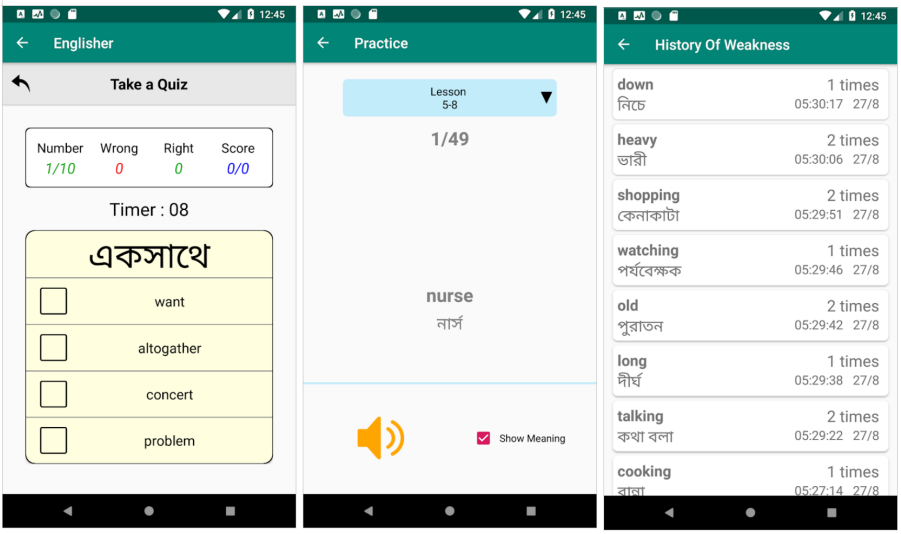
The survey was conducted over a period of four weeks, with 50 High schools in Dhaka and Bogura district of Bangladesh. It encompasses only English subject areas Teachers who teaches in high schools from class six to class Ten and teaches regularly in the school. A questionnaire was distributed to teachers allowing us to gather questionnaire answer.

10.1.1 Survey participants

During the survey standard participants were chosen emphasizing infrastructure quality, teaching experience, class size etc. At the beginning from 100 institution were selected. Then half of them were excluded since those institutions infrastructure's overall quality and condition were not above average. Among the chosen samples 76% were good and 26% considered average institutions. Privately held 45%, 32% partially government and 22% are government institutes. Over 1000 students study in almost 40% of these institutions and



(a) Lesson wise exercise



(b) Quiz with Vocabulary

Fig. 4: Englisher Mobile app for Learning LDA based topic model words

sizable number of pupils are present in each section and class. 38% class have a size greater than 50. So, we can presume that the participating teachers have quite a bit of experience teaching sufficient number of students.

10.2 Survey Results:

We have done extensive analysis with the survey data collected. In our data collection highest priority is given for the secondary class student teachers who teach between 6-10th class about 46%. High school, KG college and KG High school. Details about the statistics are depicted in the following figure. Adjacent chart explains the percentage of teachers who teach in which class. Hence, from these two figures we can get a vivid image of collected dataset resources about the participating teachers.

10.3 Training Dataset

In this research we have used dataset from [9]. For training classifier in this research 2019's Gaur et. al [9] dataset is used. Compared to the existing four-label classification scheme (no risk, low risk, moderate risk, and high risk), this dataset introduced 5 lavel classification suicide indicator, ideation, behavior, attempt and another extra category incorporating supportive category. Supportive category represents whenever someone shows empathy and condolance for a suicidal post.

10.4 Testing/Validation Dataset

For testing dataset is collected from kaggle. It is an opesource dataset publicly available collected from reddit website by a pushshift API contained suicide and depression category. This publicly available Reddit datasets in Kaggle Website comprised of 232,074 post annotated for binary classification as suicidal or non-suicidal in [1] for detecting suicidal ideation. The dataset is a collection of posts from the "SuicideWatch" and "depression" subreddits of the Reddit platform. All posts that were made to "SuicideWatch" from Dec 16, 2008(creation) till Jan 2, 2021, were collected while "depression" posts were collected from Jan 1, 2009, to Jan 2, 2021. In this research trained classifier is applied to detect class on this two category. Main objective isto determine the suicide categories (indicator, ideation, behavior, attempt, supportive) within this dataset. Document length frequency and token distribution is depicted in Figure **??**. From the frequency distribution we can see some of the document sizes are very large. Hence, during the training process we chopped the sentences into multiple sentences keeping the label same.

10.5 Data Processing and Models

10.5.1 Data Pre-processing

Social media dataset are mostly Text data which needs data pre-processing, cleaning, feature extraction and data mining related NLP tasks. NLP based text data contains noises such as: unnecessary quotes, special characters, punctuation etc. Moreover, morphological analysis is needed to retrieve root words followed by stemming, lemmatization. Then, sentences are divided into equal-length fragments, and null word padding is applied as needed. Words within a

phrase are now referred to as tokens or features, and the dataset is shown as a corpus. Special features/tokens are further preprocessed and filtered using text data feature extraction tools and methods. Features are passed through a process in which features are converted to corresponding IDs and sentences which contains a series of IDs are represented a vector. The embedding is another term that is frequently used in relation to vector text analysis. Various Vectorization methods are present. Traditional vectorization method provide weights to terms/words mainly based on frequency of words within the sentence and documents, rather than its importance and contextual meaning. Also, how does a particular word or term create impact on the neighboring words is not taken into consideration, since these models does not have any prior knowledge of any words. Hence, various neural network based language models are proposed which are pretrained on massive amount of dataset. These models mainly carries weight which represents word to word relationships and most cases can provide contextual meaning of given sentence based on pretrained dataset knowledge. Deep learning models recently showed remarkable achievements in this case representing corresponding knowledge.

10.5.2 Classification

Machine learning and Deep Learning models are particularly used for Text classification and ML for feature selection or extraction in several studies [5, 6, 42]. These extensive reviews reveal Deep learning methods receive more attention and perform better than traditional machine learning methods whereas in some cases when extracted or filtered features are fit into training process models are able to perform better. NLP techniques are applied for the annotated dataset collected from Twitter, Reddit, Facebook, instagram, Weibo [37] etc. After that various deep learning and machine learning models are trained for classification of suicide and depression. Then trained models are applied to determine the correct class of given text. [19] Provided a through investigation about passed research techniques, features, datasets, and performance metrics [6, 42].

10.5.3 Data Learning Models

Among Several Deep Learning approaches most successful NLP classifier for segregating Depression and suicidal task are CNN, LSTM, GRU, XLNET, BERT, Variants of BERT RoBERTa and variants of CNN such as: CNN-BiLSTM etc [1, 30, 37] also showed promising results.

10.5.4 Feature Selection

The feature selection procedure has a substantial impact on the performance of machine learning and deep learning models because it lowers noise in the trained dataset, enabling the model to accurately understand data patterns. LIWC, LDA, LSA, n-gram analysis [24, 33] etc are used as features analysis tools. Most dominant approaches are n-gram word frequency based approach

TF-IDF. Apart from the Deep learning model n-gram Traditional feature retrieve based analysis conducted in some research papers. In 2017 Shen et. al [**?**] collected several forms of features comprised of six chorots, namely, social network features, user profile features, visual features, emotional features, topic-level features, and domain-specific features and prepared a feature rich dictionary. This multimodal depressive dictionary learning model was used to detect the depressed users on Twitter using machine learning models.

Most dominant approaches are Word2Vec, XGBoost, SVM, Random Forest, other regression MLs. Typical word embedding approaches TF-IDF and Word2Vec, and CNN–BiLSTM are applied in [1]. Using LIWC features, XGBoost ML together surpasses the accuracy of CNN–BiLSTM in [1]. Others have used machine text Summarization based feature extraction strategy followed by classification for depression detection is applied in [44]. [4]built a set of baseline classifiers using lexical, structural, emotive and psychological features extracted from Twitter posts. Then baseline classifiers are updated by building an ensemble classifier using the Rotation Forest algorithm and a Maximum Probability voting classification decision method. [6] This paper provided an excellent overview of 75 studies in between 2013 and 2018 outlining the methods of data annotation for mental health status, data collection and quality management, pre-processing and feature selection, and model selection and verification.

10.6 Other Tools or Methods

Apart from the Machine Learning or statistical model approaches there are some sites, helpline numbers and apps [21] that contains various facts, statistics, tips, tools, healthline numbers and ways to handle suicidal depression [**?**].

nce, visualizing the result we can determine the suicidal tendency within depression post. Also, N-gram based analysis is conducted and frequency of Terms and connections of words or phrases are analyzed in this research scope. More often topic modeling clustering is used to determine latent topic and understand latent text network. From the network various facts can be revealed related to suicide and depression. This whole process is depicted in Figure **??**

11 Results Analysis

First we started our experiment with document length distribution. The length of document and term frequency within the corpus is visualized in Figure **??**. From the distribution we can see that some of the document length are excessive long and contains more than 1000 tokens (within Twitter and also Reddit both Dataset). Depression class document length are usually shorter in length. Depression document length are tend to be smaller than suicide document length.

Short sentence does not carry much terms and hence does not carry enough information to be classified confidently by classifier algorithms. We started

reducing the numbers of samples based on document length. By reducing the samples based on numbers of tokens present in a document (see Figure **??**). Documents length versus category frequency information is showed in this chart. This charts explains if we filter out the shorter comments suicide post become dominant class and depression post become outnumbered. The difference showed an exponential pattern as length of document increases. Test dataset Reddit data distribution among depression and suicide class distribution ratio is equal. Filtering the class we have seen an interesting fact that depressed people does not want to comment very long.

11.1 N-gram Analysis

11.1.1 Uni-gram

Dataset is split into separate tokens after preprocessing and uni-gram generated. Based on frequency of words wordcloud is generated from these unigrams. Frequency based comparison between two categories is conducted for depression and suicide for Test dataset in Figure **??**. Main objective was to get top ranked words from Depression and Suicide corpus. After experiments we have seen There are similarities between the top ranked words those are occurring frequently. They tend to use slang and abusive terms compared to suicidal attempt thinking people. Rather suicidal depressed people want to share their thoughts with others using longer post. However, it does not reveals any clues in terms of hypothetical relationships between the two category. It is difficult find pattern in which we can determine the depression and suicidal thought. So far we found some pattern

To understand the term occurring frequently in two different classes scatterplot library is used for visual analysis. From the above two scenario we can see that there is a pattern that people used to say more slang and abusive words when they are depressed. It is also interesting that there are many words have high frequency such as depression or depressed but belongs to suicide class. One important fact is revealed here is that we can see although suicide, suicidal these words has high frequency in Suicide class but depression, depressed also occurred in parallel with high frequency. Here several experiments can be conducted for exploratory analysis with scattertext library for terms significance. However, this library is computationally heavy for larger dataset for visualization. Another drawbacks is this library have significant focus on the terms based analysis. We have used simple vectorization methods by which we can have greater control on dataset and experiments code.

11.1.2 Bi-gram

First unigram is computed and analyzed then bigram is calculated for both categories. The bigram frequency showed there are some common terms like “mental health”, “feel like”, “make feel”, “high school” etc showed high occurrences in the dataset. Hence, we started to understand its pattern in the corpus. For analysis we have considered [’high school’, ’mental health’, ’best friend’,

'feel like', 'really want', 'suicide thought', 'friend family'] these bi-grams and wanted to explore its surrounding context for each category. We called this special bigrams since it showed importance in the suicidal and depression both categories appeared highly frequent matter. We want to analyze how these words have impact with its neighboring words.

To explore the impact of special bi-grams on the samples, special bi-gram terms containing samples are filtered from dataset. After that using lebel encoder bigrams are encoded as integers and then chord diagram is generated depicted in Figure **??** to find meaningful relationship within the samples between the bigram features.

From this two chord diagram interesting sentence can be inferred. Such as: from the depression class ? self centered person is depressed, having suicidal thought, want to go somewhere to live, spend happy moments and so on. For the suicidal class category suicidal attempt thinking people, have mental health issue, they want to share though with high school friends, best friends, friends and family members, having suicidal thoughts and so on.

Tri-grams or above did not reveals much meaning information, mostly does convey some meaningful information and therefore excluded for further experimental consideration.

12 Classification Results

To segregate the Reddit suicide dataset into different categories of suicide first we have created a classifier using different classification techniques. Since our objective is not making highly accurate classifier. Following approach is applied in this study

- Pre-processed and useful features are used from Twitter's 500 post CSSR dataset for Training classifier
- Used count vectorizer and TFIDF transformer to generate vectors for the dataset
- Trained classifier to determine the categories of various suicidal intensities

We have used simple gridsearch technique of sklearn library and from a list of various classifiers applied on the dataset, we have chosen highest accurate classifiers to determine different label of suicidal risk. so that it can recognize the category. We have used classifiers "K Nearest Neighbors", "Linear and RBF SVM", "Gaussian Process", "Decision Tree", "Random Forest", "Vanilla Neural Net", "AdaBoost", "Naive Bayes". Using various set of parameters, from the result and experiments we found almost 60% accuracy for SVM model to predict the suicide intensity categories. From various set of values of SVM we found degree=2, gamma=0.7, kernel=rbf showed the highest accuracy.

12.0.1 Suicidal Intensities visualization

How much depression can trigger suicidal thoughts is an interesting question. In this study classifier is trained on the suicidal intensity. Then trained classifier is applied on the Depression/Suicide class dataset. From various machine learning models we have found SVM is a good performing model. SVM classifier is applied for the TFIDF vectorizer embedding (see results in figure **??**) and also for Word2vec pretrained vectorizer model. The results are shown in figure **??**. From the results we can see that suicidal ideation between depression and suicidal categories number of samples are very similar. Within depression more number of samples are showed suicidal indicator category compared to suicide which is an interesting result. Suicidal behavior and attempt is comparatively high within the suicidal category than depression. Hence, figure **??** result seems to be pretty obvious, except for suicidal ideation category. Also for the suicidal indicator symptoms are higher within the depression category.

For the word2vec vector embedding scenario supportive and indicator categories results are almost similar in depression or suicide both classes. There is slight difference is shown for suicidal ideation and within suicide class, suicidal ideation is slight higher. Except the behavior and attempt category for the rest categories depression and suicide showed almost similar number of samples.

13 Discussion

From the result it is revealed that suicide categories shown within depression and suicide class vividly. Specially suicidal ideation, indicator showed similar patterns. The number of samples within depression and suicide is almost similar for this two categories. Hence, we can infer depressed person comments showed suicidal ideation and suicidal indicating symptoms. Suicidal behavior and attempt showed higher number of samples within the suicide category compared to depression category. All these results seems very logical results. Although from the results mathematical formulas are not derived in this research study since results are susceptible to chosen classifier, chosen dataset, pretrained models vectors or embedding provided to the classifier.

14 Conclusion

Suicidal risk estimation task and classification samples to determine suicidal risk within social websites and blogs, techniques are discussed before. According to suicidal category previous work has been done before. However, to what extent depression level triggers suicidal risk is not yet discussed before. Also it is difficult to determine since depression and suicide categorical variables are independent factor. There is not any underlying correlation. Several research conducted to segregate which post is suicidal and which one is depression various classifiers are proposed. Extensive work has been done to improve the classification accuracy by adopting most powerful vectorization techniques that uses cutting edge NLP models BERT and its various variants. Research has

also been conducted on how much severity label of suicide within a post is studied.
The input format for the above table is as follows:

References

[1] Theyazn HH Aldhyani, Saleh Nagi Alsubari, Ali Saleh Alshebami, Hasan Alkahtani, and Zeyad AT Ahmed. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. *International journal of environmental research and public health*, 19(19):12635, 2022.

[2] A Beck and M Kovacs. weisman a. *Handbook ofpsychiatric measures: Beck Scale for Suicide Ideation. American Psychiatric Association*, 2000.

[3] Aaron T Beck, Calvin H Ward, Mock Mendelson, Jeremiah Mock, and John Erbaugh. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571, 1961.

[4] Pete Burnap, Walter Colombo, and Jonathan Scourfield. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 75–84, 2015.

[5] Gema Castillo-Sánchez, Gonçalo Marques, Enrique Dorronzoro, Octavio Rivera-Romero, Manuel Franco-Martín, and Isabel De la Torre-Díez. Suicide risk assessment using machine learning and social networks: a scoping review. *Journal of medical systems*, 44(12):205, 2020.

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

[7] Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion proceedings of the the web conference 2018*, pages 1653–1660, 2018.

[8] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137, 2013.

[9] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, pages 514–525, 2019.

[10] Jana M Havigerová, Jiří Haviger, Dalibor Kučera, and Petra Hoffmannová. Text-based detection of the risk of depression. *Frontiers in psychology*, 10:513, 2019.

[11] Keith Hawton, Carolina Casañas i Comabella, Camilla Haw, and Kate Saunders. Risk factors for suicide in individuals with depression: a systematic review. *Journal of affective disorders*, 147(1-3):17–28, 2013.

[12] Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, et al. Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, 80:56–86, 2022.

[13] Thomas E Joiner Jr, M David Rudd, and M Hasan Rajab. The modified scale for suicidal ideation: Factors of suicidality and their relation to clinical and diagnostic variables. *Journal of abnormal psychology*, 106(2):260, 1997.

[14] Sören Kliem, Anna Lohmann, Thomas Mößle, and Elmar Brähler. German beck scale for suicide ideation (bss): psychometric properties from a representative population survey. *BMC psychiatry*, 17(1):1–8, 2017.

[15] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

[16] Kuiliang Li, Xiaoqing Zhan, Lei Ren, Nan Liu, Lei Zhang, Ling Li, Ting Chen, Zhengzhi Feng, and Xi Luo. The association of abuse and depression with suicidal ideation in chinese adolescents: a network analysis. *Frontiers in psychiatry*, 13, 2022.

[17] Yolanda López-Del-Hoyo and Pedro Cerbuna. Exploring the risk of suicide in real time on spanish twitter: Observational study. *JMIR Public Health and Surveillance*, 8(5), 2022.

[18] Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. Community-level research on suicidality prediction in a secure environment: Overview of the clpsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, 2021.

[19] Anshu Malhotra and Rajni Jindal. Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, page 109713, 2022.

[20] Paulo Mann, Aline Paes, and Elton H Matsushima. See and read: detecting depression symptoms in higher education students using multimodal

social media data. In *Proceedings of the International AAAI Conference on Web and social media*, volume 14, pages 440–451, 2020.

[21] Laura Martinengo, Louise Van Galen, Elaine Lum, Martin Kowalski, Mythily Subramaniam, and Josip Car. Suicide prevention and depression apps' suicide risk assessment and management: a systematic assessment of adherence to clinical guidelines. *BMC medicine*, 17(1):1–12, 2019.

[22] Alexander McGirr, Johanne Renaud, Monique Seguin, Martin Alda, Chawki Benkelfat, Alain Lesage, and Gustavo Turecki. An examination of dsm-iv depressive symptoms and risk for suicide completion in major depressive disorder: a psychological autopsy study. *Journal of affective disorders*, 97(1-3):203–209, 2007.

[23] Minsu Park, David McDonald, and Meeyoung Cha. Perception differences between the depressed and non-depressed users in twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 476–485, 2013.

[24] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[25] Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. The columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry*, 168(12):1266–1277, 2011.

[26] Lenore Sawyer Radloff. The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3):385–401, 1977.

[27] Esteban A Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. A dataset for research on depression in social media. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pages 338–342, 2020.

[28] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844, 2017.

[29] Yanmei Shen, Wenyu Zhang, Bella Siu Man Chan, Yaru Zhang, Fanchao Meng, Elizabeth A Kennon, Hanjing Emily Wu, Xuerong Luo, and Xiangyang Zhang. Detecting risk of suicide attempts among chinese

medical college students using a machine learning algorithm. *Journal of affective disorders*, 273:18–23, 2020.

[30] Nisha P Shetty, Balachandra Muniyal, Arshia Anand, Sushant Kumar, and Sushant Prabhu. Predicting depression using deep learning and ensemble algorithms on raw twitter data. *International Journal of Electrical and Computer Engineering*, 10(4):3751, 2020.

[31] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 25–36, 2018.

[32] Om P Singh. Startling suicide statistics in india: Time for urgent action, 2022.

[33] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893, 2019.

[34] Julio C Tolentino and Sergio L Schmidt. Dsm-5 criteria and depression severity: implications for clinical practice. *Frontiers in psychiatry*, 9:450, 2018.

[35] M Vuorilehto, HM Valtonen, T Melartin, P Sokero, K Suominen, and ET Isometsä. Method of assessment determines prevalence of suicidal ideation among patients with depression. *European Psychiatry*, 29(6):338–344, 2014.

[36] MS Vuorilehto, Tarja K Melartin, and ET Isometsä. Suicidal behaviour among primary-care patients with depressive disorders. *Psychological Medicine*, 36(2):203–210, 2006.

[37] Xiaofeng Wang, Shuai Chen, Tao Li, Wanting Li, Yejie Zhou, Jie Zheng, Qingcai Chen, Jun Yan, Buzhou Tang, et al. Depression risk prediction for chinese microblogs via deep-learning methods: content analysis. *JMIR medical informatics*, 8(7):e17958, 2020.

[38] Owen Whooley. Diagnostic and statistical manual of mental disorders (dsm). *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*, pages 381–384, 2014.

[39] Janet BW Williams. A structured interview guide for the hamilton depression rating scale. *Archives of general psychiatry*, 45(8):742–747, 1988.

[40] Ying Xu, Juan Qi, Yi Yang, and Xiaozhong Wen. The contribution of lifestyle factors to depressive symptoms: A cross-sectional study in chinese college students. *Psychiatry research*, 245:243–249, 2016.

[41] Jiayu Ye, Yanhong Yu, Qingxiang Wang, Wentao Li, Hu Liang, Yunshao Zheng, and Gang Fu. Multi-modal depression detection based on emotional audio and evaluation text. *Journal of Affective Disorders*, 295:904–913, 2021.

[42] Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46, 2022.

[43] Le Zheng, Oliver Wang, Shiying Hao, Chengyin Ye, Modi Liu, Minjie Xia, Alex N Sabo, Liliana Markovic, Frank Stearns, Laura Kanov, et al. Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. *Translational psychiatry*, 10(1):72, 2020.

[44] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: learning multi-modalities with user post summarization for depression detection on social media. In *proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 133–142, 2021.