

Sentiment Analysis of Thought-Related Language in the Context of Cognitive Disorders Using Machine Learning, Transformer-based Architecture and Large Language Models

Tasnim Niger

Islamic University of Technology

Md. Kamrul Hasan

hasank@iut-dhaka.edu

Islamic University of Technology

Hasan Mahmud

Islamic University of Technology

Research Article

Keywords: Machine Learning, Natural Language Processing, Large Language Model, Thought classification, Sentiment Analysis, Textual Dataset

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10942885/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Distorted cognitive behaviors significantly impair decision-making and mental well-being, highlighting the need for innovative methods for addressing negative thought patterns. This study introduces a framework for analyzing sentiment polarity in thought-related text, classifying textual expressions into positive and negative sentiments, with potential applications in computational mental health research. A large-scale dataset ($n = 156,785$) was curated by collecting and annotating social media texts, to capture both positive and negative sentiment polarity in thought-related expression. A comprehensive evaluation was conducted across various machine-learning (ML) models, from traditional classifiers to advanced transformer-based architectures. The study utilized the Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) feature representations, with 5-fold cross-validation for robustness. Among the traditional models, Logistic Regression (LR) and Stochastic Gradient Descent (SGD) achieved comparable performance, with BoW-based Logistic Regression yielding the highest accuracy of $92.87\% \pm 0.13\%$ (95% confidence interval (CI): 0.927, 0.930). Among the transformer models, Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) provided the best overall performance with an accuracy of $84.07\% \pm 0.15\%$ (95% CI: 0.839–0.843), demonstrating stable precision-recall trade-offs across classes. In addition, a decoder-only language model (Llama-2) analysis achieved moderate classification performance with $80.20\% \pm 2.35\%$ accuracy using instruction-based prompting. Overall, the findings demonstrate the effectiveness of lightweight machine learning models and distilled transformers for large-scale sentiment analysis of thought-related social media texts. Practically, the framework is positioned as a research-oriented tool that requires human oversight for future mental health-related applications.

1 Introduction

Negative thinking patterns are widely studied in psychology due to their association with emotional distress and reduced well-being (Gilbert, 1998; Moulds & McEvoy, 2025). Such patterns have been shown to influence how individuals interpret their experiences, often leading to pessimistic self-appraisal and heightened self-criticism (Beck, 1979; Ehret et al., 2015). In particular, self-critical thinking styles increase the frequency of negative automatic thoughts and are associated with elevated depressive symptoms (Pedro et al., 2019). It is also important to note that negative thoughts can be indicators of various psychological conditions: repetitive negative thinking is now recognised as a transdiagnostic process implicated in the onset and maintenance of multiple disorders (Spinhoven et al., 2018; Moulds & McEvoy, 2025), and the frequency of cognitive distortions expressed in patients' language has been shown to track depression symptom severity (Lalk et al., 2024). For instance, cognitive distortions such as catastrophizing, overgeneralization, and personalization frequently go undetected in everyday self-expression, yet they are recognized precursors and symptoms of anxiety and depressive disorders (Simms et al., 2017), and recent longitudinal evidence confirms that individuals with anxiety and depression express more distorted thinking in their everyday social media language (Rutter et al., 2025). Left unchecked, such self-reinforcing irrational thought patterns can trap individuals in negative feedback loops of unhealthy thinking (Shickel et al., 2020).

Moreover, manually identifying these patterns requires clinically trained raters, which is time-consuming and difficult to scale, highlighting the need for automated approaches to detect negative thinking patterns in text (Tauscher et al., 2023). A thorough evaluation is vital for accurate diagnosis and support in detecting these negative thinking patterns. To this end, this study proposes a supervised NLP-based classification model, on the premise that systematically identifying and challenging these dysfunctional thoughts can help individuals develop healthier thinking patterns (Wiemer-Hastings et al., 2004) and improve their emotional well-being (Persons & Burns, 1985).

According to the World Health Organization (WHO, 2021), one in seven people worldwide faces mental health challenges, highlighting the crucial need to enhance access to resources and improve treatment options. Given that the existing literature offers no convincing insights in this regard (Aboraya et al., 2006; Alavi et al., 2024), there is a clear opportunity to improve patient care through automated, language-based identification of negative thought patterns. This study addresses this gap by developing and evaluating an NLP-based classification approach that detects negative thinking patterns directly from unstructured text, thereby supporting earlier recognition and more scalable intervention.

In these perspectives, *Natural Language Processing (NLP)* techniques which involve different supervised learning algorithms and models can help to tackle the unstructured text classification of users' emotional thought in social spaces (Gasparetto et al., 2022; Jurasky & Martin, 2008; Razno, 2019). Nowadays, different learning models analyze categorical data in behavioral sciences, extract relevant features, and learn patterns from labeled examples to make predictions or assign appropriate labels to new, unseen text documents (Dehghan et al., 2022). Empirically, this study shows that these algorithms can include different ML classification techniques, or even pre-trained Bidirectional Encoder Representations from Transformers (BERT)-based models to enhance the analysis and results. Also, generative models, such as *LLMs* (Chang et al., 2023), can be used for such in-context learning or unstructured data analysis. For instance, Sun et al. (2023) produce unique textual responses via diverse sampling strategies over multiple runs.

Along these lines, this research explores the use of NLP-based text classification tasks for the automatic identification of positive and negative thought patterns expressed in social media text. To address the identified research gap, this study develops a large-scale thought polarity dataset constructed from publicly available social media posts and systematically evaluates conventional machine learning algorithms, transformer-based encoder models, and a prompt-based Llama-2-13B-Chat model for binary thought classification. Model performance is assessed using stratified five-fold cross-validation and evaluated in terms of accuracy, precision, recall, F1-score, mean $\pm$ standard deviation, and 95% confidence intervals. In this study, a distinction is made between a thought-related language and its expressed sentiment polarity. A thought represents cognitive content involving ideas, beliefs, interpretations, or mental representations, whereas sentiment polarity concerns the positive or negative evaluative valence expressed through language (VandenBos, 2007; Hutto & Gilbert, 2014). Thinking itself is not directly observable and must generally be inferred from behaviour or self-report (VandenBos, 2007). For example, VADER is a lexicon- and rule-based sentiment analysis method designed to quantify sentiment intensity in text, particularly social media text, and therefore does not directly identify underlying cognitive thought patterns or psychological disorders (Hutto & Gilbert, 2014). Accordingly, the terms *positive* and *negative thought polarity* in this study refer specifically to the sentiment-derived valence of self-expressed thought statements collected from social media. These labels are treated as an operational proxy for the polarity of expressed thoughts rather than as direct measurements of underlying cognition, cognitive distortions, or mental-health conditions. Expert evaluation is subsequently used to examine whether these automatically assigned polarity labels are consistent with human judgement.

The main contributions of this research are listed below:

- A large-scale dataset was curated by collecting and annotating thought-related social media texts in the context of cognitive and mental health research, focusing on positive and negative sentiment polarity. This dataset is made publicly available to support further research initiatives.
- A comprehensive evaluation of multiple machine learning models, ranging from traditional classifiers to advanced transformer-based architectures and a large language model, was conducted for sentiment classification of thought-related text, identifying the most effective modelling approaches.

The rest of the paper is organized as follows. Section 2, reviews the related research on social media analysis, computational thought classification, and machine learning techniques for mental health text analysis. Section 3 covers the Methodology and experimental setup phase of this study, including dataset collection and preprocessing, features, and learning algorithms used for classifying the positive and negative thoughts. Section 4 presents the results obtained; Section 5 covers the discussion and implications of the study; Section 6 includes the limitations and future work, and ethical considerations. Finally, Section 7 provides the conclusion of the research.

2. Related Work

2.1 Social Media Analysis for Mental Health Detection

Identifying mental health disorders requires innovative computational approaches to support timely assessment and intervention in social media texts and mental health detection (Anish & George, 2024; Kilbourne et al., 2018). Although substantial progress has been made in applying artificial intelligence (AI) to mental health, existing studies continue to highlight the need for more effective computational techniques for identifying and supporting individuals experiencing psychological difficulties across media (Vandoni et al., 2024).

Social media has emerged as an important source of behavioural and linguistic information because users frequently share their daily experiences, opinions, emotions, and health-related concerns. Consequently, Social Media Analysis (SMA) has received increasing attention for monitoring public opinion and detecting abnormal events, including disease outbreaks, natural disasters, terrorism, and health-related conditions (Goswami & Kumar, 2016). The rapid growth of Online Social Networks (OSNs) has further motivated the development of information extraction methods capable of analysing large volumes of user-generated content (Chancellor & De Choudhury, 2020; Umair et al., 2017). Within mental health research, SMA has been used to detect behavioural disorders and support AI-assisted interventions by analysing linguistic patterns in social media posts (Raslescu et al., 2022).

Computationally, several studies have explored machine learning techniques for analysing mental health using multimodal data. For instance, Joshi and Kanoongo (2022) reviewed AI and machine learning methods for depression and emotion detection, covering approaches based on facial expressions, medical images, conversational agents, and social media text. Similarly, Nova (2023) demonstrated that machine learning models can effectively classify mental disorders from social media text, highlighting their potential for automated mental health screening. Rojas-Barahona et al. (2018) proposed a deep learning framework using word embeddings and a corpus of over 500,000 peer-support posts to identify both cognitive distortions and the associated emotions and situations.

More recently, Niger et al. (2022) introduced a virtual Cognitive Behavioural Therapy (vCBT) framework for university students during the COVID-19 pandemic. Their machine learning system classified depression, anxiety, and internet addiction, demonstrating the feasibility of AI-assisted behavioural disorder screening. Likewise, Bathina et al. (2021) analysed linguistic patterns associated with depression on Twitter using principles derived from Cognitive Behavioural Therapy (CBT). Rather than relying solely on predictive modelling, their study investigated how distorted thinking is reflected in online language and showed consistency with established clinical observations regarding depression.

2.2 Computational Thought Classification and Cognitive Distortion Analysis

Although sentiment analysis has become a well-established natural language processing (NLP) task, sentiment and thought represent different psychological constructs. Sentiment reflects emotional states influenced by personal experiences and cultural context, whereas thought involves cognitive processes related to reasoning, interpretation, and decision-making (Tice et al., 2018). Positive thinking further represents an optimistic cognitive orientation characterised by confidence, resilience, and hopeful expectations about future events (Garg, 2022). Thoughts are expressions of underlying appraisal processes, not merely of affective states. Consequently, sentiment analysis datasets, including those proposed by Kabir et al. (2023) and Rahman et al. (2024), are not directly suitable for computational thought classification because they primarily capture emotional polarity rather than cognitive thought patterns.

Recent research has increasingly focused on identifying cognitive distortions and maladaptive thinking patterns using NLP. Bathina et al. (2021) demonstrated that individuals with depression express significantly more distorted thinking on social media, providing evidence that online language can reveal meaningful cognitive characteristics. However, their work primarily examined cognitive distortions associated with depression rather than developing a general framework for large-scale thought polarity classification. More broadly, existing studies have primarily concentrated on depression detection, emotion recognition, or cognitive distortion identification rather than developing expert-validated datasets and systematically benchmarking conventional machine learning, transformer-based models, and large language models for large-scale positive and negative thought classification (Rojas-Barahona et al., 2018; Bathina et al., 2021; Sharma et al., 2023a, 2023b).

These studies demonstrate the growing potential of NLP for computational mental health while also revealing an important research gap. Existing work has largely concentrated on depression detection, emotion recognition, or cognitive distortion identification, whereas large-scale computational classification of positive and negative thought patterns remains relatively underexplored.

2.3 Large Language Models for Mental Health NLP and Research Gap

Beyond thought classification, recent advances in large language models (LLMs) have enabled automated cognitive reframing. Sharma et al. (2023a) developed a language model system to support cognitive restructuring using reinforcement learning with a GPT-2-based model. Their framework was evaluated through field studies and randomised controlled trials involving more than 15,000 participants, demonstrating reductions in users' negative emotions through AI-assisted reframing.

In a related study, Sharma et al. (2023b) investigated Human–Language Model Interaction for cognitive reframing. They proposed a framework based on seven linguistic attributes and trained a GPT-3 model using a dataset of approximately 600 situations and thoughts. A field study involving 2,067 participants showed that users generally preferred reframed responses that were empathetic, specific, and psychologically supportive.

Direct quantitative comparison with the present study is challenging because the datasets, objectives, and evaluation protocols differ substantially. Whereas Sharma et al. (2023a, 2023b) focused on practitioner-curated datasets containing between 600 and 15,531 instances and evaluated user-reported psychological outcomes, including reductions in negative emotions, the present study constructs a substantially larger dataset of 129,080 social media thoughts and evaluates classification performance using stratified five-fold cross-validation, confidence intervals, accuracy, precision, recall, and F1-score. Furthermore, Sharma et al. primarily proposed design principles for effective cognitive reframing rather than developing a large-scale benchmark for thought polarity classification.

Compared with previous studies, the present work examines a large publicly available dataset with expert-supported label validation and provides a systematic comparison of conventional machine learning, encoder-based transformer models, and a decoder-only large language model within a unified evaluation framework (Sections 3 and 4). The proposed framework also employs rigorous statistical evaluation through stratified five-fold cross-validation and confidence intervals, providing a robust benchmark for large-scale computational thought polarity classification. While Sharma et al. demonstrated the practical effectiveness of large language models through user-centred field studies, the present work establishes a computational foundation that can support future research on thought classification and downstream applications, including automated thought conversion and AI-assisted cognitive support.

2.4 Theoretical Framework Guiding the Study

This study is guided by cognitive theories of emotion and psychopathology, principally Beck's cognitive model and Ellis's rational emotive behaviour framework. Beck's cognitive theory posits that emotional distress does not arise from events themselves but from systematically

biased interpretations of those events, organised around negative cognitive schemas concerning the self, the world, and the future - the "negative cognitive triad" (Beck, 1979; Beck et al., 1979). These schemas manifest as automatic negative thoughts: spontaneous, self-referential verbalisations that individuals often accept as accurate reflections of reality (Beck & Haigh, 2014). Ellis's (1962) ABC model similarly holds that beliefs (B) mediate between activating events (A) and emotional consequences (C), implying that maladaptive emotional outcomes can be traced to identifiable, expressible beliefs. Burns (1980) later operationalised these constructs into a widely used taxonomy of cognitive distortions, which underpins the thought-identification stage of Cognitive Behavioural Therapy (CBT).

Two premises of these theories make the present study computationally tractable. First, both frameworks treat verbalised thoughts as observable expressions of underlying cognition, meaning that the language individuals produce constitutes legitimate empirical evidence of their thinking patterns (Tausczik & Pennebaker, 2010). Second, CBT's clinical procedure of identifying and labelling thoughts as adaptive or maladaptive, traditionally performed manually through thought records, in essence, a classification task. The present study operationalises one linguistic dimension of this process at scale by using sentiment-derived polarity as a computational representation of the evaluative valence expressed in thought-related language. Previous research has shown that cognitive distortions can be reflected in social media language (Bathina et al., 2021) and that cognitive distortion categories can be reliably detected in text (Shickel et al., 2020; Tauscher et al., 2023). Building on this theoretical context, the present study focuses on the polarity of thought-related expressions, rather than clinical diagnosis or specific cognitive distortion categories, as the unit of computational analysis. Importantly, sentiment-derived polarity is treated as an operational linguistic proxy and not as a direct measurement of underlying cognition.

In summary, although previous studies have demonstrated the effectiveness of large language models for cognitive reframing and user-centred interventions, they relied on relatively small, practitioner-curated datasets and did not establish a scalable computational framework integrating large-scale thought classification with expert-validated data and systematic benchmarking across model families. This gap motivates the framework proposed in this study.

3 Methods and Tools

This study proposes a computational framework for classifying the sentiment polarity of thought-related language expressed in social media text. To enable large-scale supervised learning, initial positive and negative polarity labels were generated using the VADER (Valence Aware Dictionary and sEntiment Reasoning) sentiment analyzer (Hutto & Gilbert, 2014), which is specifically designed for sentiment analysis of social media and informal text and has demonstrated robust performance in prior validation studies (Bharathi et al., 2024). In this study, VADER serves as an initial sentiment-based labeling method, and the resulting polarity labels were subsequently evaluated by psychology experts to assess their consistency with human judgement.

As shown in Fig. 1, the overall thought classification workflow and methodology of this paper which consists of four main stages are represented and explained.

Social media texts are labeled with VADER and preprocessed, then passed to three branches: traditional ML classifiers with stratified 5-fold cross-validation, transformer-based encoder models, and a prompt-based Llama-2-73B-Chat classifier. All models are evaluated using stratified five-fold cross-validation, with performance reported as accuracy, precision, recall, F1-score, mean $\pm$ SD, and 95% confidence intervals before comparative model selection.

The thought classification workflow and methodology are explained as follows:

1. Data acquisition and thought labeling: using VADER (Valence Aware Dictionary and sEntiment Reasoning) sentiment analyzer and expert validation,
2. Text preprocessing, including cleaning, normalization, contraction expansion, custom stop-word removal, and lemmatization),
3. Model training using three branches- traditional machine-learning classifiers namely, Logistic Regression (LR) (Stoltzfus, 2011), Decision Tree (DT) (Song & Ying, 2015), Multinomial Naive Bayes (MNB) (Abbas et al., 2019), Random Forest (RF) (Belgiu & Drăguț, 2016), and Stochastic gradient descent (SGD) (Newton et al., 2018) evaluated using stratified k-fold cross-validation (Zeng et al., 2000); transformer encoder models (BERT, DistilBERT, RoBERT and ALBERT) and a prompt-based decoder-only LLM Llama-2 model evaluated using the same cross-validation procedure.
4. Performance evaluation based on accuracy, precision, recall, F1 score, and 95% confidence intervals, followed by comparative model selection. The overall system architecture is illustrated in Fig. 1, and each stage is described in the following subsections.

3.1 Data Description

3.1.1 Data Collection and Sample Construction

The data collection procedure involved web scraping from Reddit and using online social networks (OSNs) such as Facebook, Twitter, and Instagram, which enabled this study to search for news, statuses, and discussions that matched both positive and negative thinking. We used the Python library (Python Reddit API Wrapper) *PRAW* to scrape the data from Reddit (PRAW) (Boe, 2014). We used the polarity score method of the *VADER (Valence Aware Dictionary and sEntiment Reasoning) sentiment analyzer* to return the probability of a given input sentence being positive, negative, or neutral (Das et al., 2021). The study then used VADER's compound score, which ranges from - 1 negative to + 1 positive, and applied conservative thresholds to obtain high-confidence polarity labels. Specifically, we defined texts with a compound score greater than 0.2 as positive and less than - 0.2 as negative, treating all intermediate values in the interval [-0.2, 0.2] as neutral. This wider neutral band allowed us to discard borderline or ambiguous cases and retain only clearly positive or negative thoughts, thereby reducing potential label noise in the automatically annotated dataset.

Furthermore, texts were collected from August 2022 to January 2024 using specific positive keywords such as #Success, #Life, #Happiness, #Dream, etc., and negative keywords such as #Difficulties, #Feeling, #Failure, #Worry, and many more, to build the dataset. To find these keywords, as a reference, we used the dataset of Raslescu et al. (2022) consisting of 150 positive and 150 negative adjectives describing personality characteristics, and another database of *English EMotional TErms (EMOTE)* (Grühn, 2016), which provides positive and negative subjective ratings for 1287 nouns and 985 adjectives. After the construction of our dataset, we found that some data had a negative-to-positive correspondence. For example, the negative thought "I am feeling bad" has a possible positive convertible thought, "I am so good". Therefore, we can use this dataset for negative-to-positive thought conversion.

3.1.2 Dataset Validation and Label Reliability

VADER was selected for the initial labeling due to its established effectiveness in sentiment analysis, particularly with social media text (Reshi et al., 2022). Comparative studies have indicated its superior performance in addressing nuanced expressions, including slang, emojis, and variations in sentiment polarity (Jain et al., 2023; Oliveira et al., 2023). To assess the reliability of the automatically generated sentiment labels, an expert validation study was conducted on a stratified random subset of the dataset. This process aimed to assess the reliability of automated labeling and identify potential biases. A total of 500 randomly selected thought statements (250 positive and 250 negative) were independently reviewed by two licensed psychology practitioners with formal training in Cognitive Behavioral Therapy (CBT) and Dialectical Behavior Therapy (DBT). The psychologists evaluated each statement using a five-point Likert rating scale. For negative statements, the response categories were 1 = Not Negative, 2 = Slightly Negative, 3 = Moderately Negative, 4 = Negative, and 5 = Strongly Negative. For positive statements, the response categories were 1 = Not Positive, 2 = Slightly Positive, 3 = Moderately Positive, 4 = Positive, and 5 = Strongly Positive. This approach enabled assessment beyond binary classification by capturing borderline cases.

Expert validation generally supported the VADER-assigned sentiment polarity labels. As shown in Appendix Table C.1, for the negative samples, 80.8% and 70.0% of statements were rated negative or strongly negative (ratings 4–5) by the first and second validators, respectively. Likewise, 84.4% and 64.8% of the positive samples were rated as positive or strongly positive ratings (4–5). Differences between the validators were primarily related to the perceived strength of sentiment rather than the sentiment polarity itself, as shown in the analysis presented in Appendix C. Overall, these findings provide human-evaluation support for the VADER-derived polarity labels and their use as classification targets for the computational analysis of sentiment polarity in thought-related social media text.

3.2 Data Preprocessing

3.2.1 Data Cleaning

The constructed dataset underwent meticulous preparation to utilize text data for predictive modeling (Muller & Guido, 2016). This procedure entailed a thorough and systematic approach to organizing and structuring the data to be effectively analyzed. The preparatory stage was crucial since it formed the foundation for all the modeling process and analysis. By preparing and cleaning the text data carefully, researchers and analysts can ensure that the findings are accurate and meaningful (De Casterlé et al., 2012).

Therefore, text preprocessing involved cleaning the raw data to transform it into a presentable format which was both analyzable and predictable for the desired task. To better organize the text, diacritical accent characters such as non-ASCII characters were excluded, and converted text into lowercase. Further, irrelevant emojis, URLs, and unnecessary spaces were removed. Common contractions, for explanation, can't = cannot, were expanded manually as a part of the cleaning process. A pre-defined dictionary was used to replace contractions with their expanded forms. Furthermore, punctuation and special characters were removed. Additionally, spelling was verified with spell checker function, and an auto spell function was applied to correct them. Personal pronouns were consolidated while preserving their form to maintain the subject's agency. For example, "I", "we", "me" can all be represented as "me" for uniformity.

Notably, in classifying negative thoughts, we were meticulous about handling negative words such as "cannot" and "don't," recognizing their role as key indicators of negativity and cognitive distortions. Including them in stopwords would strip valuable polarity information,

potentially misclassifying phrases like "not happy" as positive. Therefore, we developed a customized stopwords list that specifically excludes such negation terms and their contractions (e.g., 'not', 'no', 'can't', 'don't', 'didn't', 'doesn't', 'won't', 'couldn't', 'hadn't', 'hasn't', 'haven't', 'isn't', 'mustn't', 'shan't', 'shouldn't', 'wasn't', 'weren't', 'wouldn't') while also preserving select pronouns (e.g., 'i', 'we', 'me', 'you', 'they'). The list was derived from a base set of common English stopwords (aligned with NLTK's defaults), with these exclusions to prioritize sentiment preservation.

Moreover, we implemented pronoun consolidation to normalize the remaining personal pronouns, grouping semantically similar ones into unified representatives while maintaining the subject's agency and contextual perspective in the text. Stopwords removal before consolidation allows preserved pronouns to be accurately mapped without premature filtering, thereby reducing lexical variation and feature sparsity in vectorized representations without losing key signals for binary positive/negative classification. Specifically, the consolidation was applied using a rule-based function that processes a space-separated string of tokens using the Python package of Natural Language Toolkit (NLTK) (Bird et al., 2009) .

For each token:

- First-person pronouns ('i', 'we', 'me') are mapped to 'me', consolidating singular/plural and subject/object forms to uniformly capture self-referential sentiment.
- The second-person pronoun ('you') is retained as 'you', due to its limited variation and importance in directive or interpersonal contexts.
- Third-person pronouns ('they', 'them') are mapped to 'they', combining subject and object forms for consistency.

The unmatched tokens remained unchanged, and the processed list was rejoined into a string. This implementation assumes lowercase input for case-insensitive handling and focuses on high-frequency personal pronouns observed in our dataset. The choices were justified by exploratory data analysis, which showed these pronouns frequently correlate with negative sentiments (e.g., self-focused negatives like "I can't"), and consolidation improved model generalization by 2–5% in F1-scores during ablation tests, particularly for the negative class. We limited the scope to these groups to avoid over-normalization, excluding possessives (e.g., 'my') and gender-specific pronouns (e.g., 'he', 'she') to retain potential nuanced signals in rare cases.

We then tokenized the text, removing matches from this custom list while preserving pronouns and essential words that indicated cognitive distortions. This tailored approach enabled our model to effectively capture both positive and negative sentiments within the text. Finally, we performed lemmatization on the text to find the root form of words or lemmas in NLP (Khyani et al., 2021; Lane et al., 2019).

ML algorithms require target variables to be encoded as numerical values (Muller & Guido, 2016). Thus, we converted the target variables "positive" and "negative" into numeric labels 1 and 0. Table 1 shows the top five rows from our data set after the data cleaning and encoding processes. The "Thought" column contains the original thought-related textual features. Therefore, we applied the data cleaning and transformation steps to this column. The "Label" column contains the sentiment-derived target labels encoded as numerical values, where 0 represents negative sentiment polarity and 1 represents positive sentiment polarity. The classification task is therefore to predict the sentiment polarity label associated with each textual expression. Following preprocessing, the cleaned text was stored in the "Thought_clean" column, while the "n_of_words" column represents the number of words in the original text.

Table 1
Top 5 rows from the data set after the cleaning and encoding process

Thought	Label	Thought_clean	n_of_words
Everyone is probably mad at me	0	everyone probably mad me	6
You're braver than you believe	1	you brave you believe	6
When I open my eyes, you're all that I want to...	1	me open eye you all me want see	14
They clearly don't want to be friends with me ...	0	they clearly not want friend me anymore	12
I've never done it before.	0	me never done	7

3.2.2 Data Synthesis and Classification

Further data cleaning, statistical analysis, and lemmatization processes were conducted to prepare the dataset for exploratory data analysis (EDA). The resulting set of documents was then handled with null values and duplicate data were removed. The original dataset consisted of 156785 texts. There were 27705 duplicate data. After removing duplicate values, positive thoughts were 71185 i.e. 55.15% of the dataset, and negative thoughts were 57895 i.e. 44.85% of the dataset. Therefore, we have a balanced dataset. Figure 2 illustrates the distribution of negative and positive sentiment polarity labels in the thought-related social media dataset.

3.3 Model Building for Thought Classification

3.3.1 Traditional Machine Learning Process

Feature Extraction. Machines cannot process text data in raw form, so we must break it down into numeric vectors (Muller & Guido, 2016). To utilize text data within a ML algorithm, it must first be transformed into numerical values, such as integers or floating-point values. This conversion process allows the algorithm to process and analyze the data effectively, ultimately leading to more accurate results. Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) are frequency-based word embedding techniques that help convert text sentences into numeric vectors (Kozhevnikov & Pankratova, 2020). All ML classifier models' results were evaluated on TF-IDF and BoW vectorized data (Qader et al., 2019).

As shown in Fig. 1, after tokenization in the preprocessing step, the text data were transformed into a normalized form and allowed for accurate analysis using the TF-IDF method. The most frequently occurring token found was 'like,' with a count of 5025. The BoW Count Vectorizer (CV) method was then applied to convert the dataset into a matrix format, adapting to the BoW model. It is important to note that this model generates vectors indicating the frequency of each word's occurrence in a thought, while the TF-IDF model provides insight into the importance of each word. After preprocessing, we examined the size of the training feature set for both models (103264, 24882), and the testing feature set was calculated to be (25816, 24882).

ML Classification Models. After vectorization was performed (see previous step), the next process in our data analysis was to split the data into training and test sets. The final and crucial step was to build models with different ML techniques (as shown in Fig. 1) to train the classifiers. This step involves feeding the extracted features into the classifier, which is then trained to identify the patterns and relationships between the features and the corresponding labels. In this study, model performance was evaluated using 5-fold stratified cross-validation rather than a single train-test split. This approach provides a more robust and reliable assessment by training and testing the models across multiple data partitions, thereby reducing variance associated with any single split. Stratification ensured that class distributions were preserved in each fold. This strategy is especially recommended when evaluating and comparing machine learning models to obtain stable performance estimates (Raschka, 2018). As illustrated in Fig. 1, models used *Logistic Regression (LR)* (Stoltzfus, 2011), *Decision Tree (DT)* (Song & Ying, 2015), *Multinomial Naive Bayes (MNB)* (Abbas et al., 2019), *Random Forest (RF)* (Belgiu & Drăguț, 2016), and *Stochastic gradient descent (SGD)* (Newton et al., 2018) to classify text.

3.3.2 Transformer-based LLM

Transformers represent a distinct neural network architecture that comprises an encoder and a decoder (Vaswani et al., 2017; Wolf et al., 2019). The model is anchored in the concept of self-attention, enabling it to scrutinize the complete input sequence while computing the output for every element in the series. The self-attention mechanism employed by transformers calculates the weighted sum of values, depending on the queries and keys' similarity. This feature empowers the model to concentrate selectively on distinct portions of the input sequence as it processes data. Large Language Models (LLMs) started gaining prominence after the invention of Transformer Models. Currently, there are broadly three types of Transformer-based LLMs, based on their usage of Encoder and Decoder blocks: 1) Encoder-Decoder, 2) Encoder-only, and 3) Decoder-only architecture.

The study used only the encoder blocks for classification tasks. With the help of the hugging-face library, as shown in Fig. 1, this study employed various pre-trained BERT models (Tsai et al., 2019), including *BERT* (Devlin et al., 2019), *DistilBERT* (Sanh et al., 2019), *RoBERTa* (Liu et al., 2019), and *ALBERT* (Lan et al., 2019). When working with the BERT variants, text sequence was inputted along with the corresponding lengths of each sequence. The maximum sequence length was 128. The input text was first tokenized and encoded using a pre-trained BERT transformer, which generated contextualized token representations. The final hidden state corresponding to the [CLS] token was used as a sentence-level representation. This representation was passed through a fully connected layer with 128 hidden units and a Rectified Linear Unit (ReLU) activation function (Agarap, 2018), followed by a sigmoid-activated output layer to produce a binary prediction between 0 and 1 (Pratiwi et al., 2020). All transformer models were fine-tuned using 5-fold stratified cross-validation for up to five epochs. Early stopping was implemented to monitor validation loss and reduce the risk of overfitting. The models were trained with the Adam optimizer, applying a learning rate of 1.5×10^{-5} , while binary cross-entropy loss was applied, and accuracy was designated as the primary performance metric. As defined in Fig. 1, this study used bert-base-uncased, distilbert-base-uncased, roberta-base, and albert-base-v2 as the tokenizer and pre-trained transformer models for BERT, DistilBERT, RoBERTa, and ALBERT respectively. While cross-validation increases computational cost, especially for larger transformer-based models such as BERT, it provides a fairer and statistically robust comparison across different models.

3.3.3 Decoder-only LLM

Decoder-only transformers are the most commonly used transformers in modern Language Models (LLMs). These transformers use only decoder blocks to make predictions from large, unsupervised corpora of text prompts in an auto-regressive manner. The use of Decoder-only

LLMs has allowed Generative AI to explore the earlier methods and problem statements of NLP and to create a more generalizable AI solution that can adapt to any NLP task required, whether it is Classification, Question Answering, or Translation.

Using a Decoder-only block simplifies the training process and creates a more general-purpose solution. LLM Response acts as a Discriminative Classifier by generating classification labels. The *Llama* models are good examples of Decoder-Only LLMs (Touvron et al., 2023). As described in Fig. 1, for this research purposes, the Llama-2-13B-Chat model was implemented using the llama-cpp-python inference framework. The quantized GGUF model was executed with a context window of 4096 tokens, and GPU acceleration was enabled by offloading all compatible model layers to the available GPU. Few-shot prompting was used for binary thought classification (discussed in Appendix A and B), and model performance was evaluated using stratified five-fold cross-validation. This feature makes LLMs particularly useful for generating responses to a wide range of prompts and queries.

4 Results

4.1 Model Performance Analysis

Selecting the most suitable model for the specific dataset was crucial for accurately assessing performance. The study examined several traditional ML models as well as transformer-based models by employing standard classification metrics such as precision, recall, F1-score, accuracy, and 95% confidence intervals (CI). To ensure a robust estimate of the model performance, we calculated the mean and standard deviation across the cross-validation folds.

Table 3

Performance Analysis of TF-IDF and BoW vectorized data with a range of classifiers

Classifier	Thought	TF-IDF vectorized data					BoW vectorized data				
		Precision	Recall)	F1-score	Accuracy	95% CI	Precision	Recall	F1-score	Accuracy	95% CI
LR	Positive	0.9282 ± 0.0009	0.9365 ± 0.0018	0.9323 ± 0.0011	0.9250 ± 0.0012	0.9236, 0.9265	0.9311 ± 0.0012	0.9403 ± 0.0019	0.9357 ± 0.0012	0.9287 ± 0.0013	0.9271, 0.9303
	Negative	0.9211 ± 0.0021	0.9109 ± 0.0011	0.9160 ± 0.0013			0.9257 ± 0.0022	0.9145 ± 0.0016	0.9201 ± 0.0014		
DT	Positive	0.8555 ± 0.0019	0.8943 ± 0.0025	0.8745 ± 0.0021	0.8584 ± 0.0023	0.8555, 0.8613	0.8503 ± 0.0021	0.8989 ± 0.0041	0.8739 ± 0.0021	0.8570 ± 0.0022	0.8542, 0.8597
	Negative	0.8623 ± 0.0031	0.8143 ± 0.0025	0.8376 ± 0.0026			0.8663 ± 0.0046	0.8054 ± 0.0034	0.8348 ± 0.0024		
MNB	Positive	0.8937 ± 0.0025	0.9115 ± 0.0023	0.9025 ± 0.0023	0.8914 ± 0.0026	0.8882, 0.8946	0.9072 ± 0.0017	0.8967 ± 0.0023	0.9019 ± 0.0018	0.8925 ± 0.0020	0.8900, 0.8949
	Negative	0.8885 ± 0.0029	0.8667 ± 0.0032	0.8774 ± 0.0030			0.8748 ± 0.0026	0.8872 ± 0.0021	0.8810 ± 0.0022		
RF	Positive	0.9129 ± 0.0021	0.9343 ± 0.0016	0.9235 ± 0.0016	0.9146 ± 0.0018	0.9124, 0.9168	0.9137 ± 0.0021	0.9326 ± 0.0026	0.9231 ± 0.0017	0.9143 ± 0.0018	0.9120, 0.9166
	Negative	0.9169 ± 0.0019	0.8904 ± 0.0029	0.9034 ± 0.0021			0.9150 ± 0.0030	0.8917 ± 0.0029	0.9032 ± 0.0020		
SGD	Positive	0.9237 ± 0.0019	0.9275 ± 0.0020	0.9256 ± 0.0011	0.9178 ± 0.0012	0.9163, 0.9192	0.9290 ± 0.0023	0.9411 ± 0.0031	0.9350 ± 0.0011	0.9279 ± 0.0012	0.9264, 0.9293
	Negative	0.9104 ± 0.0022	0.9057 ± 0.0026	0.9081 ± 0.0013			0.9265 ± 0.0034	0.9115 ± 0.0033	0.9189 ± 0.0013		

Note: Precision, recall, F1-score, and accuracy are reported as mean ± SD over five stratified cross-validation folds. The 95% confidence interval (CI) refers to the mean classification accuracy. BoW = Bag-of-Words; TF-IDF = Term Frequency–Inverse Document Frequency; LR = Logistic Regression; DT = Decision Tree; MNB = Multinomial Naïve Bayes; RF = Random Forest; SGD = Stochastic Gradient Descent.

Regarding the performance of the ML classifiers on TF-IDF vectorized data, Table 3 shows that the LR model achieved the highest accuracy (0.9250 ± 0.0012), followed closely by SGD (0.9178 ± 0.0012) and RF (0.9146 ± 0.0018). As mentioned in Table 3, LR also demonstrated consistently strong performance across both thought classes. For the positive class, it achieved a precision of 0.9282 ± 0.0009, recall of 0.9365 ± 0.0018, and F1-score of 0.9323 ± 0.0011. For the negative class, the corresponding precision, recall, and F1-score were 0.9211 ± 0.0021, 0.9109 ± 0.0011, and 0.9160 ± 0.0013, respectively. The narrow confidence interval (0.9271, 0.9303) further indicates stable model performance across the five cross-validation folds.

On the other hand, for the BoW vectorized data, Table 3 reveals that LR also achieved the highest overall accuracy of 0.9287 ± 0.0013, closely followed by SGD at 0.9279 ± 0.0012 and RF at 0.9143 ± 0.0018. Although the difference between LR and SGD was minimal and fell within overlapping confidence intervals, SGD exhibited a slightly more balanced precision-recall profile for both the positive and negative classes. SGD showed only marginally higher positive recall (0.9411 vs. 0.9403) and negative precision (0.9265 vs. 0.9257). Since their 95% confidence intervals overlap, there is no strong statistical evidence that one model substantially outperformed the other. Therefore, it is more accurate to state that LR and SGD exhibited comparable performance under the BoW representation, with LR maintaining a slight overall advantage.

Among the transformer-based models, DistilBERT achieved the highest overall accuracy (0.8407 ± 0.0015; 95% CI: 0.8389–0.8425). For the positive thought class, it obtained a precision of 0.8587 ± 0.0031, recall of 0.8512 ± 0.0048, and F1-score of 0.8549 ± 0.0016. For the negative thought class, the corresponding precision, recall, and F1-score were 0.8190 ± 0.0041, 0.8278 ± 0.0052, and 0.8233 ± 0.0017, respectively. These results indicate that DistilBERT achieved consistently strong performance across both thought classes while maintaining balanced

precision and recall. These results indicate that DistilBERT achieved consistently strong performance across both thought classes while maintaining balanced precision and recall.

Table 4
Performance Analysis of the various transformer-based LLM models

Classifier	Thought	Precision	recall	F1-score	Accuracy	95% CI
BERT	Positive	0.8413 ± 0.0040	0.8519 ± 0.0060	0.8466 ± 0.0024	0.8297 ± 0.0024	0.8267, 0.8327
	Negative	0.8151 ± 0.0053	0.8024 ± 0.0068	0.8087 ± 0.0029		
DistilBERT	Positive	0.8587 ± 0.0031	0.8512 ± 0.0048	0.8549 ± 0.0016	0.8407 ± 0.0015	0.8389, 0.8425
	Negative	0.8190 ± 0.0041	0.8278 ± 0.0052	0.8233 ± 0.0017		
RoBERTa	Positive	0.6763 ± 0.0063	0.9402 ± 0.0040	0.7867 ± 0.0029	0.7188 ± 0.0060	0.7113, 0.7262
	Negative	0.8586 ± 0.0033	0.4465 ± 0.0180	0.5873 ± 0.0147		
ALBERT	Positive	0.8119 ± 0.0043	0.8392 ± 0.0068	0.8253 ± 0.0020	0.8041 ± 0.0019	0.8018, 0.8064
	Negative	0.7939 ± 0.0054	0.7609 ± 0.0084	0.7770 ± 0.0028		
Note: Precision, recall, F1-score, and accuracy are reported as mean ± SD over five stratified cross-validation folds.						

Notably, BERT also demonstrated competitive performance with an accuracy of 0.8297 ± 0.0024 , whereas ALBERT achieved 0.8041 ± 0.0019 . In contrast, RoBERTa produced the lowest accuracy (0.7188 ± 0.0060), primarily due to its substantially lower recall (0.4465 ± 0.0180) and F1-score (0.5873 ± 0.0147) for the negative thought class, despite achieving a high positive-class recall (0.9402 ± 0.0040).

Furthermore, as shown in Table 5, the Llama-2 model achieved a mean classification accuracy of $80.20\% \pm 2.35\%$ across the five cross-validation folds, with a 95% confidence interval of 78.14%, 82.26%. For the positive thought class, the model obtained a precision of 0.7920 ± 0.0536 , recall of 0.8788 ± 0.0440 , and F1-score of 0.8309 ± 0.0107 . For the negative thought class, the corresponding precision, recall, and F1-score were 0.8301 ± 0.0294 , 0.7074 ± 0.1039 , and 0.7589 ± 0.0477 , respectively. These results indicate that the prompt-based Llama-2 model demonstrated reasonable capability in distinguishing positive and negative thought patterns, although its overall performance remained lower than that of the fine-tuned encoder-based transformer models. This outcome is expected because Llama-2 is primarily designed as a generative large language model rather than a discriminative classifier. Consequently, it may require task-specific fine-tuning or more extensive prompt optimization to fully exploit its classification potential.

Table 5
Performance Analysis of Llama-2 model

Thought	Precision (mean ± SD)	Recall (mean ± SD)	F1-score (mean ± SD)	Accuracy (mean ± SD)	95% CI (Accuracy)
Positive	0.7920 ± 0.0536	0.8788 ± 0.0440	0.8309 ± 0.0107	0.8020 ± 0.0235	0.7814, 0.8226
Negative	0.8301 ± 0.0294	0.7074 ± 0.1039	0.7589 ± 0.0477		

Overall, the results (Tables 3 to 5) show that traditional ML models such as LR and SGD, outperform the BERT-based transformers and Llama-2. This can be attributed to several factors. First, the dataset consists of relatively short social media texts, where lexical features captured by TF-IDF and BoW were sufficient for binary classification. This allowed simpler models to be effectively generalized without the risk of overfitting (Black et al,2025; Graff et al, 2025).

Although transformers are powerful for capturing contextual information, they require extensive fine-tuning and larger datasets to achieve optimal performance. In this study, computational limitations restricted the training to only five epochs, which could have led to underfitting. Additionally, previous comparative studies have shown that traditional linear classifiers, particularly Logistic Regression with TF-IDF or Bag-of-Words representations, remain highly competitive for short-text sentiment classification when lexical polarity cues are strong, while offering lower computational complexity and greater interpretability than deep neural models (Diri et al, 2025). These models often outperform deep learning models, which may introduce unnecessary complexity when dealing with smaller or noisy datasets (Brown et al, 2024). This aligns with the findings (Table 3), that show the BoW approaches LR can achieve comparable or superior results in resource-constrained environments because of their efficiency and interpretability.

4.2 Error analysis

Figures 3 and 4 represent the normalized confusion matrices for the LR and the DistilBERT models, respectively, highlighting their class-wise prediction behavior and error distribution. Based on the confusion matrix (Figs. 3 and 4), a higher level of accuracy indicates a more effective identification of the positive class. This is a crucial aspect of the evaluation metrics employed in the various classification models. Furthermore, this implies that the model correctly identifies instances belonging to the positive class, which is a critical factor in determining the efficiency of the classification algorithm (Sokolova & Lapalme, 2009).

To explain the results, Fig. 3, which depicts the Logistic Regression model, shows that it accurately classified 91.45% of negative thoughts and 94.03% of positive thoughts, resulting in high recall values for both classes, as indicated earlier in Table 3. The misclassification rates were relatively low, with 8.55% of negative thoughts incorrectly predicted as positive (false positives) and 5.97% of positive thoughts misclassified as negative (false negatives). This balanced error distribution accounts for the LR's strong and stable precision-recall performance (Table 3), particularly when combined with Bag-of-Words features.

On the other hand, Fig. 4 illustrates the DistilBERT model, which shows a different distribution owing to its normalization over the full dataset. DistilBERT correctly identified 37.13% of all samples as true negatives and 46.94% as true positives, yielding a combined accuracy of approximately 84%, which is again consistent with the accuracy values reported in Table 4. Misclassifications accounted for 7.73% of false positives and 8.21% of false negatives, indicating a slightly higher confusion between classes compared with LR. However, the error rates are relatively symmetrical, Table 4 suggesting that DistilBERT does not exhibit strong class bias, which aligns with its balanced precision and recall scores across both the positive and negative classes.

Overall, this comparison (Figs. 3 and 4) indicates that while Logistic Regression achieves a higher absolute classification accuracy using simpler feature representations, DistilBERT demonstrates robust and balanced class discrimination by leveraging contextualized representations. The analyses of these confusion matrices (Figs. 3 and 4) reinforce the quantitative results presented in Tables 3 and 4 and clarify the sources of classification errors for both traditional and transformer-based models.

Finally, Fig. 5 presents the training and validation accuracy and loss curves for DistilBERT during the first fold of the 5-fold stratified cross-validation. As shown in Fig. 5, the model converges rapidly within the first two epochs, with training accuracy increasing from approximately 0.62 to 0.78, followed by gradual improvements until epoch five. The validation accuracy consistently remained higher than the training accuracy across epochs, indicating strong generalization behavior rather than overfitting.

As illustrated in Fig. 5, the loss curves exhibit a smooth and monotonic decline for both the training and validation sets, with the validation loss remaining lower than the training loss throughout the optimization. This stable gap suggests effective regularization and well-behaved fine-tuning dynamics. The absence of divergence between the curves implies that the model does not memorize training samples excessively, even when trained for multiple epochs.

Importantly, this training behavior was consistent across all five folds. The final validation accuracies ranged narrowly between 0.8390 and 0.8427, with losses between 0.3714 and 0.3784, demonstrating low variance across folds. Precision, recall, and F1-score also showed minimal variation, with a mean positive-class F1-score of 0.8549 ± 0.0016 and a negative-class F1-score of 0.8233 ± 0.0017 . These results indicate that most misclassifications are systematic rather than fold-specific, and that DistilBERT learns stable and transferable contextual representations across different data partitions.

5. Discussion

This study demonstrates that the sentiment-derived polarity of self-expressed thought statements or languages in social media text can be classified using conventional machine learning, transformer encoders, and a prompt-based Llama-2 model (see Results Section). Importantly, the classification target should not be interpreted as a direct measurement of cognition or psychological disorder. Rather, positive and negative polarity represent operational labels derived initially from the evaluative valence of the textual expressions and subsequently assessed through expert evaluation. The findings therefore concern computational classification of the polarity expressed in thought-related social media text, rather than clinical identification of positive or negative cognitive states. Unlike many previous studies that focus on a single modeling approach, this work provides a comprehensive comparison across traditional machine learning algorithms, encoder-based transformers, and a decoder-only large language model under a consistent evaluation framework. The key findings and outcomes of this study are discussed in triangulation with existing literature in this section.

5.1 Key Findings

A key contribution of this work is the use of VADER-assisted automatic annotation combined with expert validation (see Methodology). Research has shown that automatic sentiment labeling enables efficient construction of large-scale datasets for social media analysis, although concerns remain regarding annotation quality, particularly for ambiguous, sarcastic, or context-dependent expressions (Hutto & Gilbert, 2014; Jain et al., 2023; Oliveira et al., 2023). In this study, to assess label reliability, two psychology experts independently reviewed a stratified sample of 500 statements using a five-point agreement scale. The majority of statements received high agreement ratings (scores of 4 or 5), with the first expert agreeing with 80.8% of the negative samples and 84.4% of the positive samples, while the second expert reported agreement rates of 70.0% and 64.8%, respectively. These findings indicate that a substantial proportion of the VADER-derived polarity labels were consistent with expert assessments of the expressed valence of the statements. The expert evaluation provides additional evidence concerning the reliability of the polarity annotations used as classification targets, while the distinction between linguistic sentiment and underlying cognition remains important.

Compared with previous studies, the proposed framework (Fig. 1) offers a scalable approach to thought polarity classification. Bathina et al. (2021) demonstrated that individuals with depression exhibit more distorted thinking in social media posts, highlighting the value of analysing online thought patterns. However, their work focused on cognitive distortions associated with depression rather than developing a general computational framework for positive and negative thought classification. In contrast, this study constructs a large automatically labelled dataset and evaluates multiple classification models using five-fold cross-validation, confidence intervals, and comprehensive performance metrics.

Furthermore, among all evaluated models, conventional machine learning methods, particularly Logistic Regression, achieved the highest classification performance (Table 3). This suggests that short social media thoughts often contain explicit lexical indicators of positive or negative sentiment that can be effectively captured using Bag-of-Words and TF-IDF representations (Table 3). Although transformer models provide richer contextual representations, they did not outperform the simpler linear classifiers in this binary classification task. Table 4 shows that DistilBERT produced the strongest performance among the encoder-based transformer models, demonstrating stable results across the five validation folds, but remained below the best-performing conventional machine learning models. These findings (Table 3, 4) indicate that increased model complexity does not necessarily translate into improved performance when discriminative lexical features dominate the classification task.

The prompt-based Llama-2 model achieved a mean classification accuracy of $79.10\% \pm 3.21\%$, with positive and negative class F1-scores of 0.7913 ± 0.0565 and 0.7852 ± 0.0213 , respectively. Although these results demonstrate reasonable classification capability, Llama-2 underperformed compared with both conventional machine learning and fine-tuned encoder-based transformers. This outcome (Table 5) is expected because Llama-2 is primarily designed as a generative language model rather than a supervised discriminative classifier. Without task-specific fine-tuning, its predictions rely heavily on prompt design and in-context learning, making classification performance more variable than that of dedicated encoder models.

The findings are consistent with recent advances in computational mental health. Sharma et al. (2023) demonstrated the potential of large language models for cognitive reframing, while Lin et al. (2024) developed a framework for detecting and reconstructing cognitively distorted sentences. However, these studies primarily focused on cognitive distortion identification and reframing tasks using relatively smaller datasets. In contrast, the present study focuses on large-scale thought polarity classification and provides a systematic comparison between conventional machine learning, transformer encoders, and a decoder-only LLM using a unified evaluation protocol. Furthermore, Wang et al. (2023) introduced the C2D2 dataset for cognitive distortion analysis, whereas our work addresses a broader binary thought polarity classification problem using substantially larger social media data, making it more suitable for scalable computational analysis.

Overall, the results demonstrate that carefully designed conventional machine learning methods remain highly competitive for large-scale thought classification despite the rapid development of transformer models and large language models. The combination of automatic annotation, expert-supported validation, rigorous five-fold cross-validation, and comprehensive benchmarking provides a practical framework for computational analysis of thought polarity. Nevertheless, this work should be interpreted as computational NLP research rather than a clinical diagnostic system. Social media expressions may reflect temporary emotions, contextual experiences, or individual communication styles rather than persistent psychological conditions. Future research should incorporate multi-expert annotation with formal inter-rater reliability analysis, multilingual and cross-cultural datasets, clinical benchmark corpora, and parameter-efficient fine-tuning of large language models to further improve the robustness and generalisability of automated thought classification.

5.2 Theoretical and Practical Implications of this Study

The findings of this study have important theoretical and practical implications for computational mental health analytics and natural language processing. Theoretically, this study is grounded in cognitive theories of emotion and psychopathology. Beck's cognitive theory posits that emotional distress arises from biased, negatively valenced interpretations of experience, organised around underlying cognitive schemas (Beck, 1979; Beck & Haigh, 2014). Similarly, Ellis's ABC model of rational emotive behaviour therapy holds that it is not events

themselves but the beliefs individuals hold about them that determine emotional consequences (Ellis, 1962). Both theories treat verbalised thoughts as observable expressions of underlying cognition, which provides the theoretical justification for using language as an empirical window into thought patterns (Tausczik & Pennebaker, 2010). The present framework operationalises these theoretical premises computationally: classifying the polarity of self-expressed thoughts at scale constitutes a computational analogue of the thought-identification stage of cognitive restructuring in cognitive behavioural therapy (Sharma et al., 2023).

Within this theoretical framing, it is important to distinguish between sentiment analysis and the computational analysis of thought-related language. Sentiment analysis estimates the evaluative valence expressed in text, whereas a thought encompasses broader cognitive content, including beliefs, interpretations, expectations, and appraisals (VandenBos, 2007; Hutto & Gilbert, 2014). Consequently, the present results should not be interpreted as demonstrating that sentiment polarity is equivalent to positive or negative cognition. Rather, this study treats sentiment-derived valence as an operational proxy for one observable linguistic dimension of self-expressed thoughts. This distinction is particularly important in computational mental health research because high predictive performance against sentiment-derived labels demonstrates reliable classification of the operational target but does not, by itself, establish the psychological construct validity of underlying thought patterns. Future research should therefore develop independently annotated cognitive categories and examine their relationship with sentiment polarity to determine the extent to which linguistic valence corresponds to broader patterns of cognition.

Within this defined scope, the study contributes to computational analysis of thought-related language by demonstrating that the polarity of self-expressed statements can be analysed at scale through sentiment-assisted annotation, expert evaluation, and comparative machine learning. Unlike many previous studies that focus primarily on depression detection (Bathina et al., 2021; Kabir et al., 2023), cognitive distortion identification (Wang et al., 2023; Lin et al., 2024), or emotion classification (Rahman et al., 2024), the present study systematically compares conventional machine learning, transformer-based models, and a decoder-only large language model for classifying the expressed polarity of thought-related social media text within a unified evaluation framework (see Sections 3 and 4).

A key contribution is the comparison of conventional machine learning, encoder-based transformers, and decoder-only large language models (See Tables 3–5). Within the scope of this study, higher model complexity was not consistently associated with improved classification performance, indicating that simpler models may be sufficient when lexical polarity cues are prominent (Table 3). This is consistent with broader evidence that classical and deep learning approaches each retain distinct advantages across mental health text-classification tasks (Zhang et al., 2022). The findings emphasise the importance of selecting models based on the characteristics of the data and the classification task, supported by robust evaluation methods including five-fold cross-validation and confidence intervals.

From a cognitive-theoretical perspective, the polarity of self-referential language may provide a useful linguistic signal for investigating how positive and negative evaluations are expressed in thought-related text. Within cognitive theories such as Beck's cognitive model, negatively valenced self-appraisals are relevant to understanding maladaptive patterns of thinking (Beck & Haigh, 2014). However, linguistic polarity alone cannot establish the presence of a cognitive distortion or psychological disorder. Practically, the framework provides an efficient method for analysing such positive and negative thought patterns in large social media datasets (Fig. 1). The strong performance of Logistic Regression (Table 3) indicates that high classification accuracy is achievable without complex models, making the approach suitable for large-scale monitoring. The validated annotated dataset further supports future research in sentiment analysis and computational psychology.

Different model families may be suited to different stages of a computational thought-analysis framework. Conventional machine learning is effective for classification, while transformer models enhance contextual understanding. Although the decoder-only Llama-2 model was less effective for classification in this study (Table 5), previous research has demonstrated the potential of generative language models for tasks such as cognitive reframing and guided generation (Sharma et al., 2023; Xiao et al., 2024). These findings suggest a possible direction for future research in which generative LLMs complement discriminative classifiers; however, such applications are outside the scope of the present study (Section 3).

Lastly, from a policymaking and digital-platform governance perspective, the framework (Fig. 1) highlights both the opportunities and limitations of automated sentiment analysis of thought-related language in the context of mental health and cognitive research. At an aggregate level, such analysis may support researchers and organisations in examining patterns in online discourse and identifying changes in positively and negatively valenced expressions across large datasets. Previous studies have demonstrated the potential of computational language models for analysing mental-health-related indicators in online text at scale (Xu et al., 2024; Zhang et al., 2022), supporting the broader relevance of such approaches for large-scale research. However, because the classification target in the present study represents sentiment-derived linguistic valence rather than independently validated cognitive or diagnostic categories, the outputs should not be used to infer an individual's psychological condition, identify a cognitive disorder, assign clinical risk, or support automated decisions about individual users. Policies governing such systems should therefore distinguish population-level linguistic analysis from individual-level psychological or clinical assessment and promote transparency regarding label generation, prediction uncertainty, privacy, and the

limitations of sentiment-derived proxies. Where computational outputs could affect individuals or inform health-related decisions, appropriate human oversight, ethical safeguards, independently validated psychological constructs, and relevant professional expertise should remain essential.

6 Limitations and Future Directions

6.1 Limitations

This study has several limitations. First, the dataset was derived from publicly available social media texts, which may not accurately reflect the clinical presentations of cognitive or behavioral disorders. Online expressions of negative thought patterns are context-dependent, vary over time, and are influenced by cultural and platform-specific factors. Therefore, these findings should not be interpreted to support any clinical diagnosis.

Second, constructing a large annotated dataset for supervised learning is inherently challenging because manual annotation is time-consuming and resource-intensive, particularly for large-scale NLP applications. To enable scalable dataset creation, this study employed the VADER sentiment analyzer to automatically label sentiment polarity and exclude neutral thought data (Hutto & Gilbert, 2014). Although prior studies have demonstrated VADER's effectiveness for short, informal texts, automatic annotation may introduce residual label noise, particularly in ambiguous or sarcastic expressions (Jain et al., 2023; Oliveira et al., 2023). Consequently, the dataset is intended for the computational analysis of positive and negative thought patterns rather than clinical diagnosis, and any remaining label noise is acknowledged as a limitation.

Third, the dataset consists exclusively of English-language content collected from a limited set of social media platforms, which restricts the generalizability of the findings to other languages, cultures, and communication contexts. Future research should explore multilingual and cross-cultural validation.

Finally, although stratified cross-validation was applied to the selected models, computational constraints limited exhaustive hyperparameter optimization and prompt-level exploration for the Llama-2 model. Accordingly, the reported performance metrics should be interpreted as indicative rather than definitive.

6.2 Future Directions

Future research should build upon the proposed classification framework by exploring automated thought conversion and cognitive reframing using fine-tuned large language models. Additionally, human-centered evaluation protocols should be employed to assess the quality, appropriateness, and potential psychological benefits of the generated responses. Moreover, future studies should validate the framework with clinically annotated datasets and multimodal mental health data, such as clinical narratives, speech, and behavioral signals, to examine its generalizability and potential applicability in AI-assisted mental health research.

6.3 Ethical Considerations

The proposed framework is intended solely as a research tool for analyzing publicly available thought-related social media texts and their sentiment polarity. The classification results should not be viewed as indicators of an individual's mental health status or used as a substitute for professional psychological assessment. All data utilized in this study were collected from publicly accessible platforms and anonymized during preprocessing to protect user privacy and confidentiality.

7 Conclusions

This study demonstrates that text classification and controlled text generation using NLP and ML techniques are effective tools for analyzing positive and negative thought-related large-scale social media texts. By systematically comparing traditional ML classifiers and transformer-based models, the study showed that lightweight linear models such as LR and SGD perform competitively when combined with BoW and TF-IDF representations, whereas DistilBERT achieves the most stable and accurate performance among transformer-based approaches. In contrast, the applied Llama-2 model exhibited a moderate classification performance under instruction-based prompting, reflecting its primary optimization for generative rather than discriminative tasks. The curated dataset and classification framework are intended to facilitate the computational analysis of thought-related sentiment polarity. The results highlight the importance of model simplicity, feature representation, and evaluation rigor when analyzing noisy real-world texts or data. Expert validation and cross-validation strengthen confidence in the reported findings. Overall, this work contributes a validated dataset and a comparative evaluation of classification models, that supports further research at the intersection of NLP and mental-health related text analysis, while maintaining clear ethical and clinical boundaries. Future work should focus on broader validation, multilingual and cross-cultural analysis, structured human-in-the-loop evaluations, and safety-aware deployment frameworks.

Declarations

Acknowledgement

This work received partial support from the ICT fellowship of the Information and Communication Technology (ICT) Division, Ministry of Posts, Telecommunications and IT, Government of the People's Republic of Bangladesh.

We gratefully acknowledge Dr. Mohammad Afzal Hossain and Md. Shahinoor Rahman, PhD, Professors, Department of Psychology, University of Chittagong, for conducting expert validation of a randomly sampled subset of the dataset. His insights from cognitive behavioral therapy and mental health assessment were invaluable in ensuring the reliability of our thought pattern annotations.

Ethical Approval and accordance

Ethical approval for this study was obtained from the Research, Extension, Advisory Services and Publications (REASP) Committee of the Islamic University of Technology (IUT), Bangladesh. The committee reviewed the data sources and methodology and confirmed that the research used publicly accessible online data and complied with applicable ethical guidelines.

Consent to participate

Not applicable. The study used publicly available online data and did not involve direct recruitment or interaction with the individuals whose social media texts were analysed. The REASP Committee of the Islamic University of Technology (IUT) confirmed that no additional consent was required for the publicly available data used in this research.

Data Availability Statement

The dataset generated and analysed during the current study is publicly available in the Mendeley Data repository (Niger et al., 2024) at <https://doi.org/10.17632/9xd46s6hvjv.3>

References

1. VandenBos, G. R. (2007). *APA dictionary of psychology*. American Psychological Association.
2. Gilbert, P. (1998). The evolved basis and adaptive functions of cognitive distortions. *British Journal of Medical Psychology*, 71(4), 447–463.
3. Moulds, M. L., & McEvoy, P. M. (2025). Repetitive negative thinking as a transdiagnostic cognitive process. *Nature Reviews Psychology*, 4(2), 127–141.
4. Beck, A. T. (1979). *Cognitive therapy and the emotional disorders*. Penguin.
5. Ehret, A. M., Joormann, J., & Berking, M. (2015). Examining risk and resilience factors for depression: The role of self-criticism and self-compassion. *Cognition and Emotion*, 29(8), 1496–1504.
6. Pedro, L., Branquinho, M., Canavarro, M. C., & Fonseca, A. (2019). Self-criticism, negative automatic thoughts and postpartum depressive symptoms: The buffering effect of self-compassion. *Journal of reproductive and infant psychology*, 37(5), 539–553. [oclc.org/10.1080/02646838.2019.1597969](https://doi.org/10.1080/02646838.2019.1597969). <https://doi-org.surrey.idm>.
7. Spinhoven, P., van Hemert, A. M., & Penninx, B. W. (2018). Repetitive negative thinking as a predictor of depression and anxiety: A longitudinal cohort study. *Journal of Affective Disorders*, 241, 216–225.
8. Lalk, C., Steinbrenner, T., Pena, J. S., Kania, W., Schaffrath, J., Eberhardt, S., & Rubel, J. (2025). Depression symptoms are associated with frequency of cognitive distortions in psychotherapy transcripts. *Cognitive Therapy and Research*, 49(3), 588–600.
9. Simms, T., Ramstedt, C., Rich, M., Richards, M., Martinez, T., & Giraud-Carrier, C. (2017, August). Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE international conference on healthcare informatics (ICHI)* (pp. 508–512). IEEE.
10. Rutter, A., Edinger, L., Lorenzo-Luaces, A., Ten Thij, L., Valdez, M., D., & Bollen, J. (2025). Anxiety and depression are associated with more distorted thinking on social media: A longitudinal multi-method study. *Cognitive therapy and research*, 49(4), 712–720.
11. Shickel, B., Siegel, S., Heesacker, M., Benton, S., & Rashidi, P. (2020, October). Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 275–280). IEEE.
12. Tauscher, J. S., Lybarger, K., Ding, X., Chander, A., Hudenko, W. J., Cohen, T., & Ben-Zeev, D. (2023). Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. *Psychiatric services*, 74(4), 407–410.

13. Wiemer-Hastings, K., Janit, A. S., Wiemer-Hastings, P. M., Cromer, S., & Kinser, J. (2004). Automatic classification of dysfunctional thoughts: a feasibility test. *Behavior Research Methods Instruments & Computers*, 36, 203–212.
14. Persons, J. B., & Burns, D. D. (1985). Mechanisms of action of cognitive therapy: The relative contributions of technical and interpersonal interventions. *Cognitive Therapy & Research*, 9, 539–551.
15. World Health Organization (2021, November 17). Mental health of adolescents. Retrieved January 20, 2024, from <https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health>
16. Alavi, M., Hungerford, C., Ghelbash, Z., Le Lagadec, D., & Cleary, M. (2024). Promoting high-quality research to improve global mental health and well-being. *Issues in Mental Health Nursing*, 1–2. <https://doi.org/10.1080/01612840.2024.2309524>
17. Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: from text to predictions. *Information*, 13(2), 83. <https://doi.org/10.3390/info13020083>
18. Jurasky, D., & Martin, J. H. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Computational Linguistics*, 26(4), 638–641. <https://doi.org/10.1162/089120100750105975>
19. Razno, M. (2019). Machine learning text classification model with NLP approach. *Computational Linguistics and Intelligent Systems*, 2, 71–73.
20. Dehghan, P., Alashwal, H., & Moustafa, A. A. (2022). Applications of machine learning to behavioral sciences: focus on categorical data. *Discover Psychology*, 2(1), 22.
21. Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., et al. (2023). A survey on evaluation of large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2307.03109>
22. Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text Classification via Large Language Models. *arXiv preprint arXiv:2305.08377*.
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
24. Ruthotto, L., & Haber, E. (2021). An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2), e202100008.
25. Anish, K. R., & George, P. G. (2024). Community Mental Health: Evolution, Current Status, and Future Perspectives. In *Mental Health Care Resource Book* (pp. 101–114). https://doi.org/10.1007/978-981-97-1203-8_6
26. Kilbourne, A. M., Beck, K., Spaeth-Rublee, B., Ramanuj, P., O'Brien, R. W., Tomoyasu, N., & Pincus, H. A. (2018). Measuring and improving the quality of mental health care: a global perspective. *World Psychiatry*, 17(1), 30–38.
27. Vandoni, M., D'Avanzo, B., & Barbato, A. (2024). The Transition towards Community-based Mental Health Care in the European Union: Current Realities and Prospects. *Health Policy*. <https://doi.org/10.1016/j.healthpol.2024.105081>
28. Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., & Hasan, K. (2023). DEPTWEET: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139, 107503.
29. Rahman, A. B. S., Ta, H. T., Najjar, L., Azadmanesh, A., & Gönlül, A. S. (2024). DepressionEmo: A novel dataset for multilabel classification of depression emotions. *arXiv preprint arXiv:2401.04655*.
30. Tice, D. M., Bratslavsky, E., & Baumeister, R. F. (2018). Emotional distress regulation takes precedence over impulse control: If you feel bad, do it! *Self-regulation and self-control* (pp. 267–298). Routledge.
31. Garg, M. (2022). An event detection technique using social media data. *arXiv*. /abs/2208.13101
32. Goswami, A., & Kumar, A. (2016). A survey of event detection techniques in online social networks. *Social Network Analysis and Mining*, 6, 1–25.
33. Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine*, 3(1), 43.
34. Umair, A., Nanda, P., & He, X. (2017, August). Online social network information forensics: A survey on use of various tools and determining how cautious Facebook users are? *2017 IEEE Trustcom/BigDataSE/ICCESS* (pp. 1139–1144). IEEE.
35. Raslescu, A., Kreicker, S., Gillespie, A., Berners-Lee, W., & Murphy, S. E. (2022). UK dataset of valence, imageability and frequency ratings of 300 adjectives for use in cognitive-emotional tasks. *Mendeley Data*, V1. <https://doi.org/10.17632/kgk3jbx9xb.1>
36. Joshi, M. L., & Kanoongo, N. (2022). Depression detection using emotional artificial intelligence and machine learning: A closer review. *Materials Today: Proceedings*, 58, 217–226.
37. Nova, K. (2023). Machine learning approaches for automated mental disorder classification based on social media textual data. *Contemporary Issues in Behavioral and Social Sciences*, 7(1), 70–83.
38. Rojas-Barahona, L., Tseng, B. H., Dai, Y., Mansfield, C., Ramadan, O., Ultes, S., & Gasic, M. (2018). Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. *arXiv preprint arXiv:1809.00640*.

39. Niger, T., Rayhan, H., Islam, R., Noor, K. A. A., & Hasan, K. (2022). Framework for behavioral disorder detection using machine learning and application of virtual cognitive behavioral therapy in COVID-19 pandemic. *Minerva Psychiatry*, 63(4), 284–294. <https://doi.org/10.23736/S2724-6612.22.02322-3>
40. Bathina, K. C., Thij, T., Lorenzo-Luaces, M., Rutter, L., L. A., & Bollen, J. (2021). Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, 5(4), 458–466.
41. Beck, A. T., Rush, A. J., Shaw, B. F., Emery, G., DeRubeis, R. J., & Hollon, S. D. (2024). *Cognitive therapy of depression*. Guilford.
42. Burns, D. D., & Good, M. F. (1980). *The new mood therapy*. New York, William Morrow.
43. Ellis, A. (1962). Reason and emotion in psychotherapy.
44. Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1), 24–54.
45. Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1), 46.
46. Sharma, A., Rushton, K., Lin, I. W., Nguyen, T., & Althoff, T. (2023). Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. *arXiv preprint arXiv:2310.15461*.
47. Sharma, A., Rushton, K., Lin, I. W., Wadden, D., Lucas, K. G., Miner, A. S., & Althoff, T. (2023). Cognitive reframing of negative thoughts through human-language model interaction. *arXiv preprint arXiv:2305.02466*.
48. Zeng, X., & Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1), 1–12.
49. Boe, B. (2024). (n.d.). PRAW 7.7.1 documentation. Retrieved May 6, from <https://praw.readthedocs.io/en/stable/>
50. Das, N., Gupta, S., Das, S., Yadav, S., Subramanian, T., & Sarkar, N. (2021, September). A Comparative Study of Sentiment Analysis Tools. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES)* (pp. 1–7). IEEE.
51. Bharathi, S. T., Raj, J. R. F., Krishnan, R. S., Rajalakshmi, V., Kumar, V. V., & Anchitaalagammai, J. V. (2024, December). MentalHealthAI: Utilizing Deep Learning for Accurate Stress Level Prediction. In *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 1705–1712). IEEE.
52. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216–225).
53. Raslescu, A., Kreicker, S., Gillespie, A. L., Berners-Lee, W., Murphy, S. E., & Harmer, C. J. (2023). Positive and negative personality descriptors: UK dataset of self-referential valence, imageability and subjective frequency ratings of 300 adjectives for use in cognitive-emotional tasks. *Data in Brief*, 46, 108831.
54. Grün, D. (2016). An English Word Database of EMOtional TErms (EMOTE). *Psychological Reports*, 119(1), 290–308. <https://doi.org/10.1177/0033294116658474>
55. Reshi, A. A., Rustam, F., Aljedaani, W., Shafi, S., Alhossan, A., Alrabiah, Z., & Ashraf, I. (2022, February). Covid-19 vaccination-related sentiments analysis: a case study using worldwide twitter dataset. *Healthcare* (Vol. 10, p. 411). MDPI.
56. Jain, R., Kumar, A., Nayyar, A., et al. (2023). Explaining sentiment analysis results on social media texts through visualization. *Multimed Tools Appl*, 82, 22613–22629. <https://doi.org/10.1007/s11042-023-14432-y>
57. Silva, D., Oliveira, da Silva, R. M. A., Lima, P. V., & Xavier, S. F. A. (2024). Reddit Comment Analysis: Sentiment Prediction and Topic Modeling using VADER and BERTopic. *Socioeconomic Analytics*, 2(1), 130–136. <https://doi.org/10.51359/2965-4661.2024.265074>
58. Muller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists*. O'Reilly Media, Inc.
59. De Casterlé, B. D., Gastmans, C., Bryon, E., & Denier, Y. (2012). QUAGOL: A guide for qualitative data analysis. *International journal of nursing studies*, 49(3), 360–371.
60. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. Available at: <https://www.nltk.org/book/>
61. Khyani, D., Siddhartha, B. S., Niveditha, N. M., & Divya, B. M. (2021). An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology*, 22(10), 350–357.
62. Lane, H., Howard, C., & Hapke, H. (2019). *Natural language processing in action*. Pearson Professional.
63. Kozhevnikov, V. A., & Pankratova, E. S. (2020). Research of the text data vectorization and classification algorithms of machine learning. *ISJ Theoretical & Applied Science*, 5(85), 574–585.
64. Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June). An overview of bag of words; importance, implementation, applications, and challenges. In *2019 International Engineering Conference (IEC)* (pp. 200–204). IEEE.

65. Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.
66. Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic Emergency Medicine*, 18(10), 1099–1104.
67. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2), 130.
68. Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS International Journal of Computer Science and Network Security*, 19(3), 62.
69. Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31.
70. Newton, D., Yousefian, F., & Pasupathy, R. (2018). Stochastic gradient descent: Recent trends. In *Recent Advances in Optimization and Modeling of Contemporary Problems* (pp. 193–220).
71. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *NIPS*.
72. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., & Brew, J. (2019). Transformers: State-of-the-art natural language processing.
73. Tsai, H., Riesa, J., Johnson, M., Arivazhagan, N., Li, X., & Archer, A. (2019). Small and practical BERT models for sequence labeling. In *EMNLP-IJCNLP*.
74. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
75. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
76. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
77. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
78. Agarap, A. F. (2018). Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375*.
79. Pratiwi, H., Windarto, A. P., Susliansyah, S., Aria, R. R., Susilowati, S., Rahayu, L. K., & Rahadjeng, I. R. (2020, February). Sigmoid activation function in selecting the best model of artificial neural networks. In *Journal of Physics: Conference Series* (Vol. 1471, No. 1, p. 012010). IOP Publishing.
80. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T. (2023). Llama: Open and efficient foundation language models.
81. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55, 1–35. <https://doi.org/10.1145/3560815>
82. Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 8086–8098). Dublin, Ireland: Association for Computational Linguistics.
83. Diri, G., Obiorah, P., & Du, H. (2025, July). Comparative Study of Sentiment Analysis Techniques: Traditional Machine Learning vs. Deep Learning Approaches. In *InSITE 2025: Informing Science + IT Education Conferences: Hiroshima* (p. 23).
84. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901).
85. Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). *Calibrate before use*. Improving few-shot performance of language models.
86. Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 human language technology conference of the North American Chapter of the Association for Computational Linguistics* (pp. 213–220).
87. Seershika, M., Aishwarya, V., Vani, V., & Karthik, N. (2024, July). Identifying The Established Semantic Kinds In Text Using Named Entity Recognition. In *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IconSCEPT)* (pp. 1–6). IEEE.
88. Zhou, Y. (2024). Analysis of Thematic Context in English Reading Teaching. *Scientific Journal of Technology*, 6(7), 125–129. <https://doi.org/10.54691/xmnwg891>
89. Qu, S. (2024). A Thematic Analysis of English and American Literature Works Based on Text Mining and Sentiment Analysis. *Journal of Electrical Systems*. <https://doi.org/10.52783/jes.3076>

90. Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. *Natural language processing using very large corpora* (pp. 157–176). Springer Netherlands.
91. Brown, K. E., Yan, C., Li, Z., Zhang, X., Collins, B. X., Chen, Y., & Malin, B. A. (2024). Not the models you are looking for: Traditional ML outperforms LLMs in clinical prediction tasks. *medRxiv*.
92. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
93. Wang, B., Deng, P., Zhao, Y., & Qin, B. (2023, December). C2D2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10149–10160).
94. Lin, S., Wang, Y., Dong, J., & Ni, S. (2024, August). Detection and positive reconstruction of cognitive distortion sentences: Mandarin dataset and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 6686–6701).
95. Xiao, M., Xie, Q., Kuang, Z., Liu, Z., Yang, K., Peng, M., & Huang, J. (2024, August). HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1707–1725).
96. Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., & Wang, D. (2024). Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(1), 1–32.
97. Niger, T., Hasan, D. K., & Mahmud, H. (2024). Dataset to Identify Behavioral Disorder, *Mendeley Data*, V3, doi: 10.17632/9xd46s6hvjv.3.

Figures

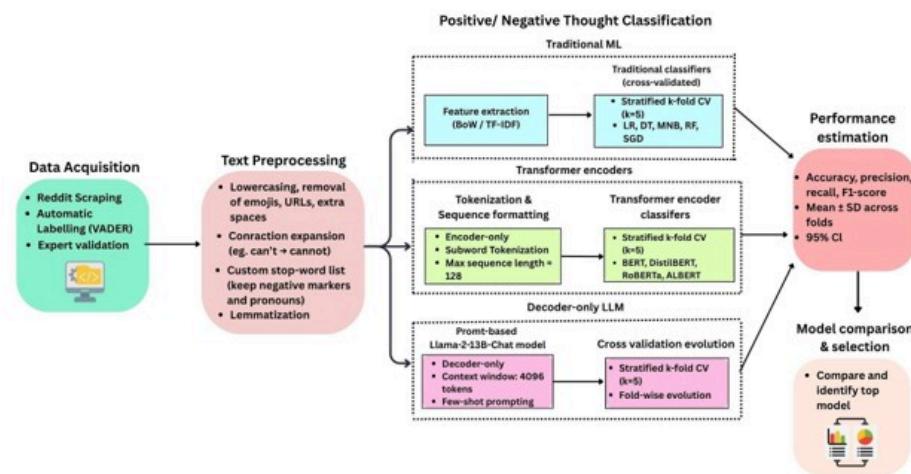


Figure 1

Architecture of the thought classification framework and methodology.

Social media texts are labeled with VADER and preprocessed, then passed to three branches: traditional ML classifiers with stratified 5-fold cross-validation, transformer-based encoder models, and a prompt-based Llama-2-13B-Chat classifier. All models are evaluated using stratified five-fold cross-validation, with performance reported as accuracy, precision, recall, F1-score, mean $\pm$ SD, and 95% confidence intervals before comparative model selection.

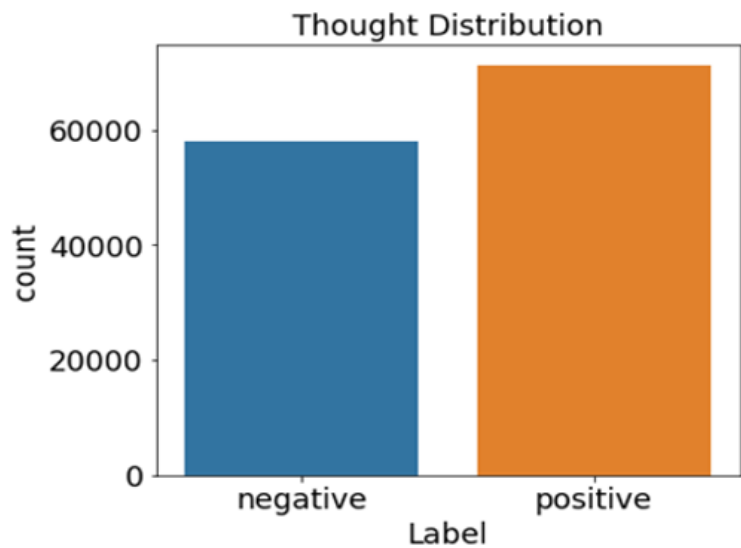


Figure 2

Positive and Negative thought distribution

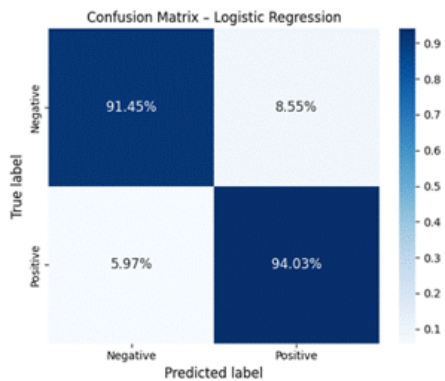


Figure 3

Normalized confusion matrix for the Logistic Regression (LR) classifier using Bag-of-Words features. Percentages indicate the proportion of true negative and true positive predictions

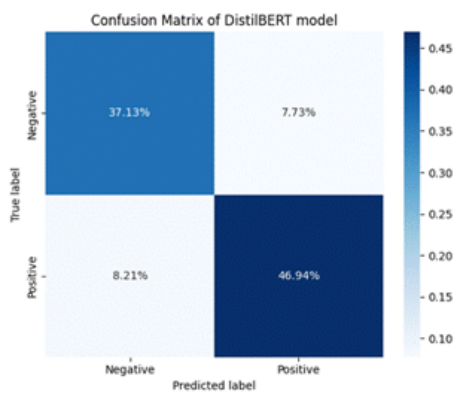


Figure 4

Normalized confusion matrix for the DistilBERT classifier. Percentages represent the proportion of samples correctly and incorrectly classified across positive and negative thought classes

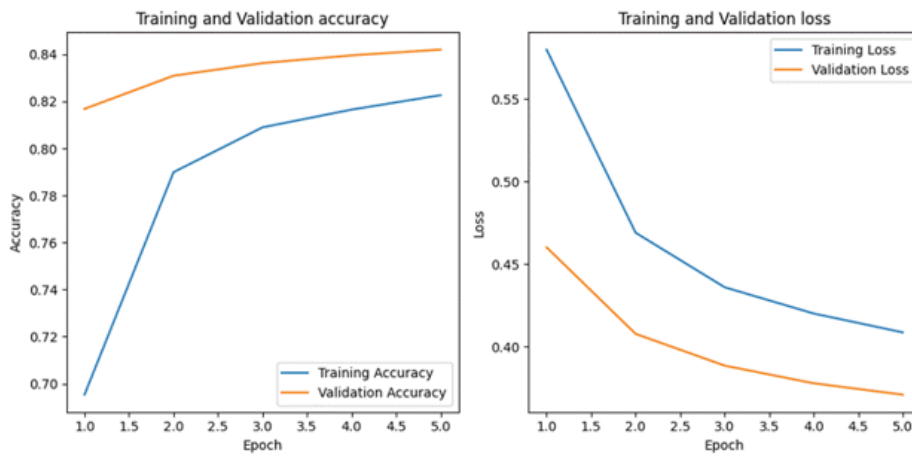


Figure 5

Training and validation accuracy and loss curves for DistilBERT during Fold 1 of 5-fold cross-validation, illustrating rapid convergence and stable generalisation behavior.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Appendix.docx](#)