Data-Driven or Dataless? Detecting Indicators of Mental Health Difficulties and Negative Life Events in Financial Resilience Using Prompt-Based Learning

Xia Cui*†‡, Terry Hanley†, Muj Choudhury‡, Tingting Mu†
*Manchester Metropolitan University, Manchester, United Kingdom
Email: x.cui@mmu.ac.uk
†The University of Manchester, Manchester, United Kingdom
Email: {terry.hanley, tingting.mu}@manchester.ac.uk
‡VoiceIQ Ltd., Manchester, United Kingdom
Email: muj@voiceiq.ai

***Abstract*—Financial resilience has been an important area of focus for the business sector since the outbreak of the pandemic. Currently, the assessment of financial resilience is typically completed through the review of financial statements. However, such resilience is commonly linked to negative life events and may be further impacted by the presence of mental health difficulties. As such, identifying and understanding these elements may provide a more complete understanding of an individual's financial situation. We discuss the development of a challenging automated financial resilience detection system that aims to identify factors that may have negative impacts upon the resilience of individuals. This makes use of textual data to identify elements, such as the occurrence of negative life events or mental health difficulties, that indicate that individuals may be vulnerable to exploitation through the misselling of products. In addition to a traditional data-driven supervised approach, this work also demonstrates applying prompt-based learning to these tasks without the need for the training data (i.e., a dataless approach).**

***Index Terms*—financial resilience, mental health, natural language processing, prompt-based learning, transfer learning**

I. Introduction

Since the financial crisis in 2008, and with the additional effects of the COVID-19 pandemic, financial stress has expanded to many individuals around the world. Such difficulties can extend beyond the financial world, and many individuals have seen their mental health and well-being impacted as a consequence of their stress. The World Health Organisation (WHO) [1] reported that almost 1 billion people have suffered from mental, neurological and substance use disorders worldwide. The pandemic and financial stress appear to have had a significant impact on the mental health of large numbers of individuals. In addition to the pandemic, commonplace negative life events are consistently associated with the onset of mental health difficulties. Under major crises and outbreaks of diseases, more people are experiencing life events such as losing their jobs and reduced hours at work, loss of family, heavy caring responsibilities, and relationship breakdowns. In the context of the financial industry, resilience is investigated through a financial statement. First, not all organisations have access to these statements. Second, these statements do not contain information about life events or mental health difficulties. Since the pandemic, more people have started to seek advice online, and several forms of online communication are text-based, which leads many organisations to a shortage of domain experts to access and investigate massive amounts of textual information. There is a great demand to develop an automated text-based system to help identify people at risk of financial resilience. Typically, developing an automated system requires a large amount of data to train a Machine Learning (ML) model to perform the identification task. Several prior works in detecting mental health difficulties [2], [3] trained a supervised model using social media data. More specifically, they used textual content (e.g., user posts) and user-defined labels (e.g., a hashtag on Twitter or a subreddit on Reddit) to form a labelled dataset. Without human investigation, the data quality is not guaranteed to construct an accurate model. Another issue arises from the poor quality of labels. Using such a labelled dataset would confuse the model and cause it to make an incorrect decision. In this paper, we discuss the development and the role of data in text-based detection of mental health difficulties and negative life events in financial resilience. As an alternative to a data-driven approach, we model the problem as a multi-task classification task to detect the indicators of mental health difficulties and negative life events using prompt-based learning (i.e., a dataless approach).

Our main contributions[1] can be summarised as follows:

- We discuss developing a text-based system to detect the indicators of mental health difficulties and negative life events using two approaches: data-driven supervised learning and dataless prompt-based learning.

[1] Resources: https://github.com/summer1278/Reddit-MH-LS-Data

- To develop a data-driven approach and evaluate performance, we collect two datasets from Reddit to detect mental health difficulties (Reddit-MH-2021) and negative life events (Reddit-LS-2021).
- Given a piece of input text, we address the problem as a multi-task classification task that provides a decision and evidence. We conduct a comparative study on each sub-task with tests and case studies.
- We find the data-driven approach obtains a promising performance on decisions but fails to reproduce the same success in providing evidence. It suffers from data quality and overfitting problems, whereas the dataless approach overcomes such problems.
- An anxiety and depression dictionary is constructed using a hybrid approach that combines an automatic process to generate candidate words from professional assessments with post-verification by a psychology researcher.

II. Related Work

Data-driven supervised text classification has been extensively used for numerous studies in the field of mental health, for example, detecting the severity of posts on mental health forums [4], risks of self-harm [5], and depression [2], [6], [7]. These supervised approaches consist of three critical steps: collect datasets, extract features from the collected data, and train a model using a learning algorithm. Apart from the data, features play a vital role in the data-driven approach. Prior works extracted general features such as frequency-based features [2], [4], lexicon features [2], [4], [7], and word representations [4], [5]. In addition, several works extracted more features related to the specific task [7], [8]. These features were either expensively hand-crafted [8] or automatically extracted from social media posts [7]. In contrast, we develop a hybrid anxiety and depression dictionary that combines the advantages of a rich vocabulary from automated generation and the reliability of human post-verification.

A data-driven approach is straightforward when we have relevant data to build a model. However, most real-world problems are more challenging, so we could end up with poor-quality annotations or even without any task-specific data. Prompt-based learning has caught colossal popularity recently. The idea originates from *dataless classification* [9] that projects the given text and labels into a shared space without training data. It finds the most semantically similar label to fit the prediction result. Yin et al. [10] proposed a paradigm *zero-shot classification* that considers the classification task as a text entailment task by incorporating task description into a prompt and then transfers the knowledge from a pre-trained language model such as BART [11]. With the help of prompts, unseen tasks are modelled in a way that the language models can be used for prediction directly. Without fine-tuning the language model with a task-specific dataset, Radford et al. [12] demonstrated the ability of a language model to perform a translation task using a prompt (e.g., "English: `[Z]`, French: `[X]`.", where `[Z]` and `[X]` are sentences in English and French respectively). Similar prompt-based learning methods have been adapted to various NLP tasks such as emotion detection [13], intention detection [14], [15], topic classification [10], and question answering [16]. These prior works have provided a spotlight on the applications that are struggling with the availability of training data.

This paper discusses developing a text-based detection system for financial resilience to identify mental health difficulties and negative life events using prompt-based learning compared to the traditional data-driven approach using social media data.

III. Methods

This section discusses the development of a system to detect the negative impacts of financial resilience in the given text. More specifically, the system has two detection tasks: detect the occurrence of (1) mental health difficulties and (2) negative life events. Two approaches are presented in this section: (a) a data-driven supervised learning method (Section III-B), and (b) a prompt-based zero-shot learning method (Section III-C). The former method uses extracted features such as topic features and anxiety and depression dictionary features to construct a model using a labelled dataset (Section III-A). The latter uses a defined prompt (e.g., the life event is *negative*) to transfer knowledge from a pre-trained Natural Language Inference (NLI) model. Each approach models the task as a multi-task problem and consists of two sub-tasks for the given task: predict the existence of the given task (decision) and the associated indicators (evidence). Considering detecting mental health difficulties as an example, in the given text, we define their existence as a decision and associated symptoms as evidence.

A. Data

We collect two datasets (Reddit-MH-2021 and Reddit-LS-2021) for training data-driven supervised learning models and validating two methods (i.e., data-driven and dataless). We use the Pushshift Reddit API[2] to collect user discussions in the related topics (subreddit).

1) Reddit-MH-2021: Following prior works on mental health detection [3], [5], we collect submissions from 17 mental health related subreddits (*COVID19_support, suicidewatch, mentalhealth, EDAnonymous, adhd, alcoholism, socialanxiety, schizophrenia, healthanxiety, addiction, anxiety, depression, bpd, ptsd, autism, lonely and bipolarreddit*) and 18 unrelated subreddits (*guns, jokes, relationships, personalfinance, divorce, teaching, technology, meditation, worldnews, AskReddit, parenting, fitness, politics, legaladvice, soccer and conspiracy*). We limit the maximum number of posts from each subreddit to 50,000 to minimise the effect of the data imbalance problem and collected 8-month submissions between 01/01/2021 and 01/09/2021. The collected dataset contains 428,937 mental health related and 404,160 unrelated submissions.

[2] https://github.com/pushshift/api

2) Reddit-LS-2021: Following the indicators of life events defined by the Financial Conduct Authority (FCA) [17], we collect submissions on financial issues (*personalfinance, UKPersonalFinance, PersonalFinanceCanada, PersonalFinanceNZ, Money, Debt, debtfree, povertyfinance, jobs, Unemployment*), relationship breakdown (*relationships, relationship_advice, Marriage, Divorce*), caring responsibilities (*Parenting, Parents, elderlycare*) and bereavement (*SuicideBereavement, bereavement, GriefSupport, Separation*). The statistics of collected posts are financial issues (154,017), relationship breakdown (105,311), caring responsibilities (28,864) and bereavement (12,833). To construct a dataset for classification, we also collect posts from 16 subreddits that are irrelevant to negative life events (*anime, Jokes, houseplants, running, MakeupAddiction, movies, teaching, technology, Meditation, socialism, worldnews, AskReddit, Fitness, politics, soccer, gaming*). The collected dataset contains 301,025 negative life events related and 325,280 unrelated submissions.

3) Data Pre-processing: We collect the text content of the submission (post) and its subreddit (topic). We remove the submission with a *[deleted]* author or content. We also use Hierarchical Density based Clustering (HDSCAN) [18] to detect outliers and further remove general and unrelated information posted by the forum moderators. Then, we use NLTK Toolkit[3] to tokenise, stem, and lemmatise the words in the submissions. We create bigrams and trigrams (i.e., two-word and three-word phrases). We remove HTML markups, URLs, email addresses, non-ASCII digits, extra white spaces and English stop words.

B. Data-Driven Approach

Given a textual input x from social media posts, a classification task is defined to learn a model that predicts the conditional probability $P(\boldsymbol{x};\theta) \rightarrow y$ where $\boldsymbol{x}$ is the input vector mapped to a predicted label of the given text y [15]. y belongs to a fixed seen set of n labels $Y = \{y_1, y_2, ..., y_n\}$. In order to learn the parameters θ, in the *supervised learning* setting, we use a set of m train posts $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_m, y_m)\}$. A data-driven approach is based on the collected datasets to identify the negative impacts of financial resilience through two problems: mental health difficulties and negative life events detection. We use social media posts from Reddit-MH-2021 for mental health difficulties or Reddit-LS-2021 for life events, and address each problem as a multi-task problem having two sub-classification tasks: (a) predict the presence of negative impacts on the given problem (a decision model), and (b) show the evidence associated with the given problem (an evidence model). Fig. 1 shows the overview of the designed system. The detail of the training process can be found in Section III-B2. During the testing process, the evidence model is triggered when the given problem exists (i.e., y_d is *true*) and y_e belongs to a list of possible evidence.

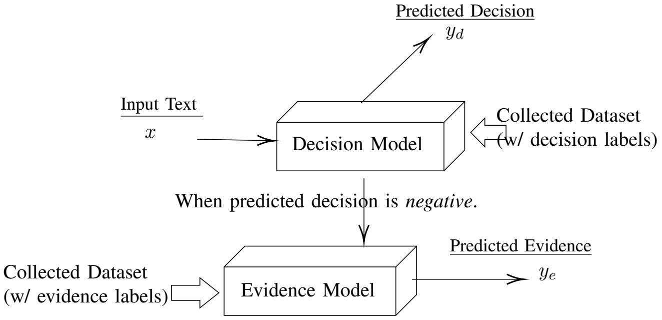

Fig. 1: System overview using a data-driven approach. To note, the mental health detection system contains only the decision model due to incomplete information about mental health symptoms. ⇔ indicates the route of training process, and ↔ indicates the route of the test process. w/ · denotes the dataset with a specific set of labels.

1) Feature Extraction: A data-driven approach is heavily based on a feature extractor. We extract Term Frequency–Inverse Document Frequency (TF-IDF) features, topic features, dictionary-based features and pre-trained word representations. More specifically, we computed the TF-IDF features (n = 1024, n is the number of features) to capture the important words that occur frequently in a document. Topic modelling enables grouping sets of words to form common topics in a collection of text in an unsupervised manner. It is commonly used to extract information from large datasets. We use Latent Dirichlet allocation (LDA) [19] to learn topics from mental health related subreddits on Reddit-MH-2021. A model with 40 topics is constructed (n=40). In this system, we focus on its usage as a part of the feature extractor, we use the learned model to build 40 topic features for a given document.

In addition, we extract two sets of dictionary-based features. The first set is the features extracted from existing manually-built lexicon databases: Linguistic Inquiry and Word Count (LIWC) [20] (n=61) and Empath [21] (n=194). The second set is the features extracted from our depression (n=21) and anxiety dictionaries (n=7). Unlike prior works, either manually built or automatically built the dictionaries using information extracted from social media posts [6]–[8], our dictionary is constructed using professional assessments and combined with human post-verification. We use Beck's Depression Inventory (BDI) [22] that includes 21 groups of statements such as sadness, pessimism, past failure, and the Generalized Anxiety Disorder questionnaire (GAD-7) [23] that includes seven questions to monitor the state of anxiety of patients. Each group has 4 statements for scoring the level of severity (score 0 to 3). We develop a tool to automatically extract related words or phrases to construct the base candidates of the dictionary and then expand them by their synonyms from thesaurus.com[4]. Score 0 indicates no symptoms of mental health issues with the patient, the tool only considers the candidates from the

[3] https://www.nltk.org/

[4] http://thesaurus.com/

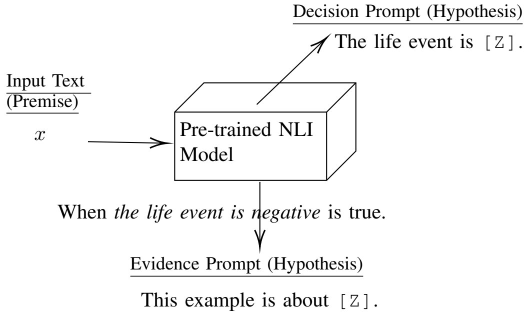
Decision Prompt (Hypothesis)
The life event is [Z].
Input Text
(Premise)
x
Pre-trained NLI Model
When *the life event is negative* is true.
Evidence Prompt (Hypothesis)
This example is about [Z].

Fig. 2: System overview of life event detection using a prompt-based dataless approach. No training or fine-tuning from the collected datasets is involved. The labels of decision z_d and evidence z_e are predicted from the pre-trained NLI model, where two associated prompt templates are filled with possible answers [Z], $z_d \in \mathcal{Z}_{\text{decision}}$ and $z_e \in \mathcal{Z}_{\text{evidence}}$. The lists of answers can be found in Table II.

level of severity score ≥ 1. After the automated expansion, a therapist was hired to post-verify the dictionary (4 hours). The constructed dictionary consists of 3,711 depression words (in 21 groups) and 716 anxiety words (in 7 groups). Using word embeddings from pre-trained models to present a document has been a popular option to overcome the out-of-vocabulary issue in TF-IDF. These pre-trained models were trained on massive datasets. They can provide a dense representation for a word in which a similar word should have a similar representation [24]. In our preliminary experiments, we find that FastText [25] trained on Common Crawl[5] outperforms the other pre-trained word embedding models (Section IV-A), we use it to produce word representations (n=300).

2) Training: We train three sub-models for the given multi-task problem using 60% of the collected datasets. The number of instances of training for each sub-model is summarised as follows: (a) mental health decision model (499,858), (b) life event decision model (375,786) and life event evidence model (180,615). The mental health model is trained on Reddit-MH-2021. Due to the incomplete information about the symptoms shown in the posts, we were unable to train a separate evidence model. The life event models are trained on Reddit-LS-2021, and the decision model is trained on a combination of related and unrelated submissions. In contrast, the evidence model is trained on a subset of related submissions that are associated with the four life event categories. We use the Random Forest algorithm to train classifiers because it reported promising results in depression classification with 89% accuracy and 90% sensibility [26]. Two training strategies are used in the development: one is to construct a joint model with the concatenation of feature sets, and the other is to construct a voting model from a set of separate sub-models trained on several single sets of features.

C. Prompt-based Dataless Approach

As an alternative to the data-driven approach, we define it as a zero-shot multi-task classification problem [10], which is more flexible and challenging since the labels are unseen during model development. Given a *premise*, NLI is a task of determining whether a *hypothesis* is true (i.e., entailment), false (i.e., contradiction), or undetermined (i.e., neutral) [15]. Table I shows the examples adapted to the mental health domain with each label. To use a model that was pre-trained for NLI to transfer the knowledge to a classification task, we convert the classification task description into a way that an NLI model could interpret. Without the need of the training dataset, *zero-shot learning* uses a pre-trained NLI model that models the input text x itself, in which $\boldsymbol{x}' = f_{\text{prompt}}(\boldsymbol{x})$. To distinguish from a seen label y, we use z to denote a label that can be unseen during the pre-training. We extend the prediction function to include the space of these labels $f_{\text{filled\_prompt}}(x', z)$ fills in the position of an answer [Z] given a prompt. The detail of the prompt design is presented in Section IV-B. To find the best answer $\hat{z}$ to the filled prompt, we follow Yin et al. [10] to introduce an argmax search function,

$$\hat{z} = \operatorname*{argmax}_{z \in \mathcal{Z}} P(f_{\text{filled\_prompt}}(x', z); \theta), \tag{1}$$

where it searches for the highest-scoring z that randomly generates outputs following the probability distribution of the pre-trained NLI model. Similar to supervised learning, we address the problem as two separate classification tasks for decision and evidence. Accordingly, we design prompt and answer templates for each sub-task. Fig. 2 shows the system overview of detecting life events using a prompt-based dataless approach. Mental health detection follows a similar structure with different sets of prompts and answers.

1) Prompt and Answer Engineering: Prompt engineering plays a vital role in the zero-shot classification [15]. A prompt is expected to interpret the information for a machine to understand. Following tests and case studies on a validation dataset, we design prompt templates and corresponding answers for mental health and life event detection. In each detection problem, two prompts, namely the decision prompt and evidence prompt, are constructed respectively, as shown in Table II. First, we define the decision task as detecting the presence of negative mental health symptoms. To find a specific symptom in mental health, we model the evidence prompt as a topic classification task. 26 symptoms were extracted from BDI and GAD-7 as the answers to fill the prompt template (Overlapped symptoms were removed). We also define the decision prompt as detecting its presence to detect negative life events. The evidence prompt is designed to perform a topic classification task with the four evidence categories as answers.

2) Pre-trained NLI Model and Fine-Tuning: Regarding the definition of the dataless prompt-based approach, we do not use any data from our collected datasets for training or fine-tuning in the experiments. We use a BART-large model [10] fine-tuned on the Multi-Genre Natural Language Inference

[5] https://commoncrawl.org/

TABLE I: Examples of NLI.

Premise	Hypothesis	Label
The man is sleeping.	A man has a mental health issue.	contradiction
An older man and a younger man are crying.	Two men are crying because of the smoke.	neutral
The lady lost her job.	The woman is suffering from a negative life event.	entailment

TABLE II: Prompt design. Given an input text x (premise), a pre-trained NLI model is used to map x to a prompt filled with an answer [Z] (hypothesis), where [Z] belongs to a list of possible labels in the specific model $\mathcal{Z}_{\text{decision}}$ or $\mathcal{Z}_{\text{evidence}}$.

Model	Type	Prompt Template	Answers / Labels ([Z])
Mental Health	Decision	The symptoms of mental health issues are [Z].	positive, negative
	Evidence	This example is about [Z].	sadness, pessimism, past failure, loss of pleasure, guilty feelings, punishment feelings, self-dislike, self-criticalness, suicidal thoughts or wishes, crying, agitation, loss of interest, indecisiveness, worthlessness, loss of energy, changes in sleeping pattern, irritability, changes in appetite, concentration difficulty, tiredness or fatigue, loss of interest in sex, mental control, anxious, relaxedness, restlessness, fearfulness
Life Event	Decision	The life event is [Z].	positive, negative
	Evidence	This example is about [Z].	caring responsibilities, bereavement, financial issues, relationship breakdown

(MultiNLI) corpus [27] for the experiments. The MultiNLI corpus[6] contains 392,702 sentence pairs with text entailment information from 5 categories: fiction (77,348), government (77,350), slate (77,306), telephone speech (83,348), and travel guides (77,350). We use the implementation of the BART model from Fairseq[7]. The NLI fine-tuned version is available at HuggingFace[8]. We do not change the parameters in the implementation. With the popularity of recent language model applications, a wide range of pre-trained language models could be considered, for example, BERT [28], GPT-3 [13] and GPT-4 [29].

IV. Experiments

We study extensively the development of two approaches: the data-driven approach and the prompt-based dataless approach. For the data-driven approach, we conduct experiments on various selections of features to construct supervised models of the problem (Section IV-A). For the dataless approach, we discuss how we design hand-crafted prompts and answers (Section IV-B). As we formulate the detection problem as a multi-task problem, we analyse the performance of each sub-task through classification tests. More specifically, we discuss their difference in providing decisions and evidence (Section IV-C). All experiments were conducted on an Intel i7-10750H 6-core 12-thread CPU.

A. Feature Selection in Data-Driven Approach

To select proper word representations for the task, we use a validation set of 166,619 posts (20% from the Reddit-MH-2021) to train a binary predictor on mental health/non-mental health classification. We use TF-IDF as a baseline to compare the test accuracy with several widely-used pre-trained word representation models[9]: Skip-gram [24], GloVe [31], Extended Dependency Skipgram [32], Turian [33], and two FastText variations (i.e., trained on Common Crawl or News database[10]) [25]. In addition, we conduct experiments using various single sets of other features (e.g. LIWC) to train the models and construct a majority voting classifier and a joint classifier that concatenates the combinations. Table III we observe word representation models have a similar performance. Using FastText trained on Common Crawl slightly outperforms the other models with a reasonable train time. Table IV shows the results for using a single set of features, the best majority voting model (i.e., voting from 3 separate models trained using FastText, Empath and Topic features, respectively), and the best meta-model (i.e., constructed on the concatenation of FastText, LIWC, Empath, BDI, GAD-7 and Topic features). Due to the complexity of training multiple sub-models to construct the voting candidates, using a majority voting model results in a longer training time compared to other models. We observe using the concatenation of all sets of features outperforms all other models. The results show that using a collection of task-specific features further improves the model's performance.

B. Prompt Engineering in Dataless Approach

To perform a classification task with a pre-trained NLI model, an appropriate prompt filled with suitable answers would increase the classification performance. As an early attempt, we take a test-agnostic approach. Mental health detection using NLP can generally be considered as a problem of identifying whether a given text shows the existence of mental health symptoms. The straightforward idea is to describe the task answers as *present* or *absent* of the mental health symptoms. We fill the designed answers into prompts and

[6] https://huggingface.co/datasets/multi_nli

[7] https://github.com/facebookresearch/fairseq

[8] https://huggingface.co/facebook/bart-large-mnli

[9] We use the implementation from flair [30].

[10] https://statmt.org/

TABLE III: Classification accuracy (acc) and class-balanced accuracy (b_acc) using pre-trained word representation models and TF-IDF features (baseline) on Reddit-MH-2021 dataset.

Representation Model	acc (%)	b_acc (%)	Train Time (in seconds)
TF-IDF	89.94	89.93	1.1426
Skip-gram	89.26	89.23	**0.5865**
GloVe	87.87	87.84	0.5934
Extended Dependency Skipgram	87.84	87.79	0.8155
Turian	82.78	82.75	0.6125
FastText (News)	89.32	89.28	0.7671
FastText (Common Crawl)	**91.01**	**90.97**	0.7763

TABLE IV: acc and b_acc using a single set of features or combinations of features on Reddit-MH-2021 dataset.

Feature / Combination	acc (%)	b_acc (%)	Train Time (in seconds)
LIWC	71.68	71.66	0.7588
Empath	83.82	83.80	0.9568
BDI	70.36	70.38	0.7386
GAD-7	63.47	63.41	**0.5366**
Topic	82.32	82.28	0.6204
Majority Voting (FastText, Empath, Topic)	89.31	89.27	4.6254
Concatenation (FastText+LIWC+Empath +BDI+GAD-7+Topic)	**91.30**	**91.25**	1.0073

TABLE V: acc on multiple evaluation datasets with mental health (MH) issues or negative life events. #Test denotes the number of test instances, and * denotes our collected dataset.

Dataset	Issues	#Test	Data-driven acc (%)	Dataless acc (%)
Twitter-2018	Depression	10,314	**79.14**	62.43
Twitter-2019	Depression	3,200	**75.97**	55.41
eRisk-2018	Anorexia	3,198	**90.12**	53.19
eRisk-2020	Self-harm	126,221	**78.22**	28.19
Reddit-2018	17 MH subreddits	177,089	**92.91**	57.45
Reddit-2019	17 MH subreddits	227,284	**88.90**	58.57
Reddit-MH-2021*	17 MH subreddits	166,620	**92.08**	59.12
Reddit-LS-2021*	Negative life events	125,261	**91.44**	58.58

test them on the validation set. We find the NLI model fails to interpret the task description correctly, and almost all test examples are classified as *present*. With the help of a domain expert, we slightly revise the answers to the decision prompt to capture more precise patterns. We find the model is able to improve the performance when we define the answers as *positive* and *negative* symptoms. Other prompt templates and answers are designed similarly through performance tests and case studies by inputting several hand-crafted prompt templates and answers with advice from a domain expert.

C. Data-Driven vs. Dataless

We compare the two approaches on two sub-tasks: decision and evidence classification. In addition to the task performance, we compare their runtime in practice with an API developed by us and hosted locally.

1) Decision: We evaluate the two approaches on eight datasets, including our collected datasets (Reddit-MH-2021 and Reddit-LS-2021) and six unseen datasets: Twitter-2018[11], Twitter-2019[12], eRisk-2018 [34], eRisk-2020 [5], Reddit-2018 [3], and Reddit-2019 [3]. Twitter-2018, Twitter-2019, eRisk-2018 and eRisk-2020 were collected for a slightly different task from ours. We focus on general mental health issues rather than a single mental health problem. Table V shows without fine-tuning the model to a particular mental health issue task (e.g., depression or self-harm), the data-driven approach obtained promising results on all evaluation datasets ($\geq$ 75% on two Twitter datasets and $\geq$ 78% on six Reddit datasets). The two Twitter datasets (Twitter-2018 and Twitter-2019) were collected based on hashtags or keyword matching "*depression*". Reddit datasets (Reddit-2018 and Reddit-2019) were collected similarly as in Section III-A. Using Reddit-LS-2021 as an example, we compute the classification accuracy on each subreddit. We find that the subreddits labelled as negative life events have reached 85.54% (69.57% to 92.31%), whereas the unrelated ones gain only 29.82% (14.46% to 45.47%) using the dataless approach. A similar trend was also observed in the Reddit-MH-2021 dataset. These results raise our concerns about the reliability of data annotations on social media platforms where a similar data collection strategy has been used in multiple prior works [3], [5]. We randomly select several examples from *AskReddit*, a subreddit considered to have no negative impacts on their financial resilience (either mental health difficulties or negative life events). We find a number of posts that show negative symptoms in mental health, for instance, "*how to deal with stress?*" and "*what is a moment where you honestly thought you were going to die?*". Compared to the data-driven approach, the dataless approach is able to capture these posts as related to mental health and show *mental control* and *suicidal thoughts or wishes* as their main symptom, respectively.

2) Evidence: Due to the availability of evidence labels, we evaluate two approaches for evidence classification on the Reddit-LS-2021 dataset. The data-driven approach does not achieve the same level of performance for evidence classification as in the decision classification. The test accuracy drops to 50.53% with b_acc at 25.03%. There are two major observations from our experiments using a data-driven approach.

a) Data Imbalance: First, the evidence train set is strongly imbalanced, for example, *financial issues* and *bereavement* (154,017 vs. 12,833). Although the classification accuracy on the train set has reached 99.97%, most posts are still classified as financial issues (60K) and relationship breakdown (2.3K) during testing. To reduce the effect of the imbalanced dataset, we applied two resampling methods to the train set: oversampling the data from the minority classes by Synthetic Minority Over-sampling Technique (SMOTE) [35], and undersampling the data from the majority classes by

[11] https://github.com/viritaromero/Detecting-Depression-in-Tweets

[12] https://github.com/swcwang/depression-detection

TABLE VI: Test examples in Reddit-LS-2021.

#	Submission Content [**original label**, *predicted label* using data-driven approach]
1	Anyone else just feel alienated from everyone? Idk. Knowing I'm (15F) am the girl whose mother left her for death and who found her body one morning just completely alienates me from those around me. And then there's the events that happened before. I feel like a broken mess on the floor. Anyone else feel like this? [**bereavement**, *relationship breakdown*]
2	Family has over 50million USD but I'm given nothing Long story short, my dad passed away and left his 50M USD to my mom. My mom bought a lot of luxury things for my brother like Ferraris, Lamborghinis and other luxury things. She also buys lots of properties for herself. He does work for my dad's company but he's very lazy. I'm currently living abroad and she has not given me anything. I make my own money. Is it wrong for her to not give me anything? All I want is a house, nothing fancy or expensive. Am I a bad person for asking too much? [**caring responsibilities**, *financial issues*]

TABLE VII: Average prediction probabilities in each category of life events. The highest predicted class probabilities in each defined life event category are bolded.

Method	Data-driven (Supervised Learning)				Dataless (Prompt-based Learning)			
Predicted Defined	bereavement	caring responsibilities	financial issues	relationship breakdown	bereavement	caring responsibilities	financial issues	relationship breakdown
bereavement	0.0429	0.0982	**0.5074**	0.3515	**0.4941**	0.3195	0.0419	0.1445
caring responsibilites	0.0445	0.0986	**0.5102**	0.3467	0.0740	**0.6701**	0.0655	0.1904
financial issues	0.0431	0.0963	**0.5093**	0.3513	0.0338	0.3828	**0.5030**	0.0804
relationship breakdown	0.0430	0.0979	**0.5061**	0.3531	0.1084	**0.4514**	0.0943	0.3458

NearMiss [36]. Neither of them improves the performance, we observe the test accuracy further drops to 45.25% and 45.27%, respectively. We suspect the data imbalance may not be the major cause of the performance drop in predicting evidence.

b) Multiple Labels: The second observation is that a post may have multiple labels. We randomly select 602 examples from each category in the same ratio with respect to the whole test set for the case study. We demonstrate two examples in Table VI. Example #1 is a post with a *bereavement* label, but it is classified as *relationship breakdown*. The content of losing a family member suggests that it may have ≥ 2 labels. A single label fails to cover other related labels. This example can also be considered a relationship breakdown or a financial issue for a young girl. Example #2 is a post with a *caring responsibilities* label but is classified as *financial issues*. The post discusses the relationship with family members when dealing with financial problems. We observe that a post may have multiple labels: we find 29.91% having ≥ 2 labels and 7.76% having ≥ 3 labels. To better understand when the post has more related categories in life events (or mental health symptoms), the detection system is revised to output all possible labels with their prediction probabilities instead of a single label. This modification enables us to evaluate the system by computing the average probabilities of getting all possible labels (i.e., the four life events) for the posts from each defined category. Table VII shows the average prediction probabilities for all defined categories, we observe that the data-driven approach does not capture the characteristics of posts in each category. They all show higher prediction probabilities in financial issues, whereas the dataless prompt-based approach overcomes this issue. The NLI model used in the prompt-based approach was pre-trained on more diverse datasets (Section III-C2). For this reason, it reduces the chance of overfitting to a dataset with uncertain quality.

3) Runtime and Model Size: We measure the average runtime of the two approaches over 5 random selections. Each selection contains 100 test examples with an average length of 220 words. We find that the average runtime per instance using the data-driven approach is about two times faster than using the dataless approach (2.55s vs. 4.36s). This is because loading and running a supervised learning model (size: 266MB) is much faster than a pre-trained large language model (size: 5GB).

V. Conclusions and Future Work

In this paper, we discussed two approaches to a text-based financial resilience detection system that targets the factors of mental health difficulties and negative life events. We demonstrated the development of a data-driven approach using social media data from data collection and feature extraction to training strategies. In addition to the traditional data-driven approach, we designed a dataless approach that uses prompts to incorporate the class information to transfer knowledge from existing pre-trained NLI models. Moreover, we discussed our concerns about using social media data to construct a reliable model (data-driven) in comparison to the dataless approach using a defined prompt.

There are a few unresolved problems for future research directions. One of the directions is to enhance the knowledge transfer from the pre-trained model by expanding the base prompts. As an early attempt at prompt-based learning in financial resilience detection, we used hand-crafted prompts by domain experts. We plan to use these prompts as the seed and expand them with multiple machine-generated prompts (i.e., automatic prompts). Another one is to develop trustworthy evaluation metrics for financial resilience.

References

[1] W. H. Organization *et al.*, "World health statistics 2022: monitoring health for the sdgs, sustainable development goals," 2022.

[2] J. Wolohan, M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard, "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP," in *Proceedings of the First International Workshop on Language Cognition and Computational Models*, (Santa Fe, New Mexico, USA), pp. 11–21, Association for Computational Linguistics, Aug. 2018.

[3] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study," *J Med Internet Res*, vol. 22, p. e22635, Oct 2020.

[4] S. Malmasi, M. Zampieri, and M. Dras, "Predicting post severity in mental health forums," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pp. 133–137, 2016.

[5] L. Oliveira, "Bioinfo@ uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases," in *Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece*, pp. 22–25, 2020.

[6] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 3838–3844, 2017.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, pp. 128–137, Aug. 2021.

[8] A. Kumar, A. Sharma, and A. Arora, "Anxious depression prediction in real-time social data," in *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India*, 2019.

[9] M.-W. Chang, L. Ratinov, D. Roth, and V. Srikumar, "Importance of semantic representation: Dataless classification," in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, p. 830–835, AAAI Press, 2008.

[10] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3914–3923, 2019.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[14] V. Varshney, M. Patidar, R. Kumar, L. Vig, and G. Shroff, "Prompt augmented generative replay via supervised contrastive learning for lifelong intent detection," in *Findings of the Association for Computational Linguistics: NAACL 2022*, (Seattle, United States), pp. 1113–1127, Association for Computational Linguistics, July 2022.

[15] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[16] W. Yin, N. F. Rajani, D. Radev, R. Socher, and C. Xiong, "Universal natural language processing with limited annotations: Try few-shot textual entailment as a start," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 8229–8239, Association for Computational Linguistics, Nov. 2020.

[17] "Guidance for firms on the fair treatment of vulnerable customers," Tech. Rep. 3, Financial Conduct Authority (FCA), July 2019.

[18] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[19] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of Population Structure Using Multilocus Genotype Data," *Genetics*, vol. 155, pp. 945–959, 06 2000.

[20] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[21] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, (New York, NY, USA), p. 4647–4657, Association for Computing Machinery, 2016.

[22] A. T. Beck, R. A. Steer, and G. K. Brown, *Beck depression inventory (BDI-II)*, vol. 10. Pearson, 1996.

[23] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[25] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[26] E. Souza Filho, H. Rey, R. Frajtag, D. Cook, L. Carvalho, A. L. Ribeiro, and J. Amaral, "Can machine learning be useful as a screening tool for depression in primary care?," *Journal of psychiatric research*, vol. 132, pp. 1–6, 09 2020.

[27] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, Association for Computational Linguistics, 2018.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[29] OpenAI, "Gpt-4 technical report," 2023.

[30] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.

[31] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[32] A. Komninos and S. Manandhar, "Dependency based embeddings for sentence classification tasks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 1490–1500, Association for Computational Linguistics, June 2016.

[33] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 384–394, 2010.

[34] D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk: Early Risk Prediction on the Internet," in *Proceedings Conference and Labs of the Evaluation Forum CLEF 2018*, CLEF 2018. Lecture Notes in Computer Science, vol 11018, (Avignon, France), 2018.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002.

[36] I. Mani and I. Zhang, "knn approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126, pp. 1–7, ICML, 2003.