Archives of Computational Methods in Engineering (2023) 30:1819–1842
https://doi.org/10.1007/s11831-022-09863-z

SURVEY ARTICLE

# Mental Health Analysis in Social Media Posts: A Survey

**Muskan Garg**[1]

Received: 27 August 2022 / Accepted: 5 November 2022 / Published online: 3 January 2023


**Abstract**
The surge in internet use to express personal thoughts and beliefs makes it increasingly feasible for the social NLP research community to find and validate associations between *social media posts* and *mental health status*. Cross-sectional and longitudinal studies of social media data bring to fore the importance of real-time responsible AI models for mental health analysis. Aiming to classify the research directions for social computing and tracking advances in the development of machine learning (ML) and deep learning (DL) based models, we propose a comprehensive survey on *quantifying mental health on social media*. We compose a taxonomy for mental healthcare and highlight recent attempts in examining social well-being with personal writings on social media. We define all the possible research directions for mental healthcare and investigate a thread of handling online social media data for stress, depression and suicide detection for this work. The key features of this manuscript are (i) feature extraction and classification, (ii) recent advancements in AI models, (iii) publicly available dataset, (iv) new frontiers and future research directions. We compile this information to introduce young research and academic practitioners with the field of computational intelligence for mental health analysis on social media. In this manuscript, we carry out a quantitative synthesis and a qualitative review with the corpus of over 92 potential research articles. In this context, we release the collection of existing work on suicide detection in an easily accessible and updatable repository:https://github.com/drmuskangarg/mentalhealthcare.

## 1 Background

According to World Health Organization, more than 0.8 million people die out of suicide every year. According to the recent report of Centers for Disease Control and Prevention (CDC) WISQARS in 2019 *Leading Causes of Death Reports*, *suicide* is the tenth leading cause of death in United States. According to the official data of USA, in every 11.1 min one person commits suicide.[1] According to the latest available data, the statistics of Canada estimates 4157 suicides in 2017, making it the ninth leading cause of death. The clinical psychologists and academic researchers come across increasing number of mental health problems and its exposure to the social media platforms during COVID-19 pandemic lockdown. The pandemic has long-term impacts on the mental health and wellness of masses due to economic insecurity and isolation. The suicide cases have adverse physical, economical, and emotional impact on social well-being. Early suicide risk prediction may control the suicide rate by reporting the need of necessary steps to take preventive measures.

As per reports released in August 2021,[2] *1.6 million people* in England were on waiting lists for mental health care. As per estimation, *8 million people* could not get specialist help as they were not considered *sick enough* to qualify. This situation underscores the need for automation of mental health detection from social media data where people express themselves and their thoughts, beliefs/emotions with ease. These writings contain heterogeneous, unstructured and ill-formed data which is human-readable but difficult

✉ Muskan Garg
muskanphd@gmail.com

[1] University of Florida, Gainesville, FL 32601, USA

[1] https://suicidology.org/wp-content/uploads/2021/01/2019datapgsv2b.pdf.

[2] https://www.theguardian.com/society/2021/aug/29/strain-on-mental-health-care-leaves-8m-people-without-help-say-nhs-leaders.

to interpret automatically by a system. Recent studies on predicting suicidal tendency on social media data by using machine learning, ML [1–5] models are more successful as compared to the medical records [6] and paved the way to explore deep learning, DL [7–12] and computational intelligence techniques [13] for quantifying suicidal tendency. We acknowledge that we limit the scope of our study to stress, clinical depression, and suicide risk.

### 1.1 Motivation

The labour-intensive engineering with traditional clinical psychology is a theoretical approach to identify signs of suicidal tendencies. This subjective approach follows the time consuming face-to-face interaction. 80% of people who are at risk are not comfortable in disclosing the level of stress and anxiety that they may have [14]. Further, increase in the levels of stress and anxiety may align thoughts of a person to suicidal tendencies. Progressive studies on suicide prevention [15] has enriched the research community with dataset, resources and provides motivation for new-frontiers.

In the past, we closely observe cross-sectional studies for identifying mental disorder in a given self-reported text using AI models for classification and categorization. We witness progressive studies on finding mental disorder levels from longitudinal data which provides useful insights. Based on these interesting investigations, research community may report the available and required resources in near future for medical assistance to people at risk. These developments reduces the dependency of in-person sessions with therapist/clinical psychologist and thus, cost of identifying people at risk. As evident from recent deployments of suicide risk detection model by Facebook [16], we may identifying potential users at risk and offer them help in near future.

As evident from studies in the past, social media platforms has strong association with feelings expressed by users [17–19]. About 8 out of 10 people tend to disclose their suicidal tendencies on social media [20]. Mental health prediction from social media [21] facilitates suicidal risk assessments [22] and early detection of suicidal tendencies by using emotion spectrum from social media user's historical timeline [7] due to the presence of *Papageno effect* [23]. Such path-breaking developments intensifies faith in developing learning-based mechanisms to capture mental health levels using language.

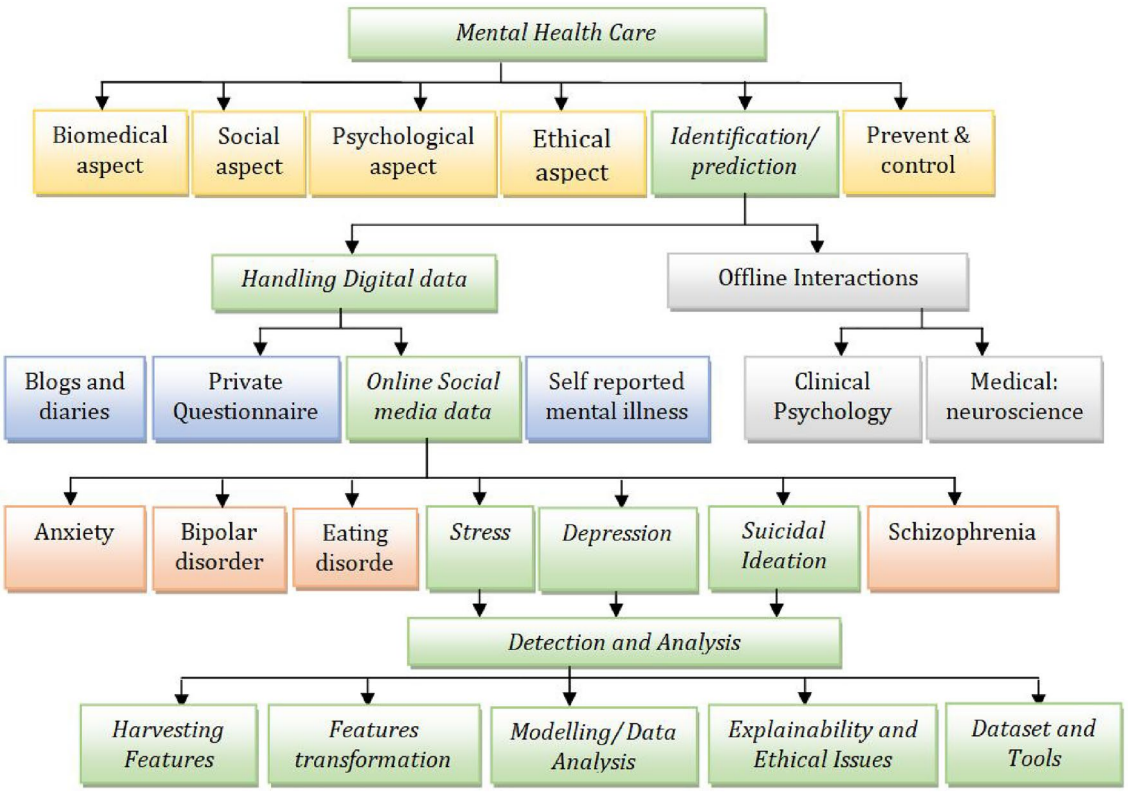


**Fig. 1** Taxonomy on mental healthcare

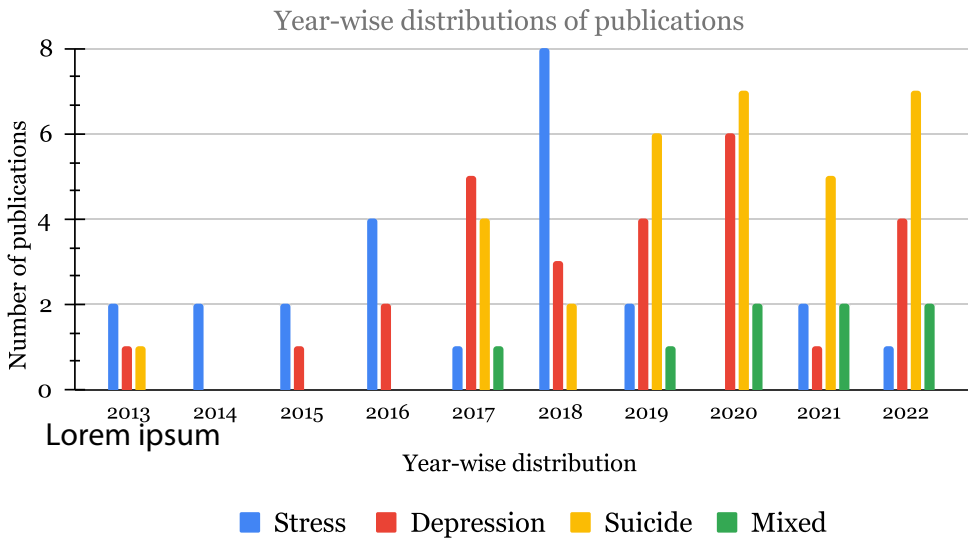


**Fig. 2** Year-wise distribution of number of publications on mental disorders

## 1.2 Mental Healthcare: A Taxonomy

After comprehensive investigation in NLP-centered problems and social computing of mental health, we introduce a unique taxonomy for mental healthcare as shown in Fig. 1. We further examine mental healthcare as an interdisciplinary domain of computational linguistics and human–computer interaction to automate the predictions.

We discover different aspects of mental health domain and observe both independent and integrated studies for each aspect. In this section, we describe different aspects of mental healthcare. Recent developments with ECG signals, Electronic Health Record (EHR), demographic information and other medical reports exemplifies the available data and resources for neuroscience-based studies known as *biomedical* domain. The *social aspect* of mental health studies is closely associated with the human-behaviour within the society. The *psychological aspect* is inclined towards theorizing the thoughts on mental health. The *ethical aspect* is concerned with the security of the data which mean *to what extent* and *in what manner* can it be used [24]. The *prevention and control* measures for mental health issues are examine independently or in association with any of the corresponding aspects.

Conventionally, identification of people at risk is the carried out on digital data and the traditional offline interactions. The use of traditional method is decreasing because of social stigma and unavailability of clinical psychologists. Digital mental health comprises of *blogs and diaries* of a user, information filled in *private questionnaires* or Google forms, *self-reported mental illness* (voluntarily), and *online social media data*. We choose to explore the online social media language resources which contains heterogeneous type of information such as linguistics, user-metadata, social metadata, and multimedia data. The scope of this manuscript is to deal with text in social media platforms (Twitter, Reddit, Sina Weibo) and it occasionally contains images for stress, depression and suicide risk on social media. There are studies over multimodal (images, audio and visual) social media platforms (Instagram, Youtube[3]) in the past which is beyond the scope of this manuscript due to different nature and semantics of available resources.

Social NLP research community investigate six other social mental health problems in social media data which may/may be directly associated with the suicidal tendencies. Moreover, among nine mental health problems, stress, clinical depression and suicidal risk detection are the most widely studied areas on social media [18]. The success of existing AI models have given new research direction to investigate this problem and motivate academic researchers to find its practical application in industry.

## 1.3 Corpus Overview

We perform in-depth analysis for 92 research articles which are further classified as 9 articles for *stress*; 32 articles for *depression*; 37 articles for *suicide risk*; 14 articles for two or more *mental disorders*. The year-wise distribution of publications is shown in Fig. 2 which and top 3 venues are CLPsych, ACL, and AAAI as observed from Fig. 3. We advocate that the research articles on stress and suicide risk detection are fewer than the article on identifying clinical depression.

The area of interest by research community has evolved from *social venues* [25–28], to *Human–Computer Interaction* venues [29, 30], and *Computational Linguistic* domain of computer science [31, 32]. Existing studies have addressed the concerns on dataset and its ethical constraints

[3] https://dcapswoz.ict.usc.edu/.

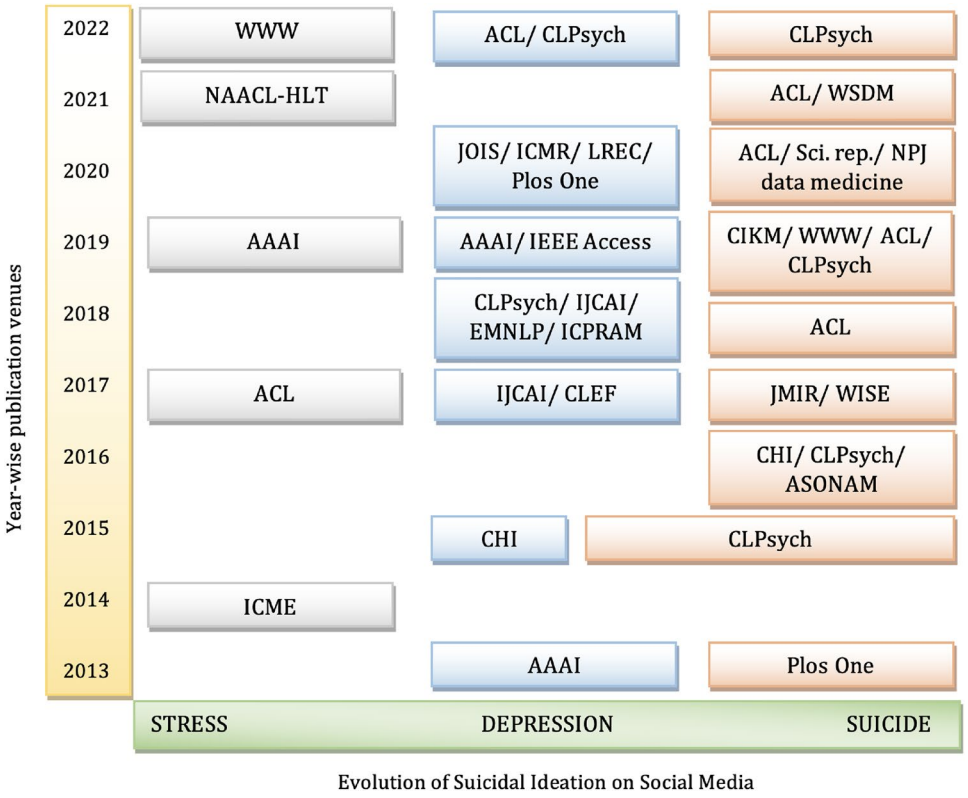


**Fig. 3** Year-wise distribution of publication venues

[17, 33–35]; multi-modal feature extraction [36–39]; classification techniques [1–4, 9, 39–41], graph learning approach [2]; use of the Internet of Medical Things for real-time applications [34]; noisy label problem in dataset annotations [42]; and improvement over the attention mechanisms [9, 11, 43, 44].

### 1.4 Scope of the Study

The research domain of Mental Illness Detection and Analysis on Social media (MIDAS) has evolved for less than a decade [45]. In the past, the honest disclosure of public opinion about privacy concerns demands the need of explainable and responsible AI models [46]. An in-depth study about dataset and its ethical issues were explored in a systematic review for statistical analysis of mental health dataset [17]. A critical review of 75 research articles on the mental health issues from 2013 to 2018 study the design and research methods [18]. A short survey address the concerns of association between social media data and mental health prediction [47]. Recent advancements yield comprehensive study of features and online behaviour patterns for mental health prediction with DL mechanisms [48].

We focus on direct contributions in the field of suicide risk detection by identifying the extent of suicidal tendencies which shows new research direction to build real-time NLP-centered applications to handle the problem of mental disorders. Our major contributions are:

- Classification of heterogeneous social media features.
- State-of-the-art AI models for stress, depression and suicide risk detection and analysis.
- Available tools, resources, and dataset in this research domain.
- Highlight the open challenges and new frontiers.

We further structure this work in different sections. Section 2 presents the classification of different features of social media data for suicide risk detection. We elaborate embedding and feature enhancement in this domain. Section 3 give summary of automated learning based techniques for quantifying mental health. We further compose a list of available dataset and other tools/resources. Section 4 highlights the open challenges and new frontiers. Finally, Sect. 5 concludes the manuscript.

## 2 Features from Social Media Data

With this background, *data curation* becomes the most challenging task as it contains unstructured/semi-structured, user-generated and ill-formed nature. Recent advances in the development of classifiers [49] enrich natural language

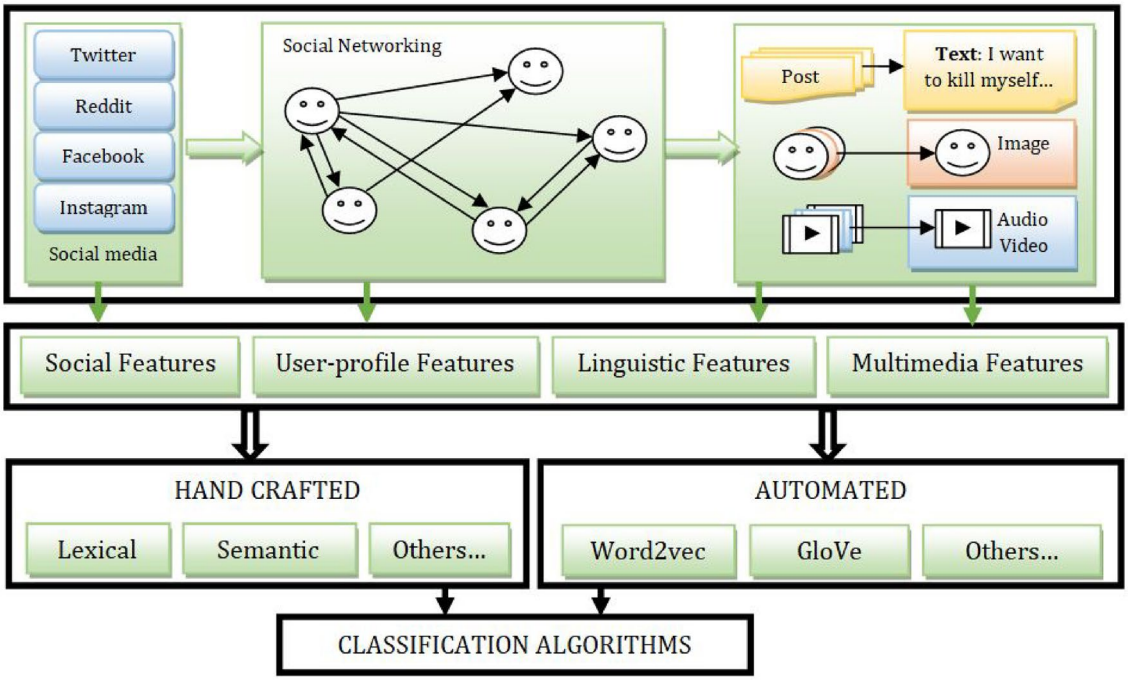


**Fig. 4** Architecture of feature harvesting from social media data for classification algorithms

understanding to infer mental states. In the past, an exclusive study for feature extraction have made headway towards finding neuropsychiatric disorders from self-reported text [13, 50].

The social media platforms are usually characterized by one-way connections (Twitter, Reddit, Instagram) and two-way connections (Facebook). The most widely used social media platforms are Twitter and Reddit followed by Instagram and Facebook. We observe multimodal models on social media data but we limit our studies to natural language processing and social features only.

When the information is limited, it start fabricating patterns among them. These patterns aid in feature extraction or transformation for both cross-sectional and longitudinal study. Recent works for feature extraction have addressed the concerns to explore dominant features [13, 50, 51].

The architecture of cross-sectional study to infer mental state from social media data is given in Fig. 4. Learning-based models are build on features extracted from data such as the *handcrafted features*, the statistical information, and *automated features* to name a few. We categories and discuss four classes of features, namely, user-profile features, linguistic features, social features, and multimedia features as given in Fig. 5.

The Social NLP research community exploit social media data for two modalities: *text* and *images*. We extract the textual features using either *a conventional approach* or via *automation*. The conventional approach contains *surface-level linguistic features* and *semantic level aspects* and is referred as *handcrafted features*. The automatic features incorporate vector representation for end-to-end pre-trained models.

## 2.1 Handling Ambiguity of Features

Although there is no ideal classification of features, we classify them into four different categories with few exceptions belonging to multiple categories. We resolve the perplexities with following guidelines:

- The metadata of posts yields information about both *user metadata*: data about the users' profile and is thus, kept under *user profile features*; and *post metadata*: data about the post and is categorized under *Social features*.
- The ruminative response style is expression of repetitive thoughts and behavior [52]. People with depression tend to express their feelings or negative experiences repeatedly by repeating the sentences in their posts. Though the ruminative response style is the part of both *user behaviour* and *linguistic styles*, it is more closely associated with user-profile features and thus, studied under *User Profile Feature*.

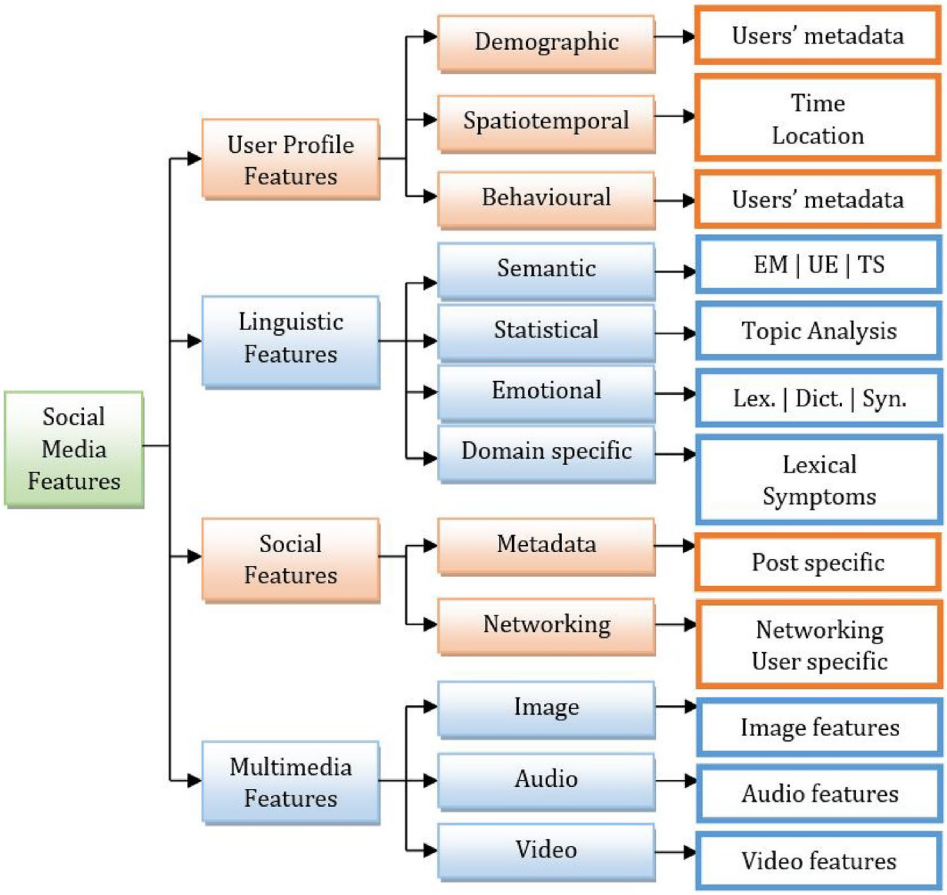


**Fig. 5** Classification of social media features for quantifying suicidal tendencies

**Table 1** User profile feature extraction for mental health state

| Category | Sub-category | Feature | Type | Research articles |
|---|---|---|---|---|
| Demographic | Users' Meta-data | Age | User | [36, 39, 53, 60] |
| | | Gender | User | [36, 36, 39, 53, 60] |
| | | Education | User | [39, 53] |
| | | Occupation | User | [39, 53] |
| | Users' Network | Follower | User | [61] |
| | | Ego-network | User | [62] |
| Spatio-temporal | Temporal | Timeline | User | [7, 36, 39] |
| | Spatial | Location | User | – |
| Behavioural | Posting Behavior | Gen. Behaviour | User | [39] |
| | | Ruminative | User | [11] |
| | | Posting Time | User | [59, 63] |

- An interesting study introduce bBridge [53], a big data based feature extraction approach from social media data which contains both user-profile features and social networking features.
- The community specific information of the user comprises of the information about followers, and favourites. We associated these features with the user's social networking and thus, are discussed in *Social Features*.

#### 2.1.1 User Profile Features

Past studies reveals the proportional impact of employment on psychiatric behaviour of a person by analyzing their college degree/type of job [54]. People sharing similar demographic, linguistic and cultural traits as those of depressed users are more at-risk than others [55]. In this context, we further classify the user-profile features in Table 1.

**Table 2** Linguistic feature extraction for mental health state

| Category | Sub-category | Feature | Type | Research articles |
|---|---|---|---|---|
| Emotional | Emotional Model | EmoBERT | Model | [7, 67] |
| | | MentalBERT | Model | [67, 68] |
| | Textual Sentiments | Emoji | Post | [39, 63, 76, 77] |
| | | Emoticons | Post | [63, 76, 77] |
| | | SentiWordNet | Post | [11] |
| | | SentiNet [78] | Post | [3] |
| Semantic | Topic Analysis | LDA | Post | [36, 39, 40, 59, 79] |
| | | Brown Clustering | Post | [70] |
| Statistical | Lexical | TFIDF | Post | [1, 7, 80, 81] |
| | | Text | Post | [38, 42, 63, 81–83] |
| | | Morphological | Post | [50, 84] |
| | | Stylometric | Post | [50] |
| | | n-gram | Post | [13, 40, 50, 59, 85] |
| | | Punctuation | Post | [76, 77] |
| | Dictionary | LIWC | Post | [13, 40, 45, 59, 81] |
| | | Suicide Dictionary | Post | [86] |
| | | ANEW | Post | [87] |
| | Syntactical | POS Tagging | Post | [11, 30, 50, 63, 81] |
| | Syntactical | POS Tagging | Post | [11, 30, 50, 63, 81] |
| Domain Specific | Lexicon | Antidepressant | Wiki | [3, 11, 39] |
| | | TensiStrength | Dict. | [59] |
| | | Dictionaries | Wiki | [3] |
| | Dep. Symptoms | DSM [88] | Dict. | [11, 39] |
| | | Plutchik | Post | [7] |
| | | VAD | Post | [36, 39, 89] |
| | | Affect and Intensity | Post | [36] |
| | | Big 5 Personality [90] | Post | [36] |
| | | Anxiety, Anger, Dep. | Post | [36] |

*Demographic* The *users' metadata* contains information about the age, gender, occupation, race, ethnicity [55]. These characteristics of people disclose their alignment towards psychiatric disorders such as mental disorders fore in old aged people more than younger ones. Social well-being of males decline more than females [56].

*Spatio-temporal* [7]: models user's emotional spectrum by tracking their historical timeline on social media platform. In their study, the patterns of irregularities among posting behaviour incorporates the time-varying component and use time-aware LSTM cell to capture patterns [57]. A shared task in eRISK workshop at CLEF forum introduce a longitudinal dataset which encourages more research contributions for early risk detection in social media [58]. Similarly, location of social media post have strong associations with economical indexes like *Ease of doing business*[4] and *World Happiness Report*[5] with mental health status of residents. In future, both temporal and location component may simulate significant information for mental health analysis.

*Behavioural features* Social media users are more likely to be expressed late night than during day time [39]. Behavioural patterns such as insomnia index, sleep cycle [45] and ruminative response style [52] affects the user's state of mind. People with depression tend to express their feelings or negative experiences repeatedly. In this context, [59] consider ruminative response style using text encoding mechanism resulting into significance of mental health analysis.

### 2.1.2 Linguistic Features

To study linguistic features [54], recapitulate the importance of words that users pick to express their feelings in their personal writings. People with depression exhibit differences with respect to linguistic styles such as the distribution of nouns, verbs and adverbs and the unconscious conceptualization of complex sentences [64]. The exclusive studies on linguistic features reveals the increased use of first person language, the current scenario and anger based terms for

[4] https://en.wikipedia.org/wiki/Ease_of_doing_business_index.

[5] https://en.wikipedia.org/wiki/World_Happiness_Report.

**Table 3** Social feature extraction for mental health state

| Category | Sub-category | Feature | Type | Research articles |
|---|---|---|---|---|
| Social Metadata | Post Specific | Length | Post | [7, 30, 43, 94] |
| | | #(Hashtags) | Post | [59] |
| | | #(URL) | Post | [59] |
| | | Metadata | Reddit | [36] |
| Social Network | Networking | Interactions | User | [37, 37, 39, 63, 77] |
| | | At-Mentions | User | [61] |
| | | Replied to | User | [61] |
| | User Specific | #(Favourites) | User | [63, 76] |
| | | #(Likes) | User | [77] |
| | | #(Posts) | User | [39, 63] |
| | | #(Comments) | User | [76, 95] |
| | | #(ReTweet) | User | [43, 76, 80] |

person's state of mind [65]. We further classify linguistic features in Table 2.

*Emotional Features* Infusing implicit and explicit emotions while encoding text is trending in current scenario. We emphasise and recommend the use of sentiments and emotions from active vocabulary of a user. The research community witnesses many emotion based pre-trained models as word embedding. Such models set strong foundation for building contextual transformer-based models [66]. We come across different pre-trained models such as EmoBERT [67], DistillBERT for emotions,[6] MentalBERT [68], and other Contextual BERT-based models [69].

*Semantic Features* The topic modelling methods such *LDA*, [28, 70] is used for clustering the posts related to similar topics. The depressed and non-depressed users discuss different topics which may help to determine potential depressed users [71]. Another interesting study aims to understand the Twitter users' discourse and psychological reactions to COVID-19 pandemic time period using topic modeling [72].

*Statistical Features* We categorize the statistical features into lexical, dictionary-based, and syntactic. The *lexical features* use tokenized form of text to calculate statistical measures such as TFIDF, n-grams, morphology and alike features. *Dictionary features* are use existing dictionaries such as LIWC,[7] Suicide dictionary[8] and ANEW[9] for assigning values. We use *syntactical features* are used to check the context of a token with respect to its neighbourhood, for instance, Part-Of-Speech tagging. The domain specific features are the lexicon of mental health specific words derived from Wikipedia, domain specific dictionaries, and depression symptoms such as Diagnostic and Statistical Manual of Mental Disorders (DSM-IV).[10]

*Domain Specific* With evolving era of 'Emotional Intelligence', we observe a clear description on emotion models in clinical psychology and psychiatric theories for affective computing [73]. Valence refers to the pleasant–unpleasant quality of a stimulus and ranges from negative to positive, whereas arousal refers to the intensity of a stimulus and ranges from dull to arousing. The past studies with MHA incorporate the Valence arousal dominance (VAD) Emotion model [36, 39, 43, 74] and Plutchik model [7, 75]. Plutchik's theory of emotion and emotional consequences for cognition, personality, and psychotherapy is derived from an evolutionary perspective [75].

### 2.1.3 Social Features

Depressed people who are conscious about their social circle on social media platforms and have limited number of friends [91]. The depressed tweet gains more attention from friends and so, important features are Retweets, comments, and favourites [76]. We further classify social features into *social metadata* and *social networking* as shown in Table 3.

**Table 4** Multimedia feature extraction for mental health state

| Category | Feature | Type | Research articles |
|---|---|---|---|
| Image | Colour Combinations | Image | [37–39, 76] |
| | Colour Ratio | Image | [37–39, 63, 76] |
| | Brightness | Image | [37–39, 63, 76, 95] |
| | Saturation | Image | [37–39, 63, 76, 77] |
| | Convolution | Image | [38, 96, 97] |

[6] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base.

[7] https://liwc.wpengine.com.

[8] https://sites.google.com/view/daeun-lee/dataset.

[9] https://github.com/sbma44/begin_anew.

[10] https://en.wikipedia.org/wiki/DSM-IV_codes.

**Table 5** Feature vector representation for social mental status detection

| Features | Category | Subcategory | Type | Papers |
|---|---|---|---|---|
| Traditional FE | Statistical | Vectorizer | Post | [13, 42, 50, 106] |
| | | Entropy | Post | [45] |
| | | Statistics | Post | [45] |
| | Dictionaries | Dict. learning | Post | [39, 63] |
| Embedding | Static | Word2vec [98] | Post | [1, 10, 11, 27, 50, 82, 97] |
| | | GLoVe | Post | [10, 11] |
| | | Fasttext | Post | [10, 11] |
| | CE: Transformer | BERT[a] [99] | Post | [7, 36, 37] |
| | | Sentence BERT [100] | Post | [36, 42] |
| | | GUSE [101] | Post | [42] |
| | Encoding Seq. | RNN [107] | Post | [11, 11] |
| | | GRU [108] | Post | [38, 43] |
| | | LSTM [109] | Post | [7, 63] |
| Image FE | Image | VGGNet [102] | Image | [38, 97] |
| | | ImageNet [110] | Image | [97] |
| | | CNN | Image | [37] |
| Dim. Red. | Linear | Filter | Vectors | [42, 45, 103] |
| | Non-linear | NMF | Vectors | [36] |
| | | t-SNE | Vectors | [1, 40] |
| | Post Feature T. | HAN [32] | Vectors | [11, 43, 44, 105, 111] |
| | | Joint Sparse Repr. | Vectors | [7, 39, 63, 76, 77] |
| | | Optimization | Vectors | [13, 82] |

*HAN* Hierarchical Attention Network

[a]https://github.com/google-research/bert

*Social Metadata* Social information about post of a user consists of the length of a post, number of hashtags in a post, number of URLs used in a post and other minute details which is termed as the metadata.

*Social Network* We observe patterns in interaction and relationships among users [92]. These networking features are gaining importance due to non-Euclidean space representation of the problem. Applying hyperbolic geometry on non-Euclidean representation has given new research direction in the field of mental health analysis [66, 93].

#### 2.1.4 Multimedia Features

The increase in use of *images* for feature extraction or transformation either consider *display picture* in Twitter (also referred as Avtars in Reddit) or images posted by user. The colour combinations, colour ratio, brightness, saturation, and convolution are few interesting features for mining social media images as shown in Table 4.

### 2.2 Feature Vector Representation

The feature vectorization is the process of representing input data in the form of a vector. We further classify feature vector representation into *text feature vectorization* and *image feature vectorization* as shown in Fig. 6. The *text feature vectorization* comprised of feature extraction and feature embedding. We enlist the past studies along with classified insights for feature vector representation in Table 5.

*Textual Feature Extraction* The traditional methods of converting text in vectors (TFEx) is performed with conventional approach of TFIDF vectorizer, Count vectorizer, and Hashing vectorizer [42]. For dimensionality reduction, the selective features are processed further by using PCA, NMF and other filter based linear feature selection algorithms. In the past, authors use one-hot encoding to encode a set of Tweets [82]. The uni-modal dictionaries evolves from text and image data separately which are further useful for joint sparse representation [39]. These traditional feature extraction techniques are convenient for converting the social media data into vector representation for classification models.

*Feature Embedding* With advancements in the word to vector conversion using neural network approach, the word-2vec [98], the GloVe [10, 11], and the Fasttext are encode the text. To handle the longer text like phrase, sentence or paragraph, the researchers use BERT [99], Sentence-BERT [100], and Google Universal Sentence Encoder (GUSE) [101] for feature vector representation [42].The use of embedding over dense layers, BERT, GUSE, and GRU [11,

**Table 6** Feature extraction and transformation for mental health detection

| Paper | Year | F1 | F2 | F3 | F4 | TFE | Emb. | DR | Output |
|---|---|---|---|---|---|---|---|---|---|
| Choudhury et al. [45] | 2013 | ✓ | ✓ | ✓ | | ✓ | ✓ | | Depression |
| Lin et al. [76] | 2014 | | ✓ | ✓ | ✓ | | | ✓ | Stress |
| Lin et al. [77] | 2017 | | ✓ | ✓ | ✓ | | ✓ | ✓ | Stress |
| Shen et al. [39] | 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | | | Depression |
| Song et al. [11] | 2018 | | ✓ | | | ✓ | ✓ | ✓ | Depression |
| Sawhney et al. [81] | 2018 | | ✓ | | | | ✓ | | Suicidal Id. |
| Orabi et al. [82] | 2018 | | ✓ | | | | ✓ | ✓ | Depression |
| Tadesse et al. [40] | 2019 | | ✓ | | | ✓ | | ✓ | Depression |
| Matero et al. [36] | 2019 | ✓ | ✓ | ✓ | | | ✓ | ✓ | Suicidal Id. |
| Gui et al. [38] | 2019 | | ✓ | | ✓ | | ✓ | ✓ | Depression |
| Guntuku et al. [59] | 2019 | ✓ | ✓ | ✓ | | ✓ | ✓ | | Stress |
| Xu et al. [112] | 2020 | ✓ | ✓ | ✓ | ✓ | ✓ | | | Mental Health |
| Lin et al. [37] | 2020 | | ✓ | ✓ | ✓ | | ✓ | | Depression |
| Sawhney et al. [7] | 2021 | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | Suicidal Id. |
| Haque et al. [42] | 2021 | | ✓ | | | ✓ | ✓ | ✓ | Suicide and Dep. |
| Zogan et al. [43] | 2021 | ✓ | ✓ | | | | ✓ | ✓ | Depression |
| Turcan et al. [83] | 2021 | | ✓ | | | | ✓ | ✓ | Stress |
| Zogan et al. [41] | 2021 | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | Depression |
| Lee et al. [86] | 2022 | | ✓ | | | | ✓ | ✓ | Suicide Risk |
| Tavchioski et al. [113] | 2022 | | ✓ | ✓ | | | ✓ | ✓ | Depression |
| Naseem et al. [114] | 2022 | | ✓ | | | | ✓ | ✓ | Depression |
| Garg et al. [114] | 2022 | | ✓ | | | | ✓ | | Depression |
| Yang et al. [85] | 2022 | | ✓ | | | | ✓ | | Depression |

*F1* User Profile Feature, *F2* Linguistic Feature, *F3* Social Feature, *F4* Multimedia Feature, *FEx* Feature Extraction, *DR* Dimensionality Reduction, *FEm* Feature Embedding, *PFT* Post Feature Transformation

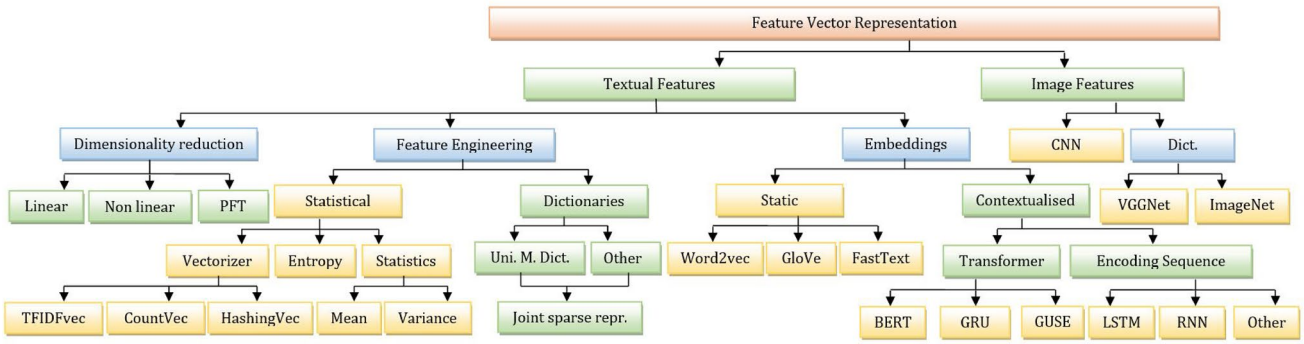


**Fig. 6** Feature vector representation for mental health analysis in social media posts

38] for sequence to sequence learning has given significant contributions in attention based mechanism to enhance the importance of feature across representation.

An image represents many characteristics of the psychological thoughts and health. The permutations and combinations of different image features extraction determines the mental health. The research community follows end-to-end feature transformation technique by using a 16-layer pre-trained VGGNet to use image as features [38, 97, 102].

*Dimensionality Reduction* One of the most promising step of social media data mining is dimensionality reduction. The dimensions of text representation in conventional feature extraction techniques are reduced by linear and non-linear methods such as Principal Component Analysis (PCA), Deep Neural Autoencoders (DNAE) [103], and Uniform Manifold Approximation and Projection (UMAP) [104] for dimensionality reduction in MIDAS. The Post Feature Transformation (PFT) approach is recommended for transformer-based end-to-end data conversion into feature

**Table 7** Results obtained for Social Media Health Detection

| Year | Dataset | Psychological Outcome | Papers | Avail. |
|---|---|---|---|---|
| 2015 | CLPsych [25] | Suicide Risk | [25–28, 36, 82, 82, 117] | S |
| 2017 | MDDL [39] | Depression | [37–39, 41] | ✓ |
| 2017 | RSDD [115] | Depression | [11, 41, 44, 115] | S |
| 2018 | SMHD [118] | Mental Health | [118, 119] | S |
| 2018 | eRISK[116] | Stress | [3, 39, 116] | ✓ |
| 2018 | *Pirina* [120] | Depression | [40, 120] | ✓ |
| 2018 | *Ji* [121] | Suicidal id. | [1, 13] | AOR |
| 2019 | Sina Weibo [10] | Suicide Risk | [2, 10] | AOR |
| 2019 | Dreaddit [35] | Stress | [83, 122] | ✓ |
| 2019 | SRAR [123] | Suicide Risk | [123] | S |
| 2019 | *Aladaug* [124] | Suicidal Id. | [1, 124] | AOR |
| 2020 | UMD-RD [22] | Suicide Risk | [22, 36, 123] | S |
| 2020 | GoEmotion [33] | Emotion | [83, 125] | ✓ |
| 2021 | SDCNL [42] | Suicide/Depression | [42] | ✓ |
| 2022 | CAMS [126] | Mental Health | [126, 127] | ✓ |
| 2022 | RHMD [128] | Mental Health | [128] | ✓ |
| 2022 | *Kayalvizhi* [129] | Depression | [113, 130] | ✓ |

✓: Available, *S* Available via Signed agreement, *AOR* Available On Request to authors

vectors. We witness existing works with attention mechanism such as Hierarchical Attention Mechanism (HAM) [32, 105] to give importance to important posts for identifying suicidal tendencies [11, 43]. The multi-attributed feature extraction is given as 3-level framework using three-level features extraction which consists of low level feature (linguistic features), middle level features (visual features) and high-level features (social features) to give as an input to the Deep Sparse Neural Network (DSNN) [76]. They argue the unavailability of all three types of features in data.

### 2.3 Summary of Feature Extraction and Transformation

The existing potential studies define and explore new features for mental health detection from social media data as shown in Table 6. Most of the recent approaches use embedding techniques and work on post-feature transformation to hypothesise better feature representation. Moreover, all existing studies are using the textual information of post and other features optionally.

## 3 Classification

The classification problem of identifying suicidal tendency on social media use many shallow learning and DL algorithms. One of the most challenging module is to handle the unstructured and semi-structured data from social media data, filling missing values and jointly represent the multimodal information. Although, data resource for this task is freely available in public domain, most of the dataset are not available due to sensitivity of the data.

### 3.1 Available Dataset

In the past, the research community witness the use of widely available datasets such as CLPsych shared task [25], Reddit Self-reported Depression Diagnosis [115], and Language of Mental Health [64], early risk prediction on the Internet (eRISK) from CLEF Forum [116]. As discussed earlier, only a few dataset are available in public domain, many of them are either reproducible or available on request. Every year we come across more than 12 dataset for predicting mental health on social media data. Limited availability of these dataset lead us to enlist either the most popular and reproducible dataset, or the dataset which are available by request or via signed agreement. A list of reproducible dataset are enumerated in Table 7. In this section, we further discuss details of each dataset.

(1) *CLPsych 2015 Shared task dataset:* The CLPsych dataset[11] contains three modules which are available via signed agreement, namely, DepressionvControl (DvC), PTSDvControl (PvC), and DepressionvPTSD (DvP). To use this dataset, the academic researchers must sign a confidentiality agreement to ensure the privacy of the data.

[11] https://github.com/clpsych/shared_task.

(2) *Multimodal Dictionary Learning (MDDL):* MDDL[12] is a depression detection dataset which comprises of three modules D1, D2, and D3. The *Depression Dataset D1* is constructed using tweets from 2009 and 2016 where users were labeled as depressed if their anchor tweets satisfied the strict pattern "(I'm/I was/I am/I've been) diagnosed depression". The *Non-Depression Dataset D2* is constructed in December 2016, where users were labeled as non-depressed if they had never posted any tweet containing the character string "depress". Although D1 and D2 are well-labeled, the depressed users on D1 are too few, thus, a larger unlabelled *Depression-candidate Dataset D3* is constructed for depression behaviors discovery which contains much more noise.

(3) *Reddit Self-reported Depression Diagnosis (RSDD)*: The RSDD dataset[13] contains the Reddit posts of approximately 9000 users who have claimed to have been diagnosed with depression ("diagnosed users") and approximately 107,000 matched control users. The introduction to Reddit dataset [115] has given a significant contribution which was used by many existing studies.

(4) *Self-Reported Mental Health Diagnoses (SMHD) dataset*: The SMHD dataset,[14] just like RSDD dataset, can be obtained via signed agreement as per the privacy policy of data. The dataset consists of Reddit posts of the users diagnosed with one or several of nine mental health conditions ("diagnosed users"), and matched control users. This dataset is also used by few studies in literature and is related to multiple mental health conditions instead of just the depression dataset.

(5) *eRISK*: The eRISK dataset[15] is available online for experiments and analysis to meet the targets of a shared task since few years. The dataset for early risk detection by CLEF Lab is given to solve the problems of detecting depression, anorexia and self-harm since few years.

(6) *Pirina*: A new dataset is proposed [120], named as Pirina to refer it in this study and is available online[16] for research purposes. A filtered data is extracted from Reddit social media platform for depression detection task. Although, this dataset is not actively maintained, it can be extracted and can be used for pilot study.

(7) *Ji*: A new Reddit dataset of 5326 suicidal posts out of 20,000 posts were extracted and 594 Suicidal Tweets out of 10,000 Tweets were extracted for experiments and evaluation of the proposed classification approach for suicidal risk detection. This dataset is referred as *Ji* dataset[17] in this study which is available on-request.

(8) *Sina Weibo*: Another dataset which is proposed for public domain and remains un-named is given the name of the social media platform, Sina Weibo,[18] to refer it for this study. The dataset with 3652 users having suicidal tendency and 3677 users not having suicidal risk is extracted from Sina Weibo, a Chinese social media platform.

(9) *Dreaddit*: Dreaddit,[19] a new text corpus of lengthy multi-domain social media data for the identification of stress. This dataset consists of 190K posts from five different categories of Reddit communities; the authors additionally label 3.5K total segments taken from 3K posts using Amazon Mechanical Turk. The lexical features which used in this dataset are Dictionary of Affect in Language [131], LIWC features [132] and patterns sentiment library [133]; syntactic features like unigrams and bigrams, the Flesch-Kincaid Grade level and the automated reliability index; social media features like timestamp, upvote ratio, karma (upvote–downvote) and the total number of comments.

(10) *Suicide Risk Assessment using Reddit (SRAR)*: The SRAR dataset[20] is available in public domain. The dataset is composed of 500 Redditors (anonymized), their posts and domain expert annotated labels. The SRAR is used along with different lexicons which are built from the knowledge base associated with mental health like SNOMED-CT, ICD-10, UMLS, and Clinical Trials. This dataset is recently used [123] and the research community is looking forward to use this in near future to enhance the proposed techniques.

(11) *Aladaug*: This dataset is built by Aladaug [124] during his study on suicidal tendency identification from the posts over social media data. Since, there is no name given to this dataset, this dataset is named as *Aladaug* to refer it in this study. Among 10,785 posts, 785 were manually labelled for this study. This dataset is available on request from authors.

(12) *The University of Maryland Reddit Suicidality Dataset (UMD-RD)*: The UMD-Reddit Dataset[21] contains one sub-directory with data pertaining to 11,129 users

[12] https://github.com/sunlightsgy/MDDL.

[13] http://ir.cs.georgetown.edu/resources/rsdd.html.

[14] http://ir.cs.georgetown.edu/resources/smhd.html.

[15] https://erisk.irlab.org/eRisk2021.html.

[16] https://files.pushshift.io/reddit/submissions/.

[17] https://github.com/shaoxiongji/sw-detection.

[18] https://github.com/bryant03/Sina-Weibo-Dataset.

[19] http://www.cs.columbia.edu/~eturcan/data/dreaddit.zip.

[20] https://github.com/AmanuelF/Suicide-Risk-Assessment-using-Reddit.

[21] http://users.umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html.

who posted on *SuicideWatch*, and another for 11,129 users who did not. For each user there is full longitudinal data from the 2015 Full Reddit Submission Corpus. The UMD-Reddit dataset have been used by academic researchers actively since 2019 as it is available via signed agreement.

(13) *GoEmotion*: The GoEmotion dataset[22] contains 58K carefully curated comments extracted from Reddit, with human annotations to 27 emotion categories or Neutral. It also contains a filtered version based on reter-agreement, which contains a train/test/validation split. This dataset is proposed [33] in 2020 for emotion detection and is used to validate the scalability of the proposed models for stress detection.

(14) *SDCNL dataset*: The SDCNL[23] dataset was collected using Reddit API and scraped from two subreddits, r/SuicideWatch and r/Depression which contains 1895 total posts. Two fields were utilized from the scraped data: the original text of the post as our inputs, and the subreddit it belongs to as labels. Posts from r/SuicideWatch are labeled as suicidal, and posts from r/Depression are labeled as depressed.

(15) *CAMS*: CAMS stand for Causal Analysis for Mental illness in Social media posts. The introduction of CAMS dataset[24] enables academic researchers to perform causal inference, causal explanation extraction and causal categorization. The dataset contains 5051 samples and categorize each sample into one of the five different causal categories, namely, bias/abuse, jobs and carers, medication, relationships, and alienation. This dataset is publicly available [126].

(16) *RHMD*: The RHMD stands for a Real-world Dataset for Health Mention classification on Reddit data.[25] The health mention is defined as a problem to find symptoms and understand its semantics. These semantics specifies the contextual perspective in which a given symptom is used in texts [128]. Every sample of this dataset categorizes a given post in five categories health mention, non-health mention, hyperbolic mention, figurative mention, and uninformative.

(17) *Kayalvizhi*: A unique dataset[26] that not only detects depression from social media but also analyzes the level of depression. Initially 20,088 instances of postings data were annotated, out of which 16,613 instances were found to be mutually annotated instances by the two judges, and thus they were considered as instances of data set with their corresponding labels [129].

### 3.2 The Historical Evolution of Classification Models

In this section, we discuss the evolution of methods developed for mental health analysis in the past. The Social NLP researchers at Microsoft, one of the leading IT based solution organization, disclose the significance with role of social media in identifying mental health problems. After comprehensive study of 92 research articles on three mental health problems of stress, depression and suicide risk; the evolution of historical timeline is represented in Fig. 7. Furthermore, the architecture of path-breaking models for mental health analysis is shown in Fig. 8.

Past studies since 2013 set preamble to investigate the significance of users' social media data for predicting depression [45] and suicidal tendencies [60]. With introduction to word-embedding and vector-space representation [98], encouraging studies over developing deep neural network classifies for psychological perspective has gained much attention from academic researchers [76, 95]. After linguistic features, Ref. [27] introduce unique features, namely, user-profile features resulting into improved performance for classifying posts. We witness exponential growth in this domain after release of initial datasets as it resolve the problem with limited availability of sensitive dataset of mental health in social media posts. CLPsych shared task data paves a way for new studies and development of new datasets for future use [47].

In 2017, we observe extended studies on different social media platforms such as Facebook [16], Sina Weibo (a Chinese online social platform), and Instagram [94, 134]. The use of social media and social network features for stress detection has enriched this domain with learning-based mechanisms [77]. Simultaneously, the dual-attention mechanism for multimodal approaches reveals the need of explainability and reliability of models [135].

In 2018, more studies revolve around the dimensionality reduction or optimizing the feature vector for ML and DL models, respectively [82]. The studies for depression detection started with the use of different social network features [45], evolved with interactions over social media [77] and cascading social networks [61] to extract reliable features, followed by ontology and knowledge graphs [2].

The observations about users' dynamic historical timeline on Twitter include improvements with interpretive Multi-Modal Depression Detection with Hierarchical Attention Network (MDHAN) [43]. The MDHAN framework is designed with multi-model features and two attention

[22] https://github.com/google-research/google-research/tree/master/goemotions.

[23] https://github.com/ayaanzhaque/SDCNL/tree/main/data.

[24] https://github.com/drmuskangarg/CAMS.

[25] https://github.com/usmaann/RHMD-Health-Mention-Dataset.

[26] https://github.com/Kayal-Sampath/detecting-signs-of-depression-from-social-media-postings.

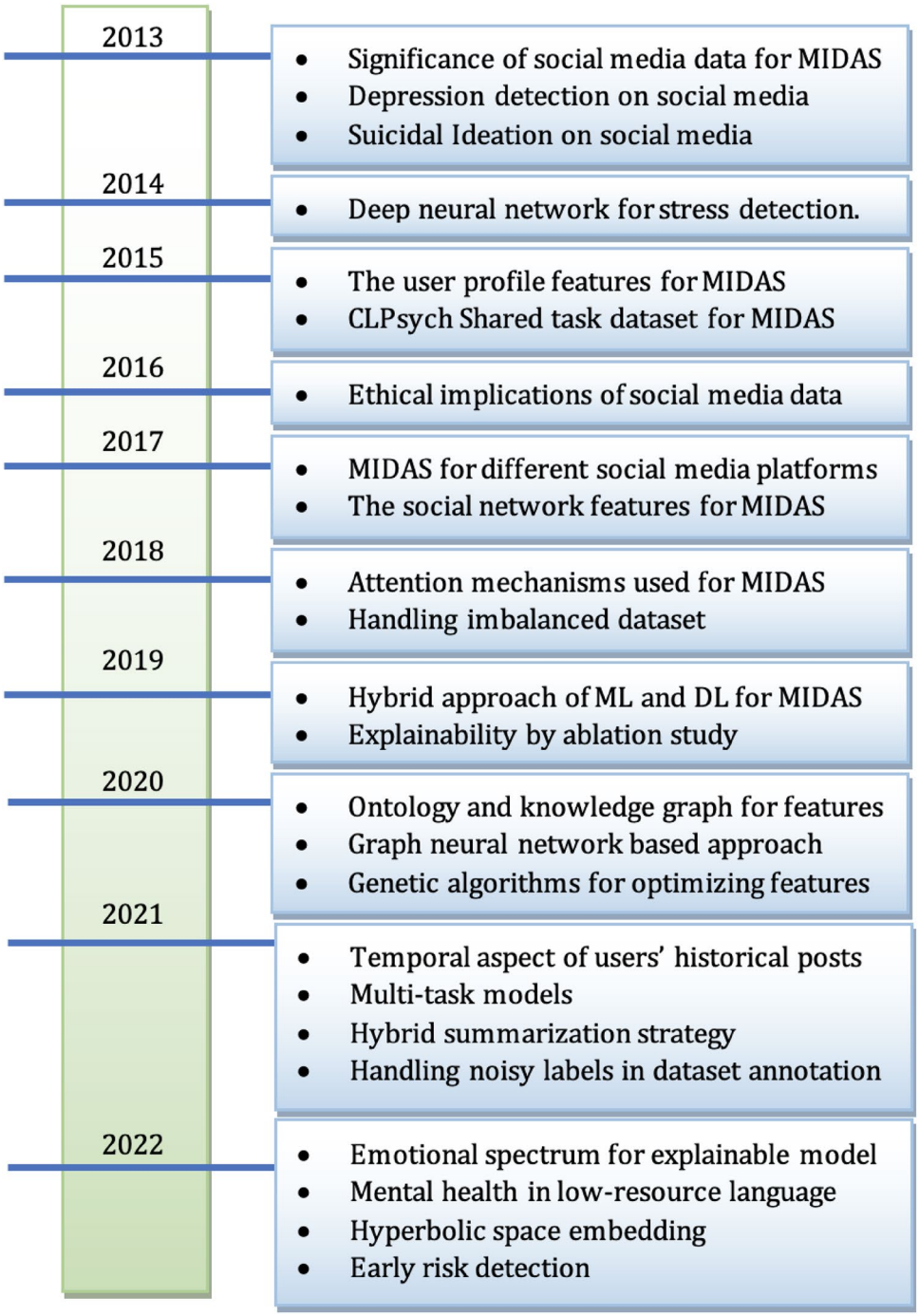


**Fig. 7** The timeline of evolving important events for quantification of suicidal tendency on social media

mechanisms are applied at tweet-level and at word-level, respectively. Ref. [38] introduce COMMA, a depression detection mechanism, to use encoded text/ visual data and their selection using GRU to apply averaged embedding on classifier. We further investigate a set of recent contextualized models such as multimodal feature extraction techniques for multiple social networking learning (MSNL) [136], Wasserstein dictionary learning (WDL) [137], and multimodal depressive dictionary learning (MDL) [39] methods. The authors in Dual-ContextBERT model [36] use multi-level analysis by removing a limitation of single-level analysis. It is the best performing model at CLPsych 2019 which feeds BERT encoded posts to attention-based RNN layer.

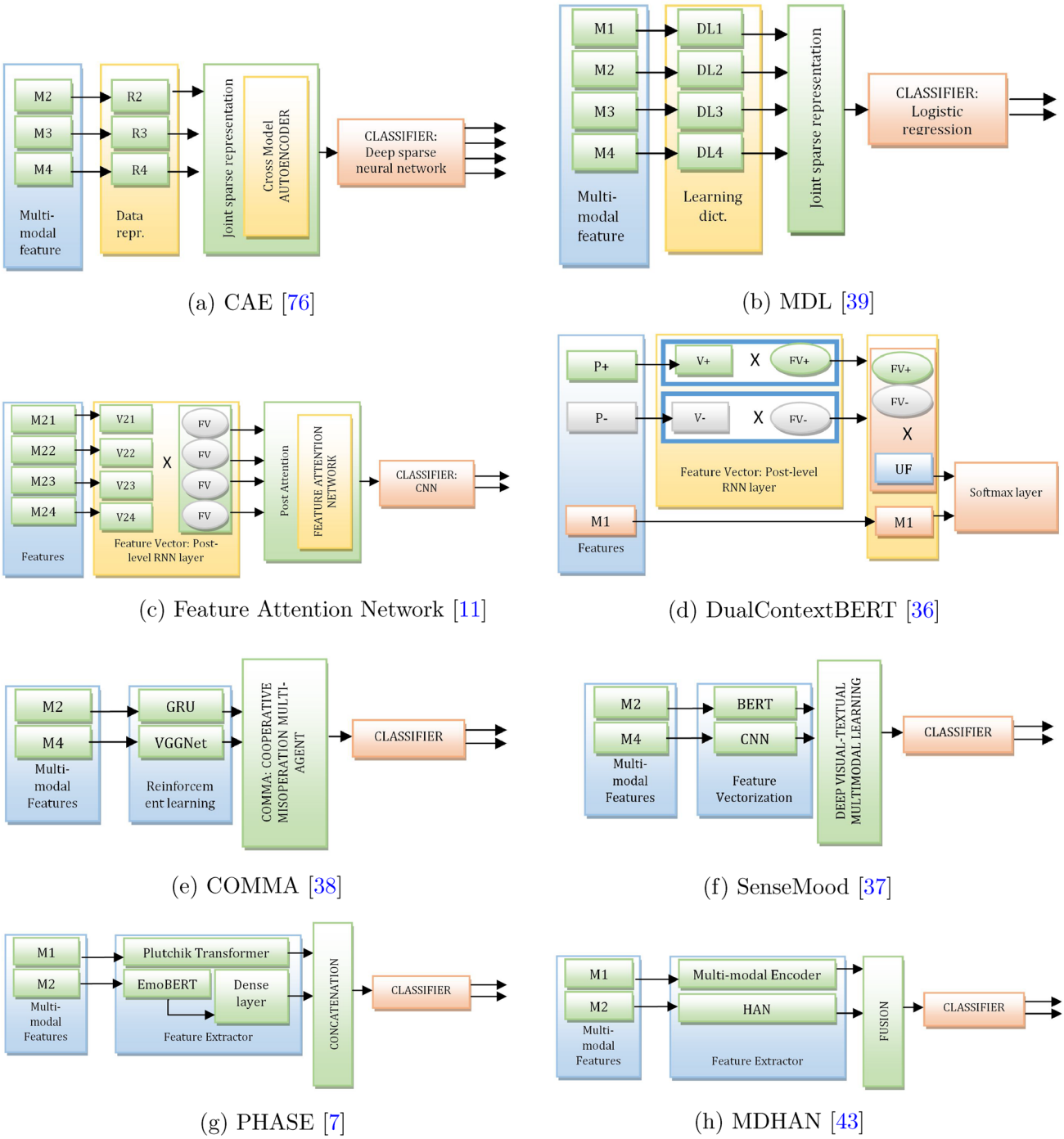


**Fig. 8** Some existing models for quantifying the suicidal tendency on social media

The performance evaluation for responsible and explainable models is carried out for mental health prediction using *Ablation study* [7, 83]. After extensive literature over ML and DL algorithms, academic researchers found interesting improvements with hybrid studies [36]. Recent investigation with Graph Neural Network results into improved early risk detection [2]. Furthermore, computation intelligence techniques for feature optimization has resolved the problem of noisy data [13].

During next transition phase, we observe research advancements with historical aspect of the users' timeline for identifying different phase of mental health [7], and hybrid extractive and abstractive summarization strategy as

**Table 8** Linguistic feature extraction for mental health status

| Outcome | Method | Dataset | Baselines | Results | C. |
|---|---|---|---|---|---|
| Depression | MDL [39] | MDDL | Naive Bayes [136, 137] | F1: 85% | |
| Suicidal Id. | PHASE [7] | Self [81] | [10, 36, 81, 123, 142, 143] | F1: 80.5% | ✓ |
| Depression | COMMA [38] | MDDL | Naive Bayes, [39, 136, 137] | F1: 90% | |
| Stress | CAE [76] | Self | (SVM, ANN, DNN) + SAE | F1: 86.12% | |
| Stress | FGM [77] | Self | LR, SVM, RF, DNN | F1: 93.40% | |
| Depression | FAN + CNN [11] | RSDD | RNN | Better P | |
| Suicidal Id. | C-LSTM [81] | Self [81] | LSTM, RNN | F1: 82.7% | |
| Suicidal Id. | SISMO [31] | SRAR | [4, 10, 36, 123, 142] | F1: 73% | |
| Suicidal Id. | D-C BERT [36] | Self | BERT, Dict. | F1: 50% | |
| Depression | WEO [82] | CLPsych | Word2vec | F1: 86.96% | |
| Dep./SI | GUSE [42] | SDCNL | BERT | F1: 95.44% | ✓ |
| Depression | ML [3] | eRISK | ML | F1: 53% | ✓ |
| Depression | MDHAN [43] | MDDL | HAN, CNN, [39], BiGRU | F1: 89.3% | |
| Stress | *Turcan* [83] | Dreaddit + 1 | RNN, BERT, Multi-task | F1: 80.34% | ✓ |
| Depression | SenseMood [37] | MDDL | [39, 63, 135, 140] | F1: 93.60% | |
| Depression | MLP [40] | Pirina | SVM, LR, RF, Ada-Boost | F1: 93% | |
| Depression | XA-Boost [44] | RSDD | SVM, LSTM, [115, 132] | F1: 60% | |
| Depression | DepressionNet [41] | MDDL | [39, 108], BiGRU, CNN | F1: 91.2% | ✓ |
| Suicidal Id. | SNAP-BATNET [61] | Self [81] | [81], ELMo, RCNN | F1: 92.6% | |
| Suicidal Id. | LSTM-CNN [1] | Self | RF, SVM, NB, XGBOOST | F1: 93.4% | ✓ |
| Suicidal Id. | SDM [10] | Sina-W | NB, LSTM, SVM | F1: 90.92% | |
| Suicidal Id. | KGbased [2] | Sina-W | SDM, LSTM, CNN | F1: 93.69% | |
| Suicidal Id. | DAM [9] | Self | NB, LSTM, SVM, CNN [10] | F1: 91.54% | |
| Depression | LSTM + RNN [144] | Self | LSTM, RNN | F1: 98% | |
| Suicide risk | C-GraphSAGE [86] | UMD | SDM, C-CNN, BERT, RF | F1: 84% | ✓ |
| Depression | KG + DR [113] | WikiData5m | autoBOT, BERT | F1: 86.27% | ✓ |
| Suicide risk | *Naseem* [114] | [123] | C-CNN, SDM, SISMO, RF | F1: 79% | |
| Suicide risk | *Garg* [126] | CAMS | LSTM, GRU, CNN | F1: 50.13% | ✓ |
| Depression | KC-Net [85] | Dreaddit | GRU, BERT, BiLSTM-Att | F1: 83.5% | ✓ |

DepressionNet [41].[27] DepressionNet is a novel approach which summarizes user posts before encoding it via embedding. They apply BiGRU model and concatenate results with encoded current post. The multitask models encode data using pre-trained models and GoEmotions dataset [83].[28] Most of the datasets collected from Reddit are labelled using sub-reddits. However, Ref. [42] suggests the problem of noisy labels and address it by introducing a new dataset on depression versus suicide. In extension to this, a data augmentation approach resolve the problem of limited data availability for mental health analysis [138].

Past studies incorporate the multi-modal feature extraction for building contextual transformer based models to resolve the problem of depression detection such Co-attention [139], Dual attention [135], and Modality attention [140]. The novel contributions for suicidal tendency predictions are comprised of implicit emotion-based features [7] and explicit commonsense knowledge [141].

People often express their feeling in native language and thus, a potential new research frontier is to build explainable language-independent models for low-resourced languages. A comprehensive study for Chinese data reveals interesting insights with semantics in language [145]. Linguistic features analysis shows significant increase over due to frequency of terms related to affect, positive emotion, anger, cognition (including the subcategory of insight), and conjunctions. A recent work with code-mixing is carried out over English and Hindi language, which shall help in implementation across multiple platforms and help in putting a stop to the ever-increasing depression rates in a methodical and automated manner [146]. We keep this as an open-research direction to examine mental health for low-resourced languages.

We further observe an improved efficiency of early risk detection with the help of bidirectional transformer based models and ordinal classification [114, 147]. Recent advances on early depression detection using attention mechanisms

[27] https://github.com/hzogan/DepressionNet.

[28] https://github.com/eturcan/emotion-infused.

**Table 9** Inferences of evolving suicidal tendency detection on social media

| Paper | Year | Contributions | Det. | BA | Str | Ex | LI | Code | L |
|---|---|---|---|---|---|---|---|---|---|
| Lin et al. [76] | 2014 | A cross-media auto-encoder for joint representation of features | ✓ | | | | | R | Chinese |
| Lin et al. [77] | 2017 | A factor graph model (FGM) with CNN for classification | ✓ | ✓ | | | ✓ | R | Chinese, English |
| Shen et al. [39] | 2017 | Public dataset, feature extraction with scalable approach for SMHP | ✓ | ✓ | | | | R | English |
| Almeida et al. [3] | 2017 | Checked different machine learning models with classification | ✓ | | | | | A | English |
| Song et al. [11] | 2018 | A feature attention network for identifying important features | ✓ | | | I | | R | English |
| Orabi et al. [82] | 2018 | The Word Embedding Optimization (WEO) for optimizing the feature vectors | ✓ | | | | | NA | English |
| Gui et al. [38] | 2019 | Introducing GRU + VGGNet + COMMA model for Depression detection | ✓ | ✓ | | | | R | English |
| Matero et al. [36] | 2019 | A dual context based approach by hybridising both ML and DL | ✓ | | | | | NA | English |
| Guntuku et al. [59] | 2019 | Implications of using social media as a tool for stress detection, studies over Facebook and Twitter | | ✓ | | | | NA | English |
| De Choudhury et al. [45] | 2013 | The use of statistics of social media data for SMHP | | ✓ | | | | NA | English |
| Tadesse et al. [40] | 2019 | Investigate machine learning techniques for depression detection | ✓ | | | | | R | English |
| Cong et al. [44] | 2018 | The model of integrating XGBoost and Attention with BiLSTM | ✓ | | | | | R | English |
| Vioules et al. [151] | 2018 | Automatic identification of user's online behaviour | ✓ | ✓ | ✓ | | | NA | English |
| Mishra et al. [61] | 2019 | The social networking features based model for identifying suicide ideation | ✓ | | | | | R | English |
| Cao et al. [10] | 2019 | A model with two-layered attention mechanism and domain specific word embedding | ✓ | ✓ | | ✓ | ✓ | R | Chinese |
| Xu et al. [112] | 2020 | Jointly analyzing language, visual, and metadata cues and their relation to mental health | ✓ | | | | | R | English |
| Lin et al. [37] | 2020 | A deep visual textual multimodal learning to map psychological state of users on social media | ✓ | ✓ | ✓ | | | R | English |

*Ex* Explainability, *A* Available, *R* Reproducible, *S* Available by Signed Agreement, *NA* Not Available, *Str* Streaming Data, *LI* Language Independent, *L* Language used, *Det*. Detection, *BA* Behavioural Analysis

over transformer-based model results into explainable AI in this domain [85, 147–149]. More work with graph convolution encoders [86] and hyperbolic space embedding has enriched this domain with new insights on recognizing patterns in graph and visualizing the problem in non-Euclidean distance, respectively. Other than improvising cross-sectional and longitudinal studies with additional attention mechanisms and semantic enhancements, we came across next level study on finding indicators to state reason behind mental disorders in self-reported texts [126]. Such studies show new research direction towards discourses and pragmatics.

To summarize the extensive study of classification models for identifying suicidal tendency, we reveal information about recent developments in Table 8 where we mention dataset, baselines, results and code availability for each study. We acknowledge that existing studies are not directly comparable. Also, before we discuss new frontiers, we enlist useful tools and resources for future research.

### 3.3 Tools and Resources

As discussed earlier, the social media data is firsthand user-generated information which is informal in nature. Thus, identifying named entities and semantics in social media posts is still a challenging task. In this section, we enlist different tools/ libraries as potential sources.

- *Python Reddit API* The Reddit social media platform can be scrapped through Python Reddit API Wrapper (PRAW)[29] and follows Reddit API rules[30] for scrapping data.

[29] https://github.com/praw-dev/praw, https://github.com/shaoxiongji/webspider.

[30] https://github.com/reddit-archive/reddit/wiki/API.

**Table 10** Inferences of recent suicidal tendency detection on social media

| Paper | Year | Contributions | Det. | BA | Str | Ex | LI | Code | L |
|---|---|---|---|---|---|---|---|---|---|
| Tadesse et al. [1] | 2020 | The LSTM + CNN classification model | ✓ | | | | | R | English |
| Cao et al. [2] | 2020 | A knowledge graph and ontology based graphical neural network for suicide risk detection | ✓ | ✓ | | ✓ | ✓ | R | Chinese, English |
| Shah et al. [13] | 2020 | Hybrid approach by using computationally intelligent techniques and other optimizations for features | ✓ | | | | | R | English |
| Sawhney et al. [7] | 2021 | Users' historical timeline encoded and mapped with other features | ✓ | | | ✓ | | A | English |
| Sawhney et al. [31] | 2021 | An ordinal attention network for suicidal ideation detection | ✓ | ✓ | | ✓ | | A | English |
| Zogan et al. [43] | 2021 | A multi-modal depression detection with HAN (MDHAN) | ✓ | | | ✓ | | R | English |
| Turcan et al. [83] | 2021 | Multi-task with emotional models for more explainable stress detection model | ✓ | ✓ | | ✓ | | A | English |
| Haque et al. [42] | 2021 | The SDCNL model with GUSE—dense over UMAP-(Kmeans, GMM) | ✓ | | | | | A | English |
| Zogan et al. [41] | 2021 | DepressionNet using hybrid extractive and abstractive summarization strategy | ✓ | ✓ | | | | A | English |
| Lee et al. [86] | 2022 | The great utility in identifying suicidality of individuals using suicide dictionary and graph neural network | ✓ | ✓ | | ✓ | ✓ | A | English |
| Tavchioski et al. [113] | 2022 | A novel method using knowledge graph and dimensionality reduction for depression detection | ✓ | ✓ | | | | R | English |
| Naseem et al. [114] | 2022 | A behaviour prediction model uses ordinal classification over transformer encoder | ✓ | ✓ | | ✓ | ✓ | NA | English |
| Garg et al. [126] | 2022 | Learning based approach for causal analysis of mental health illness in social media posts | ✓ | ✓ | | | | A | English |
| Yang et al. [85] | 2022 | A knowledge-aware module based on dot-product attention to accordingly attend to the most relevant knowledge aspects | ✓ | ✓ | | ✓ | ✓ | A | English |

*Ex* Explainability, *A* Available, *R* Reproducible, *S* Available by Signed Agreement, *NA* Not Available, *Str* Streaming Data, *LI* Language Independent, *L* Language used, *Det.*: Detection, *BA* Behavioural Analysis

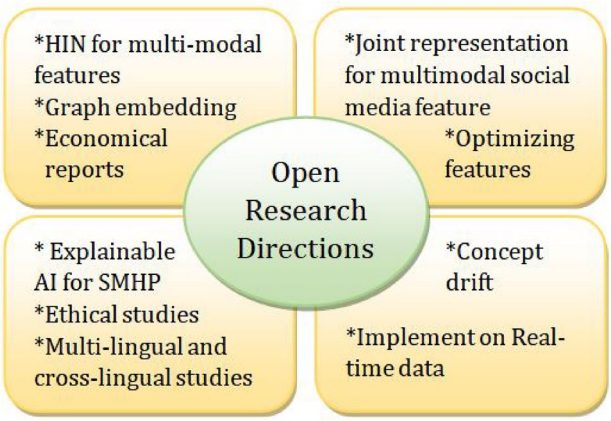


**Fig. 9** Open challenges and new research directions in identifying suicidal tendency on social media

- *PyPlutchik*: An embedding to employ emotion models as pre-built tools in Python environment [150] and trained on Plutchik model of emotions [75].
- *DLATK Python Package*: DLATK stands for Differential Language Analysis Toolkit which is an end-to-end human text analysis package which is specifically suited for social media and social scientific applications. The non-neural models may be implemented via the DLATK Python package [90].
- *Optimization*: Adaptive experimentation is the ML guided process of iteratively exploring a (possibly infinite) parameter space in order to identify optimal configurations in a resource-efficient manner. Ax[31] currently supports Bayesian optimization and bandit optimization as exploration strategies and is used for social mental health detection [83].

[31] https://github.com/facebook/Ax.

## 4 New Frontiers

After extensive study of 92 research articles related to stress, depression and suicidal tendency, we make inferences to define new research directions and future scope as shown in Tables 9 and 10. Finally, we give new frontiers in Fig. 9.

1. *Noisy Labels*: We found that the potential of some labels of data is found to be corrupted in the past which are mentioned as the *noisy labels*. To solve this problem, SDCNL model introduced a unique feature of label correction methodology to classifying posts as suicide versus depression [42].
2. *New Features*: The other factors which can be potential features are the happiness index of the country of a user; the ease of living index of the country of the user; the variation in geographical locations and multi-source distributed crawling; detection of multi sources communities by using spectral clustering over multi-level graphs [53]. Although there are studies on finding correlations among different features and map different variables for mental illness in China [152], there is the need to study this for different countries and at global-level due to much of socio-political differences in each country.
3. *Embedding for Multi-task problem*: We observe solution of multi-task mental health analysis through systematic word embedding optimizer [82]. However, there is no explainability or mathematical validation for why the results are better.
4. *Time Complexity*: Although, it is observed that the recent approaches for stress detection shows the significant improvement with F1-score for FGM approach [95] but it is computationally expensive and takes almost more than the double time as compared to the second best approach. There is need to give equal importance to the complexity in recent advancements.
5. *Behavioral Analysis*: The mental health detection is the part of integrated study of computational linguistics, human–computer interactions and clinical psychology. Few studies have observed the latent patterns among social media users which express their common but sensitive thoughts. Depressed tweets are more likely to be expressed late night than during day time [39]. This analytical part of human behavior is rarely explored in the existing literature as observed from Tables 9 and 10.
6. *Interpretability and Explainability*: There are detailed and theoretical explanations of the proposed approach to test its interpretability [11] or explainability [2, 7, 43, 83] via ablation studies. A complete section of ethical validation must be explored further to enhance the applicability of the new methods in real-time applications.
7. *Social Networks and Graph Neural Networks*: The trend of making use of text, visual, and multimedia information has given several new research directions in this domain. In the past, network features for Twitter data shows promising results [153], still there is a big room to study multi-level networks and heterogeneous information networks for multi-modal information in social media for better and integrated representation. Few studies on knowledge graph [2], ontology [2] and graph neural networks [92] validate it as a progressive domain.
8. *Multi-lingual, cross-lingual and language-independent approach*: We find limited studies with low-resourced language in this domain. There is no work found in the multi-lingual approach as observed for offensive language [154]. Few studies have made progress towards language independent approach [2, 10, 77], however, the existing techniques are not directly or indirectly not compared for language-independent or multi-lingual approach.
9. *Incremental Learning from Streaming Data*: There are some studies on Topic extraction on social media content for early depression detection on retrospective data [79] and phase change of the user [7, 92]. The existing studies have rarely use the online streaming data [37] and there is no such study which shows the concept drift [151] in streaming data. A concept drift identifies the level of changing risk in suicidal tendency.
10. *Real-time Applications*: A real-time mental health prediction is yet to be explored because to the best of our knowledge, there is only one study on integration of Internet of Medical Things (IoMT) and Social Media dataset by academic researchers [34].

## 5 Conclusion

This manuscript is an extensive literature survey on predicting suicidal tendency from social media data. The exponential progress in the field of data science for mental health prediction has shown its significance in recent years. The corpus of 92 research articles contains studies over stress, depression and suicide risk detection on social media. However, there is no substantial work on quantifying the suicide risk from the longitudinal data of the user. To handle this and to integrate the existing studies on multiple tasks, an extensive survey is given along with the open challenges and possible research directions. The major contributions of this manuscript are enlisting the available dataset (publicly,

on-request and via signed agreement); introduction to the taxonomy of the mental healthcare; classification of feature extraction and transformation techniques for vector representation; the historical evolution of suicidal tendency detection with timeline; new research directions and open challenges. This manuscript further highlights the important contributions which can be used as benchmark studies in this domain.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References


1. Tadesse MM, Lin H, Xu B, Yang L (2020) Detection of suicide ideation in social media forums using deep learning. Algorithms 13(1):7
2. Cao L, Zhang H, Feng L (2020) Building and using personal knowledge graph to improve suicidal ideation detection on social media. IEEE Trans Multimed. https://doi.org/10.1109/tmm.2020.3046867
3. Almeida H, Briand A, Meurs M-J (2017) Detecting early risk of depression from social media user-generated content. In: CLEF (working notes), 2017
4. Amini P, Ahmadinia H, Poorolajal J, Amiri MM (2016) Evaluating the high risk groups for suicide: a comparison of logistic regression, support vector machine, decision tree and artificial neural network. Iran J Public Health 45(9):1179
5. Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA (2020) A machine learning approach predicts future risk to suicidal ideation from social media data. NPJ Digit Med 3(1):1–12
6. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, Asch DA, Schwartz HA (2018) Facebook language predicts depression in medical records. Proc Natl Acad Sci USA 115(44):11203–11208
7. Sawhney R, Joshi H, Flek L, Shah R (2021) Phase: learning emotional phase-aware representations for suicide ideation detection on social media. In: Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: main volume, 2021, pp 2415–2428
8. Zogan H, Razzak I, Wang X, Jameel S, Xu G (2020) Explainable depression detection with multi-modalities using a hybrid deep learning model on social media. arXiv preprint. arXiv:2007.02847
9. Ma Y, Cao Y (2020) Dual attention based suicide risk detection on social media. In: 2020 IEEE international conference on artificial intelligence and computer applications (ICAICA), 2020. IEEE, pp 637–640
10. Cao L, Zhang H, Feng L, Wei Z, Wang X, Li N, He X (2019) Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. arXiv preprint. arXiv:1910.12038
11. Song H, You J, Chung J-W, Park JC (2018) Feature attention network: interpretable depression detection from social media. In: PACLIC, 2018
12. Ophir Y, Tikochinski R, Asterhan CS, Sisso I, Reichart R (2020) Deep neural networks detect suicide risk from textual Facebook posts. Sci Rep 10(1):1–10
13. Shah FM, Haque F, Nur RU, Al Jahan S, Mamud Z (2020) A hybridized feature extraction approach to suicidal ideation detection from social media post. In: 2020 IEEE Region 10 symposium (TENSYMP), 2020. IEEE, pp 985–988
14. McHugh CM, Corderoy A, Ryan CJ, Hickie IB, Large MM (2019) Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. BJPsych Open 5(2):e18
15. Stone DM (2021) Changes in suicide rates—United States, 2018–2019. Morb Mortal Wkly Rep 70(8):261–268
16. Vincent J (2017) Facebook is using AI to spot users with suicidal thoughts and send them help. Verge
17. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 18:43–49
18. Chancellor S, De Choudhury M (2020) Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 3(1):1–11
19. Luxton DD, June JD, Fairall JM (2012) Social media and suicide: a public health perspective. Am J Public Health 102(S2):195–200
20. Golden RN, Weiland C, Peterson F (2009) The truth about illness and disease. Infobase Publishing, New York
21. De Choudhury M (2013) Role of social media in tackling challenges in mental health. In: Proceedings of the 2nd international workshop on socially-aware multimedia, 2013, pp 49–52
22. Shing H-C, Resnik P, Oard DW (2020) A prioritization model for suicidality risk assessment. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics, 2020, pp 8124–8137
23. Niederkrotenthaler T (2017) Papageno effect: its progress in media research and contextualization with findings on harmful media effects. In: Media and suicide: international perspectives on research, theory, and policy. Routledge, London, pp 133–158
24. Chancellor S, Birnbaum ML, Caine ED, Silenzio VM, De Choudhury M (2019) A taxonomy of ethical tensions in inferring mental health states from social media. In: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp 79–88
25. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M (2015) CLPsych 2015 shared task: depression and PTSD on Twitter. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015, pp 31–39
26. Milne DN, Pink G, Hachey B, Calvo RA (2016) CLPsych 2016 shared task: triaging content in online peer-support forums. In: Proceedings of the third workshop on computational linguistics and clinical psychology, 2016, pp 118–127
27. Preotiuc-Pietro D, Sap M, Schwartz HA, Ungar LH (2015) Mental illness detection at the world well-being project for the CLPsych 2015 shared task. In: CLPsych@ HLT-NAACL, 2015, pp 40–45
28. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen V-A, Boyd-Graber J (2015) Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015, pp 99–107
29. Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H (2015) Recognizing depression from Twitter activity. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, 2015, pp 3187–3196

30. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M (2016) Discovering shifts to suicidal ideation from mental health content in social media. In: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp 2098–2110
31. Sawhney R, Joshi H, Gandhi S, Shah RR (2021) Towards ordinal suicide ideation detection on social media. In: Proceedings of the 14th ACM international conference on web search and data mining, 2021, pp 22–30
32. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, 2016, pp 1480–1489
33. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S (2020) GoEmotions: a dataset of fine-grained emotions. arXiv preprint. arXiv:2005.00547
34. Gupta D, Bhatia M, Kumar A (2021) Real-time mental health analytics using IoMT and social media datasets: research and challenges. Available at SSRN 3842818
35. Turcan E, McKeown K (2019) Dreaddit: a Reddit dataset for stress analysis in social media. arXiv preprint. arXiv:1911.00133
36. Matero M, Idnani A, Son Y, Giorgi S, Vu H, Zamani M, Limbachiya P, Guntuku SC, Schwartz HA (2019) Suicide risk assessment with multi-level dual-context language and BERT. In: Proceedings of the sixth workshop on computational linguistics and clinical psychology, 2019, pp 39–44
37. Lin C, Hu P, Su H, Li S, Mei J, Zhou J, Leung H (2020) SenseMood: depression detection on social media. In: Proceedings of the 2020 international conference on multimedia retrieval, 2020, pp 407–411
38. Gui T, Zhu L, Zhang Q, Peng M, Zhou X, Ding K, Chen Z (2019) Cooperative multimodal approach to depression detection in Twitter. In: Proceedings of the AAAI conference on artificial intelligence, 2019, vol 33, pp 110–117
39. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, Chua T-S, Zhu W (2017) Depression detection via harvesting social media: a multimodal dictionary learning solution. In: IJCAI, 2017, pp 3838–3844
40. Tadesse MM, Lin H, Xu B, Yang L (2019) Detection of depression-related posts in Reddit social media forum. IEEE Access 7:44883–44893
41. Zogan H, Razzak I, Jameel S, Xu G (2021) DepressionNet: a novel summarization boosted deep framework for depression detection on social media. arXiv preprint. arXiv:2105.10878
42. Haque A, Reddi V, Giallanza T (2021) Deep learning for suicide and depression identification with unsupervised label correction. arXiv preprint. arXiv:2102.09427
43. Zogan H, Wang X, Jameel S, Xu G (2020) Depression detection with multi-modalities using a hybrid deep learning model on social media. arXiv preprint. arXiv:2007.02847
44. Cong Q, Feng Z, Li F, Xiang Y, Rao G, Tao C (2018) XA-BiLSTM: a deep learning approach for depression detection in imbalanced data. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), 2018. IEEE, pp 1624–1627
45. De Choudhury M, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. In: Proceedings of the international AAAI conference on web and social media, 2013, vol 7
46. Ford E, Curlewis K, Wongkoblap A, Curcin V (2019) Public opinions on using social media content to identify users with depression and target mental health care advertising: mixed methods survey. JMIR Ment Health 6(11):12942
47. Conway M, O'Connor D (2016) Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 9:77–82
48. Jia J (2018) Mental health computing via harvesting social media data. In: IJCAI, 2018, pp 5677–5681
49. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J (2021) Deep learning-based text classification: a comprehensive review. ACM Comput Surv 54(3):1–40
50. Stankevich M, Isakov V, Devyatkin D, Smirnov I (2018) Feature engineering for depression detection in social media. In: ICPRAM, 2018, pp 426–431
51. Hussain J, Satti FA, Afzal M, Khan WA, Bilal HSM, Ansaar MZ, Ahmad HF, Hur T, Bang J, Kim J-I et al (2020) Exploring the dominant features of social media for depression detection. J Inf Sci 46(6):739–759
52. Nolen-Hoeksema S (1991) Responses to depression and their effects on the duration of depressive episodes. J Abnorm Psychol 100(4):569
53. Farseev A, Samborskii I, Chua T-S (2016) A big data platform for social multimedia analytics. In: Conference: the 2016 ACM, 2016
54. Park M, Cha C, Cha M (2012) Depressive moods of users portrayed in Twitter. In: Proceedings of the 18th ACM international conference on knowledge discovery and data mining, SIGKDD 2012, 2012
55. Preoţiuc-Pietro D, Eichstaedt J, Park G, Sap M, Smith L, Tobolsky V, Schwartz HA, Ungar L (2015) The role of personality, age, and gender in tweeting about mental illness. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015, pp 21–30
56. Fu S, Ibrahim OA, Wang Y, Vassilaki M, Petersen RC, Mielke MM, St Sauver J, Sohn S (2022) Prediction of incident dementia using patient temporal health status. Stud Health Technol Inform 290:757–761
57. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J (2017) Patient subtyping via time-aware LSTM networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp 65–74
58. Losada DE, Crestani F, Parapar J (2019) Overview of eRISK 2019 early risk prediction on the Internet. In: International conference of the Cross-Language Evaluation Forum for European Languages, 2019. Springer, pp 340–357
59. Guntuku SC, Buffone A, Jaidka K, Eichstaedt JC, Ungar LH (2019) Understanding and measuring psychological stress using social media. In: Proceedings of the international AAAI conference on web and social media, 2019, vol 13, pp 214–225
60. Masuda N, Kurahashi I, Onari H (2013) Suicide ideation of individuals in online social networks. PLoS ONE 8(4):62262
61. Mishra R, Sinha PP, Sawhney R, Mahata D, Mathur P, Shah RR (2019) SNAP-BATNET: cascading author profiling and social network graphs for suicide ideation detection on social media. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: student research workshop, 2019, pp 147–156
62. Burdisso SG, Errecalde M, Montes-y-Gómez M (2019) A text classification framework for simple and effective early depression detection over social media streams. Expert Syst Appl 133:182–197
63. Shen T, Jia J, Shen G, Feng F, He X, Luan H, Tang J, Tiropanis T, Chua TS, Hall W (2018) Cross-domain depression detection via harvesting social media. In: International joint conferences on artificial intelligence, 2018
64. Gkotsis G, Oellrich A, Hubbard T, Dobson R, Liakata M, Velupillai S, Dutta R (2016) The language of mental health problems in social media. In: Proceedings of the third workshop on computational linguistics and clinical psychology, 2016, pp 63–73

65. O'Dea B, Larsen ME, Batterham PJ, Calear AL, Christensen H (2017) A linguistic analysis of suicide-related Twitter posts. Crisis J Crisis Interv Suicide Prev 38(5):319
66. Sawhney R, Agarwal S, Neerkaje AT, Aletras N, Nakov P, Flek L (2022) Towards suicide ideation detection through online conversational context. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, 2022, pp 1716–1727
67. Aduragba OT, Yu J, Cristea AI, Shi L (2021) Detecting fine-grained emotions on social media during major disease outbreaks: health and well-being before and during the COVID-19 pandemic. In: AMIA annual symposium proceedings, 2021, vol 2021, p 187. American Medical Informatics Association
68. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E (2021) MentalBERT: publicly available pretrained language models for mental healthcare. arXiv preprint. arXiv:2110.15621
69. Khadhraoui M, Bellaaj H, Ammar MB, Hamam H, Jmaiel M (2022) Survey of BERT-base models for scientific text classification: COVID-19 case study. Appl Sci 12(6):2891
70. Mitchell M, Hollingshead K, Coppersmith G (2015) Quantifying the language of schizophrenia in social media. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015, pp 11–20
71. Resnik P, Armstrong W, Claudino L, Nguyen T (2015) The University of Maryland CLPsych 2015 shared task system. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015, pp 54–60
72. Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T (2020) Public discourse and sentiment during the COVID 19 pandemic: Using latent Dirichlet allocation for topic modeling on Twitter. PLoS ONE 15(9):0239441
73. Zhao S, Wang S, Soleymani M, Joshi D, Ji Q (2019) Affective computing for large-scale heterogeneous multimedia data: a survey. ACM Trans Multimed Comput Commun Appl 15(3s):1–32
74. Schlosberg H (1954) Three dimensions of emotion. Psychol Rev 61(2):81
75. Plutchik R (1980) A general psychoevolutionary theory of emotion. In: Theories of emotion. Elsevier, Amsterdam, pp 3–33
76. Lin H, Jia J, Guo Q, Xue Y, Huang J, Cai L, Feng L (2014) Psychological stress detection from cross-media microblog data using deep sparse neural network. In: 2014 IEEE international conference on multimedia and expo (ICME), 2014. IEEE, pp 1–6
77. Lin H, Jia J, Qiu J, Zhang Y, Shen G, Xie L, Tang J, Feng L, Chua T-S (2017) Detecting stress based on social interactions in social networks. IEEE Trans Knowl Data Eng 29(9):1820–1833
78. Cambria E, Olsher D, Rajagopal D (2014) SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, 2014, vol 28
79. Maupomé D, Meurs M-J (2018) Using topic extraction on social media content for the early detection of depression. In: CLEF (working notes), 2018, vol 2125
80. Saravia E, Chang C-H, De Lorenzo RJ, Chen Y-S (2016) MIDAS: mental illness detection and analysis via social media. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), 2016. IEEE, pp 1418–1421
81. Sawhney R, Manchanda P, Mathur P, Shah R, Singh R (2018) Exploring and learning suicidal ideation connotations on social media with deep learning. In: Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2018, pp 167–175
82. Orabi AH, Buddhitha P, Orabi MH, Inkpen D (2018) Deep learning for depression detection of Twitter users. In: Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic, 2018, pp 88–97
83. Turcan E, Muresan S, McKeown K (2021) Emotion-infused models for explainable psychological stress detection. In: Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, 2021, pp 2895–2909
84. Wang X, Zhang H, Cao L, Feng L (2020) Leverage social media for personalized stress detection. In: Proceedings of the 28th ACM international conference on multimedia, 2020, pp 2710–2718
85. Yang K, Zhang T, Ananiadou S (2022) A mental state knowledge-aware and contrastive network for early stress and depression detection on social media. Inf Process Manag 59(4):102961
86. Lee D, Kang M, Kim M, Han J (2022) Detecting suicidality with a contextual graph neural network. In: Proceedings of the eighth workshop on computational linguistics and clinical psychology, 2022, pp 116–125
87. Moulahi B, Azé J, Bringay S (2017) DARE to Care: a context-aware framework to track suicidal ideation on social media. In: International conference on web information systems engineering, 2017. Springer, pp 346–353
88. Whooley O (2014) Diagnostic and statistical manual of mental disorders (DSM). In: The Wiley Blackwell encyclopedia of health, illness, behavior, and society. Wiley, Hoboken, pp 381–384
89. Leiva V, Freire A (2017) Towards suicide prevention: early detection of depression on social media. In: International conference on Internet science, 2017. Springer, pp 428–436
90. Schwartz HA, Giorgi S, Sap M, Crutchley P, Ungar L, Eichstaedt J (2017) DLATK: differential language analysis toolkit. In: Proceedings of the 2017 conference on empirical methods in natural language processing: system demonstrations, 2017, pp 55–60
91. Park M, McDonald D, Cha M (2013) Perception differences between the depressed and non-depressed users in Twitter. In: Proceedings of the international AAAI conference on web and social media, 2013, vol 7
92. Sawhney R, Joshi H, Shah R, Flek L (2021) Suicide ideation detection via social and temporal user representations using hyperbolic learning. In: Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, 2021, pp 2176–2190
93. Sawhney R, Thakkar M, Agarwal S, Jin D, Yang D, Flek L (2021) HypMix: hyperbolic interpolative data augmentation. In: Proceedings of the 2021 conference on empirical methods in natural language processing, 2021, pp 9858–9868
94. Cheng Q, Li TM, Kwok C-L, Zhu T, Yip PS (2017) Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res 19(7):243
95. Lin H, Jia J, Guo Q, Xue Y, Li Q, Huang J, Cai L, Feng L (2014) User-level psychological stress detection from social media using deep neural network. In: Proceedings of the 22nd ACM international conference on multimedia, 2014, pp 507–516
96. Wang Y, Tang J, Li J, Li B, Wan Y, Mellina C, O'Hare N, Chang Y (2017) Understanding and discovering deliberate self-harm content in social media. In: Proceedings of the 26th international conference on World Wide Web, 2017, pp 93–102
97. Zhou Y, Zhan J, Luo J (2017) Predicting multiple risky behaviors via multimedia content. In: International conference on social informatics, 2017. Springer, pp 65–73

98. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint. arXiv:1301.3781
99. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805
100. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv preprint. arXiv:1908.10084
101. Cer D, Yang Y, Kong S-Y, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Céspedes M, Yuan S, Tar C et al (2018) Universal sentence encoder. arXiv preprint. arXiv:1803.11175
102. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556
103. Wang W, Huang Y, Wang Y, Wang L (2014) Generalized autoencoder: a neural network framework for dimensionality reduction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2014, pp 490–497
104. McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint. arXiv:1802.03426
105. Ive J, Gkotsis G, Dutta R, Stewart R, Velupillai S (2018) Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In: Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic, 2018, pp 69–77
106. Al Asad N, Pranto MAM, Afreen S, Islam MM (2019) Depression detection by analyzing social media posts of user. In: 2019 IEEE international conference on signal processing, information, communication and systems (SPICSCON), 2019. IEEE, pp 13–17
107. Elman JL (1990) Finding structure in time. Cogn Sci 14(2):179–211
108. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint. arXiv:1412.3555
109. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
110. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
111. Wang N, Luo F, Shivtare Y, Badal VD, Subbalakshmi K, Chandramouli R, Lee E (2021) Learning models for suicide prediction from social media posts. arXiv preprint. arXiv:2105.03315
112. Xu Z, Pérez-Rosas V, Mihalcea R (2020) Inferring social media users' mental health status from multimodal information. In: Proceedings of the 12th language resources and evaluation conference, 2020, pp 6292–6299
113. Tavchioski I, Koloski B, Škrlj B, Pollak S (2022) E8-IJS@ LT-EDI-ACL2022-BERT, AutoML and knowledge-graph backed detection of depression. In: Proceedings of the second workshop on language technology for equality, diversity and inclusion, 2022, pp 251–257
114. Naseem U, Khushi M, Kim J, Dunn AG (2022) Hybrid text representation for explainable suicide risk identification on social media. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/TCSS.2022.3184984
115. Yates A, Cohan A, Goharian N (2017) Depression and self-harm risk assessment in online forums. arXiv preprint. arXiv:1709.01848
116. Losada DE, Crestani F, Parapar J (2018) Overview of eRISK: early risk prediction on the Internet. In: International conference of the Cross-Language Evaluation Forum for European Languages, 2018. Springer, pp 343–361
117. Jamil Z (2017) Monitoring tweets for depression to detect at-risk users. PhD Thesis, University of Ottawa
118. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N (2018) SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. arXiv preprint. arXiv:1806.05258
119. Gamaarachchige PK, Inkpen D (2019) Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In: Proceedings of the tenth international workshop on health text mining and information analysis (LOUHI 2019), 2019, pp 54–64
120. Pirina I, Çöltekin Ç (2018) Identifying depression on Reddit: the effect of training data. In: Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop and shared task, 2018, pp 9–12
121. Ji S, Yu CP, Fung S-F, Pan S, Long G (2018) Supervised learning for suicidal ideation detection in online user content. Complexity. https://doi.org/10.1155/2018/6157249
122. Harrigian K, Aguirre C, Dredze M (2020) On the state of social media data for mental health research. arXiv preprint. arXiv:2011.05233
123. Gaur M, Alambo A, Sain JP, Kursuncu U, Thirunarayan K, Kavuluru R, Sheth A, Welton R, Pathak J (2019) Knowledge-aware assessment of severity of suicide risk for early intervention. In: The World Wide Web conference, 2019, pp 514–525
124. Aladağ AE, Muderrisoglu S, Akbas NB, Zahmacioglu O, Bingol HO (2018) Detecting suicidal ideation on forums: proof-of-concept study. J Med Internet Res 20(6):215
125. Burkhardt H, Pullmann M, Hull T, Aren P, Cohen T (2022) Comparing emotion feature extraction approaches for predicting depression and anxiety. In: Proceedings of the eighth workshop on computational linguistics and clinical psychology, 2022, pp 105–115
126. Garg M, Saxena C, Krishnan V, Joshi R, Saha S, Mago V, Dorr BJ (2022) CAMS: an annotated corpus for causal analysis of mental health issues in social media posts. arXiv preprint. arXiv:2207.04674
127. Saxena C, Garg M, Ansari G (2022) Explainable causal analysis of mental health on social media data. In: Proceedings of ICONIP, 2022
128. Naseem U, Khushi M, Kim J, Dunn AG (2022) RHMD: a real-world dataset for health mention classification on Reddit. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/TCSS.2022.3186883
129. Kayalvizhi S, Thenmozhi D (2022) Data set creation and empirical analysis for detecting signs of depression from social media postings. arXiv preprint. arXiv:2202.03047
130. Sivamanikandan S, Santhosh V, Sanjaykumar N, Durairaj T et al (2022) scubeMSEC@ LT-EDI-ACL2022: detection of depression using transformer models. In: Proceedings of the second workshop on language technology for equality, diversity and inclusion, 2022, pp 212–217
131. Whissell C (2009) Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. Psychol Rep 105(2):509–521
132. Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015. Technical report
133. De Smedt T, Daelemans W (2012) Pattern for Python. J Mach Learn Res 13(1):2063–2067
134. Reece AG, Danforth CM (2017) Instagram photos reveal predictive markers of depression. EPJ Data Sci 6:1–12
135. Nam H, Ha J-W, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp 299–307

136. Song X, Nie L, Zhang L, Akbari M, Chua T-S (2015) Multiple social network learning and its application in volunteerism tendency prediction. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp 213–222
137. Rolet A, Cuturi M, Peyré G (2016) Fast dictionary learning with a smoothed Wasserstein loss. In: Artificial intelligence and statistics, 2016. PMLR, pp 630–638
138. Ansari G, Garg M, Saxena C (2021) Data augmentation for mental health classification on social media. arXiv preprint. arXiv:2112.10064
139. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. arXiv preprint. arXiv:1606.00061
140. Moon S, Neves L, Carvalho V (2018) Multimodal named entity disambiguation for noisy social media posts. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics: long papers, 2018, vol 1, pp 2000–2008
141. Ghosal D, Majumder N, Gelbukh A, Mihalcea R, Poria S (2020) COSMIC: commonsense knowledge for emotion identification in conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp 2470–2481
142. Sawhney R, Manchanda P, Singh R, Aggarwal S (2018) A computational approach to feature extraction for identification of suicidal ideation in tweets. In: Proceedings of ACL 2018, student research workshop, 2018, pp 91–98
143. Sinha PP, Mishra R, Sawhney R, Mahata D, Shah RR, Liu H (2019) # suicidal—a multipronged approach to identify and explore suicidal ideation in Twitter. In: Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp 941–950
144. Amanat A, Rizwan M, Javed AR, Abdelhaq M, Alsaqour R, Pandya S, Uddin M (2022) Deep learning for depression detection from textual data. Electronics 11(5):676
145. Yu L, Jiang W, Ren Z, Xu S, Zhang L, Hu X (2021) Detecting changes in attitudes toward depression on Chinese social media: a text analysis. J Affect Disord 280:354–363
146. Belinda CM, Ravikumar S, Arif M et al (2022) Linguistic analysis of Hindi–English mixed tweets for depression detection. J Math. https://doi.org/10.1155/2022/3225920
147. Naseem U, Dunn AG, Kim J, Khushi M (2022) Early identification of depression severity levels on Reddit using ordinal classification. In: Proceedings of the ACM web conference 2022, 2022, pp 2563–2572
148. Zogan H, Razzak I, Wang X, Jameel S, Xu G (2022) Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. In: World Wide Web, 2022, pp 1–24
149. Wang X, Cao L, Zhang H, Feng L, Ding Y, Li N (2022) A meta-learning based stress category detection framework on social media. In: Proceedings of the ACM web conference 2022, 2022, pp 2925–2935
150. Semeraro A, Vilella S, Ruffo G (2021) PyPlutchik: visualising and comparing emotion-annotated corpora. arXiv preprint. arXiv:2105.04295
151. Vioules MJ, Moulahi B, Azé J, Bringay S (2018) Detection of suicide-related posts in Twitter data streams. IBM J Res Dev 62(1):7:1-7:12
152. Li H, Han Y, Xiao Y, Liu X, Li A, Zhu T (2021) Suicidal ideation risk and socio-cultural factors in China: a longitudinal study on social media from 2010 to 2018. Int J Environ Res Public Health 18(3):1098
153. Yazdavar AH, Mahdavinejad MS, Bajaj G, Romine W, Sheth A, Monadjemi AH, Thirunarayan K, Meddar JM, Myers A, Pathak J et al (2020) Multimodal mental health analysis in social media. PLoS ONE 15(4):0226248
154. Ranasinghe T, Zampieri M (2021) Multilingual offensive language identification for low-resource languages. arXiv preprint. arXiv:2105.05996