Text Mining to Understand Emotion Triggers

by

Liuyan Chen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Science

Waterloo, Ontario, Canada, 2019

Author Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In computational linguistics, most sentiment analysis builds binary classification models on customer reviews data to predict whether a review is positive or negative. In this thesis, we go a step further and build interpretable classification models to predict fine-grained emotions associated with text (such as happy, sad, productive and tired). This analysis is enabled by a unique journaling dataset containing short pieces of text and associated emotional status self-reported by writers. To further study what people feel emotional about (emotion triggers), we perform model interpretation.

We make two main contributions. First, we apply state-of-the-art text mining methodologies to extract emotion triggers from text, during which we discover and solve an issue with the attention mechanism in a popular deep learning model (DMN). Second, we obtain data-driven evidence of emotion triggers based on a group of 67,000 people, which contributes to a better understanding of emotion triggers from the perspective of public mental health.

Acknowledgements

This thesis would not have been possible without the constant support that I received from many people.

First and foremost, I would like to express my heartfelt gratitude to my amazing supervisor, Professor Lukasz Golab, for his patient guidance, encouragement, and advice throughout this research. I am fortunate to have a supervisor who is exceptionally knowledgeable and has inspired me tremendously in the research. I am also grateful for his understanding and support when I need time away to fulfill my family obligations.

I want to acknowledge Professor Olga Vechtomova and Professor Oliver Schneider for taking their time to review my thesis and provide valuable feedback.

Life of grad school would not have been such a joyful journey without the friends that I made here at the University of Waterloo. I thank all my lovely friends for always being there for me.

I am incredibly grateful to my husband and my parents for encouraging me to pursue the research area that I am passionate about.

Dedication

I dedicate this thesis to my beloved husband and parents for their unconditional love, support, and encouragement.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CNN Convolutional Neural Network 17, 18

Df Document frequency 14

DMN Dynamic Memory Network iii, ix–xii, 6–8, 10, 18, 19, 21, 24–26, 28, 29, 36–42, 44, 52, 64–69, 71–74, 77

GloVe Global Vectors for Word Representation 30, 40

GMT Greenwich Mean Time 27, 30

GRU Gated Recurrent Unit 17, 19, 20, 40

Idf Inverse document frequency 14

LSTM Long Short-term Memory 17, 18

NLP Natural Language Processing 2, 9, 11, 12, 16

QRNN Quasi-Recurrent Neural Networks 17, 18

RNN Recurrent Neural Networks 17, 18

RNTN Recursive Neural Tensor Network 17

SVMs Support Vector Machines 15, 32, 33

Tf-idf Term frequency-inverse document frequency 14, 15, 17, 31, 32, 35, 73

Tf Term frequency 14, 15, 17, 31

List of Symbols

a_0 Input to the answer module 21

b Bias parameters in neural network layers 20

C_t Final outputs of the input module 19, 20

df_t Document frequency of term t 14

E_{CE} The standard cross-entropy cost 21, 39, 41

e^i New input representation of episode i 20

g_t^i The attention gate of sentence t during memory episode i 20, 39

h_t The hidden sate at time step t in the input module 19

idf_t Inverse document frequency of term t 14

J Loss function 21, 39, 41

λ A training hyperparameter that determines regularization strength 39, 41

L The word embedding matrix 19

m^i Memory at episode i 20

N Total number of documents in a corpus 14

q_{T_Q} Final outputs of the question module 20, 40

q_t The hidden sate at time step t in the question module 20

$softmax$ Softmax function 20, 21

tanh Tanh activation function 20

T_C Number of sentences in an input journal 19, 39, 40

$tf_{t,d}$ Term frequency of term t in document d 14

T_I Number of words in an input journal 19

T_M Maximum number of memory iterations in Episodic Memory Module 21, 40, 41

T_Q Number of words in a question 20

w_k^E Word index of emotion class k 40

w All weight parameters in neural networks 39

w_t Word index of word t in an input journal 19

W Weight parameters in neural network layers 20

y_0 Softmax probabilities of all classes 21

Chapter 1

Introduction

1.1 Motivation and Problem Definition

Emotion triggers are defined as environmental stimuli that bring about an emotion (James, 1884; Lazarus, 1991). They play a vital role in emotion regulation therapy, which is a necessary treatment to help people regulate their emotions (especially negative emotions) and to prevent negative thoughts and feelings from driving our behavior[1]. Gross and Muñoz (1995) conclude that there are two ways of regulating emotions. One way is to change external environments where certain emotion triggers are much more likely to occur. The other way is to select mental environments to change the probability of certain emotions occurring. These two ways of emotion regulation inevitably require a detailed understanding of emotion triggers. For example, if we know a concrete instance of emotion triggers for happiness ("working out"), we can regulate our emotions by changing the external environments (going to the gym) and changing our mental environments to enjoy the activity.

Motivated by the importance of emotion triggers, we define our research goal as to understand emotion triggers through text mining.

Specifically, we analyze an anonymous dataset that contains over 700,000 mood journals from over 67,000 writers. Each record in our data comes with a short journal that expresses the writer's feelings and a fine-grained emotion label that is self-reported by the writer. In total, we have 16 types of emotions as follows:

[1]Sources: http://www.psychologytoday.com/ca/basics/emotion-regulation Last visited Mar 5^{th}, 2019.

Angry, Sad, Stressed, Frustrated, Down, Lonely, Anxious, Overwhelmed, Tired, Calm, Good, Productive, Accomplished, Happy, Excited, Ecstatic

For example, a synthetic journal record is:

Journal: *After talking with my parents, I wasn't happy at all. But he calmed me down and supported me. I'm so happy to have him!*

Emotion: *Happy*

Given the emotion label of the above example, an emotion trigger for *Happy* can be summarized from text through manual inspection: others' support and understanding can trigger *Happy*.

To automatically summarize emotion triggers for all emotions in our data, we develop methodologies that first build interpretable emotion classification models and then extract emotion triggers through model interpretation and manual inspection.

Note that our study presents data-driven evidence of correlations between emotion triggers and associated emotions, but does not conclude any causality as we do not conduct controlled experiments on the emotion triggers that we identify.

Our study is challenging in two ways:

- NLP itself is difficult due to the ambiguity of human languages. Our problem is more difficult than general NLP tasks because people could perceive the same language context differently. Several previous studies show that sentiment classification tasks are generally more difficult compared to traditional topic classification tasks. Unlike topics, which can be identified through keywords, sentiment relies upon more than keywords (Pang et al., 2002; Mishne, 2005).

- Unlike most open-source sentiment data that only have binary sentiment labels, our data are more fine-grained and include 16 different mood labels. While binary sentiment detection on text is hard due to language ambiguity, our fine-grained emotion detection is more challenging. Even for humans, it is not an easy task to distinguish between many different emotions. Previous studies show that the more sentiment labels the data have, the worse the model performs (Mohammad, 2012; Pang and Lee, 2005; Thelwall et al., 2010a; Socher et al., 2013).

1.2 Background

Overall, there have been three lines of work that study sentiment or emotions: 1) Psychological emotion studies, which use small survey samples. 2) Sentiment analysis in computational linguistics, which uses only binary sentiment labels and does not attempt to interpret emotion triggers. 3) Emotion analysis in computational linguistics, which does not attempt to interpret emotion triggers. We compare those studies with ours in Figure 1.1.

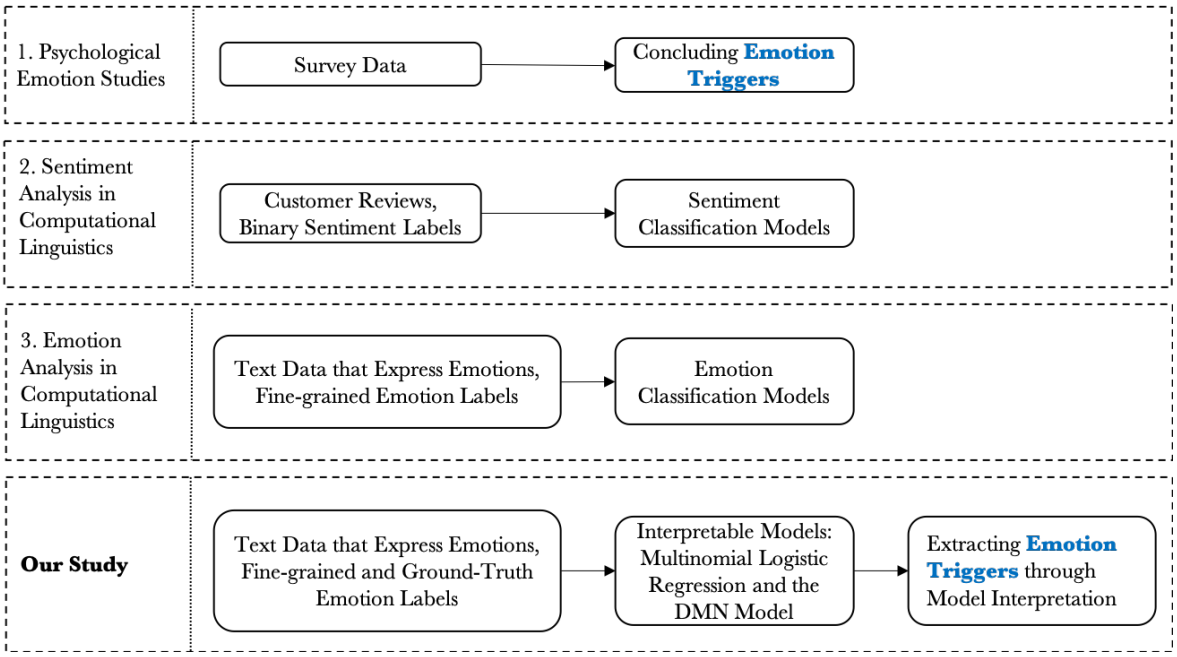

Figure 1.1: A comparison of our study and related prior work.

Most psychological emotion studies tested the statistical significance of particular emotion triggers by collecting data from surveys, and they offered either a high-level summary of possible types of emotion triggers (Ekman, 2003) or a small set of triggers that are only applicable to a small population (Liebert and Morris, 1967; Bond et al., 2001). For example, based on the results of experiments with people who live in isolated tribes, Ekman (2003) discussed possible paths of emotion generation. For example, he concluded that "things affecting welfare" and "memories of past emotional experiences" could cause emotions. Although these conclusions can offer a basic understanding of general types of

emotion triggers, we do not know what the "things" or "emotional experiences" specifically mean. It would be more helpful to know concrete examples of emotion triggers (for example, working out is a trigger for happiness), as previous research shows that being aware of potential emotion triggers can contribute to better emotional health (Gross and Muñoz, 1995). In one study by Bond et al. (2001), they surveyed 2,680 teenage students and found that victimization history can positively predict anxiety or depression syndromes among teenagers. Overall, using survey data has two main limitations: 1) Predefined survey questions can limit findings of unknown emotion triggers. 2) The relatively small scale of survey data could result in low data support.

Sentiment studies in computation linguistics have been enabled by the availability of large-scale social media text data during recent decades. They typically collect data from customer review sites, such as IMDb and Yelp, so the sentiment studied is people's opinion and attitude toward products. For example, one popular sentiment analysis data is the Stanford Sentiment Treebank (SST), which was compiled by Socher et al. (2013) from movie reviews. The dataset includes fine-grained sentiment labels — very negative, negative, neutral, positive, and very positive — for 11,855 sentences. Table 1.1 shows one data example in SST, where the label is at first manually annotated using Amazon Mechanical Turk. Then, those annotated labels were merged into five classes by the authors because some labels were rarely selected by annotators. Figure 1.2 demonstrates the interface that annotators see.

Sentence:	"This movie doesn't care about cleverness, wit or any other kind of intelligent humor."
Label:	"negative"

Table 1.1: Sample data in Stanford Sentiment Treebank in Socher et al. (2013)

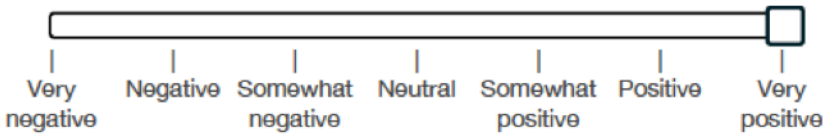

Figure 1.2: Labeling interface in Socher et al. (2013). Upon seeing a random phrase, annotators selected a corresponding sentiment.

As presented in Figure 1.1, there are two major limitations of previous sentiment analysis in computational linguistics:

- The customer reviews data and sentiment labels lack connection with real human emotions. Comparing the data in sentiment analysis (Table 1.1) with that in emotion analysis (our journal data), it is clear that customer reviews data reflect more about people's opinions towards commercial products rather than people's internal mental status. Moreover, the sentiment label "negative" is too coarse to be connected with real human emotions, such as "Fear", "Anger", and "Embarrassment".

- Most sentiment studies primarily focus on optimizing model accuracy but do not attempt to interpret emotion triggers (Socher et al., 2013).

While sentiment studies in computational linguistics mostly use customer review data, emotion analysis in computational linguistics collects data that are written to express emotion status. Wang et al. (2012) collected 2.5 million tweets using a list of 131 seed words consisting of basic emotion words and their synonyms. Table 1.2 shows several sample sentences. Words in bold are hashtags in original tweets and are treated as mood labels in their emotion classification. However, similar to sentiment studies in computational linguistics, most emotion studies in computational linguistics have the same limitation: they primarily focus on optimizing model accuracy but do not attempt to interpret emotion triggers. For example, Wang et al. (2012) built multinomial classifiers to predict emotions but did not further explore why people feel a certain emotion.

1	**Fear:** *"When I see a cop, no matter where I am or what I'm doing, I always feel like every law I've ever broken is stamped all over my body"*
2	**Anger:** *"I hate when my mom compares me to my friends."*
3	**Embarrassment:** *"I hate when I get the hiccups in class."*

Table 1.2: Sample emotion text in Wang et al. (2012)

Some emotion analysis papers involve emotion triggers, but they do not aim to obtain aggregate insights into emotion triggers for different types of emotions (Neviarouskaya and Aono, 2013; Li and Xu, 2014; Ghazi et al., 2015). For example, Li and Xu (2014) only used the extracted emotion triggers as additional word features to classify emotions.

Although the above studies in computational linguistics primarily focus on achieving good model accuracy, we find two recent papers that apply interpretable text mining models and attempt to interpret their models. Kumar et al. (2016) built a Question-Answering model with an attention module — the DMN. It allows both supervised attention training and unsupervised attention training. The former training method requires labels of actual answers and location labels of important sentences that contain supporting facts for answers, while the latter method only requires labels of actual answers. Their experiments in sentiment classification achieved a higher prediction accuracy than before and demonstrated strong interpretability of their attention mechanism (Figure 1.3). Bagroy et al. (2017) built a binary logistic regression model to classify whether text is mental-health related or not, and interpreted their models through term coefficients to conclude predictors for mental-health-related text.

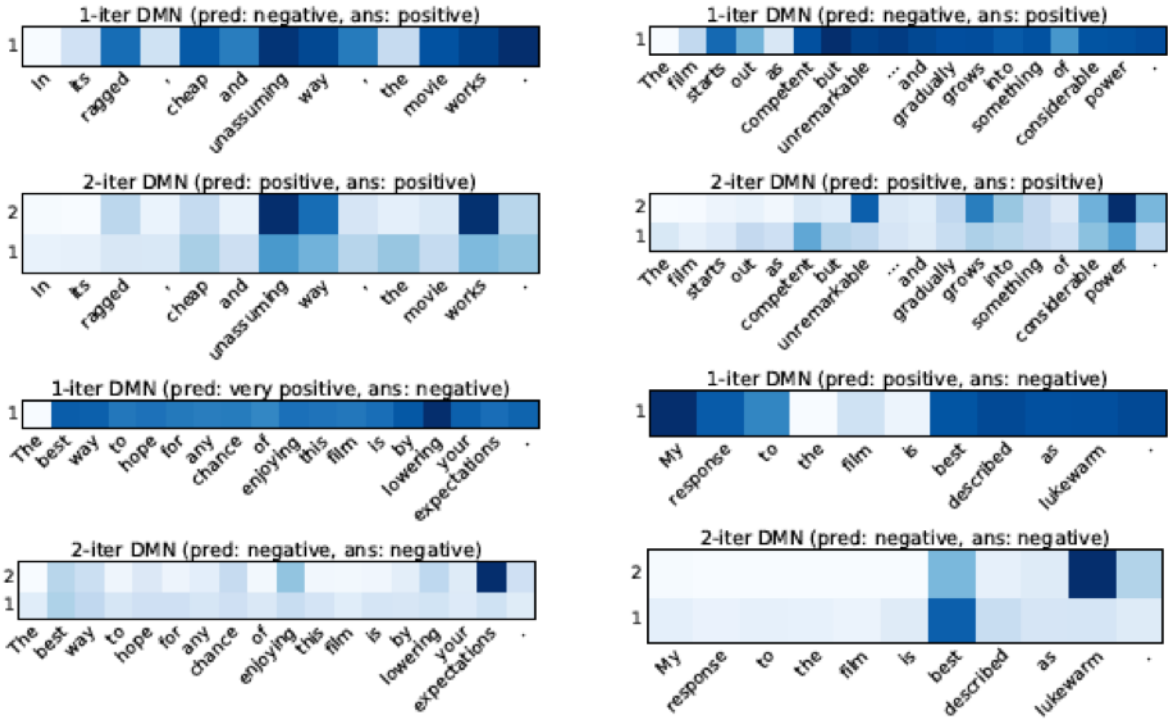

Figure 1.3: Sentiment analysis examples trained by the DMN in Kumar et al. (2016), with the question as "What is the sentiment?". Darker colors mean higher attention weight of the input word. The X-axis represents the input movie review that is tokenized by words. Y-axis represents the 1^{st} and 2^{nd} memory iterations.

However, Kumar et al. (2016) did not proceed further to analyze important predictors for sentiment on all data besides showing a few examples. Also, they did not experiment with their model on an emotion dataset as ours. For the work from Bagroy et al. (2017), as their model is not about classifying emotions, their findings are more about potential triggers for general mental health concerns instead of emotions.

Driven by the lack of findings of emotion triggers in prior work, our work is different in that we use a unique emotion dataset that has ground-truth and fine-grained emotion labels, and we focus on not only getting high classification accuracy but also extracting and summarizing emotion triggers from our models (See Figure 1.1).

1.3 Contributions

In this thesis, we make two main contributions:

1. We test and inspect two interpretable text mining models that help us extract emotion triggers from our data. We show that our multinomial logistic regression model using n-gram features produces a strong accuracy score. Further, emotion triggers are summarized from predictive terms that are selected based on term coefficients. We also apply the DMN model, and find that although our data do not have labels of sentences that contain emotion triggers, which are essential to train supervised attention, our approach that combines emotion classification and model interpretation helps us identify sentences containing emotion triggers. We first train the DMN to classify emotions using the question "What do I feel", with attention trained in an unsupervised way. Next, we feed each journal into the trained model with a new question "Why do I feel x" (x is replaced with the associated ground-truth mood label). Based on attention gates of sentences, we select important sentences that contain emotion triggers and further summarize those sentences into categories of emotion triggers.

   During the process of unsupervised attention training, we observe and solve an issue with the attention mechanism in the DMN. The issue makes the attention layer assign equal attention gates to all input sentences and defeats the purpose of the attention module — to measure which sentences contain emotion triggers. Through investigation, we discover that the problem is due to misuse of the popular L2 regularization. Further, we show that using our proposed regularization methods on attention can significantly increase attention sparsity even when attention is trained

in an unsupervised way. We suspect that all current models using attention mechanism and training attention in an unsupervised way may subject to this issue, due to the popularity of the L2 regularization.

2. We obtain useful insights regarding emotion triggers. The insights can help the public understand emotion triggers better. Also, knowing a concrete set of emotion triggers for different emotions can help emotional therapists carry out treatments. For example, we find that "job interviews" is a trigger of both "Anxious" and "Excited". Different emotional reaction upon the same trigger may result from individual differences in cognition processes. We suggest that emotion therapists can help patients who demonstrate increasing anxiety before job interviews re-evaluate impacts of "job interviews" and reduce their anxiety. Emotion triggers that we identify through text mining can also provide researchers in Psychology with future research topics.

1.4 Chapter Outline

The rest of this thesis is organized as follows:

- Chapter 2 offers a comprehensive review of previous psychological emotion analysis, as well as emotion and sentiment analysis work in computational linguistics that uses large-scale text data.
- Chapter 3 explains each step in our data processes by details.
- Chapter 4 shows findings from different stages of our analysis: exploratory data analysis, the bag-of-words modeling, and the DMN modeling.
- Chapter 5 discusses our main research findings and their implications. We also point out several limitations of our study and make recommendations for further research work.

Chapter 2

Definitions and Literature Review

In psychology, research into emotions has a long history, and some early studies include James (1884) and Lazarus (1991). While lots of these researches are qualitative in nature, there are also a large number of studies using quantitative methods. Those quantitative emotion studies in psychology typically collect a relatively small scale of data from surveys, experiments, or observational studies, and investigate the mechanism of emotions and emotion triggers. Roughly since the year of 2000, there has been a growing volume of sentiment analysis in computational linguistics, which utilizes NLP, machine learning and deep learning techniques to analyze text data collected from online customer review sites. Those sentiment analysis in computational linguistics refers sentiment as to whether people like or dislike the product rather than people's internal emotional status, and those studies focus primarily on developing computational methods to predict sentiment based on input text rather than interpreting the mechanism of sentiment. Realizing the disconnection between the concept of sentiment and emotions, more recent studies in computational linguistics in the last ten years start to emphasize the importance of emotions. Those emotion analysis studies in computational linguistics collect text data that are created to express emotions and use emotion dimensions according to basic emotions defined in psychology instead of the binary sentiment dimensions used in the computational sentiment analysis. However, most of those emotion studies do not attempt to interpret emotion triggers. Therefore, to date, there have been limited studies in computational linguistics that investigate emotion triggers and provide aggregate insights into emotion triggers for different emotions.

In this chapter, we first review the above three lines of prior work that are related to our study: quantitative emotion analysis in psychology, sentiment analysis in computational linguistics, and emotion analysis in computational linguistics. Specifically, we review two

computational models in the second line of work — the bag-of-words machine learning models and the DMN model, as we apply them in our analysis. In the end, we summarize limitations of previous studies.

2.1 Quantitative Emotion Analysis in Psychology

Quantitative emotion analysis in psychology mostly conducts surveys, experiments, or observational studies to study a specific group of people and a particular type of emotions. Many studies focus on exploring relationships between particular emotion triggers and particular emotions using statistical testing. Surveys that were conducted by Liebert and Morris (1967) found that the higher the students' confidence in performing well on the exam, the less worried they feel. Fredrickson and Joiner (2002) observed that students' initial positive emotion state would broaden their thinking and cognition and in turn predict positive emotions in the future. Bond et al. (2001) used a survey to assess relationships between a victimization history and anxiety or depression syndromes among teenagers. They concluded that the victimization history, including either been teased or deliberately excluded, have had rumours spread about them, or have experienced physical threats and violence, can positively predict anxiety or depression syndromes among teenagers. Campbell-Sills et al. (2006) demonstrated from a clinical experiment that experiencing the same emotion trigger, participants with mood disorders show a higher level of suppression than non-clinical participants, and the difference in negative mood suppression among females is higher than that of males.

While quantitative emotion analysis in psychology provides concrete examples of emotion triggers, there are three main drawbacks of their approach. First, their findings are usually lack of data support since only a relatively small scale of data are analyzed. Limited by the number of participants, the size of data is usually at most a few thousand. For example, the work of Liebert and Morris (1967) was based on a sample of just 54 students. Also, there were in total 2,680 teenage students who took the survey of Bond et al. (2001). The clinical experiment of Campbell-Sills et al. (2006) involved only 90 participants. Second, those studies usually need first to establish a hypothesis that certain emotion triggers are related to certain emotions before conducting surveys and collecting data, so the results are only specific to that particular emotion and emotion triggers in the original hypothesis. Third, the designing of the surveys and experiments usually requires a deliberate choice of studied emotions and emotion triggers according to the established hypothesis, which we think can limit findings of unknown triggers.

2.2 Sentiment Analysis in Computational Linguistics

Sentiment analysis is a popular research topic in computational linguistics. It includes studies that utilize NLP, text mining, machine learning, and deep learning to identify and analyze sentiment in text automatically. Most sentiment studies in computational linguistics typically collect data from customer review sites, such as IMDb, Yelp, Amazon and so on, so that the sentiment studied is people's opinion or attitude toward products. One earliest and most cited sentiment analysis work in computational linguistics is from Pang et al. (2002), where computational models were built to predict whether people like or dislike movies.

Since the work of Pang et al. (2002), the research on sentiment in computational linguistics has gained momentum, and text mining techniques to study sentiment have been optimized continuously. Our work also benefits hugely from the methodologies of previous computational sentiment analysis.

2.2.1 Online Sentiment Data Sources

Roughly since the year of 2000, large-scale sentiment data have become available online because of the booming development of the internet, mobile technology, and social media. Online platforms, especially social media sites and blogs, have gained massive popularity among online users. Many people tend to share their stories, express mental needs and seek help through the internet. This trend further allows researchers from computational linguistics to test and build computational models to study sentiment. First, people are facing with increasing work pressure, financial stress and interpersonal relationship issues due to the fast-paced modern lifestyle[2]. Second, there are numerous reports of mental disorders and suicidal cases, which continuously attract peoples' attention to emotional health and mental well-being[3]. Third, people's willingness and acceptance of using the internet and social media is unprecedentedly high due to the convenience brought by technologies.

[2]According to The Changing Workplace: A Survey of Employees' Views and Experiences, the percentage of employees who feel under a great deal of pressure has increased from 51% to 57% over the period 2003 to 2009. Source: https://www.researchgate.net/figure/3-shows-that-pressure-increased-in-both-the-public-and-private-sector-over-the-period_tbl22_279798945 Last visited Mar 5^{th}, 2019.

[3]The death rate for suicide among children and adolescents doubled from 2007 to 2014. Death Rates for Motor Vehicle Traffic Injury, Suicide, and Homicide Among Children and Adolescents aged 10-14 Years the United States, 1999-2014. Source: https://www.cdc.gov/mmwr/volumes/65/wr/mm6543a8.htm Last visited Mar 5^{th}, 2019.

There are three major advantages of using online data instead of survey data. First, it is easier to acquire a large-scale dataset. For example, the SST dataset contains 11,855 sentences from movie reviews (Socher et al., 2013). Pak and Paroubek (2010) utilized a dataset with 300,000 tweets. Second, it is possible to obtain a much larger population than surveys. The Amazon reviews data in Zhang et al. (2015) contain product reviews from over 6 million users. Third, as text content is created freely by online users, the response bias in surveys is greatly reduced. For example, to collect emotion-related content from surveys, researchers usually must compile a questionnaire with questions about emotions and emotion triggers. The structured questions potentially limit the findings of unknown emotion triggers.

In terms of the sentiment labels, ground-truth sentiment labels exist in most data that are collected from customer review sites. For each review, there is usually a sentiment label that is assigned by the reviewer. For example, each movie review in the IMDb data used in Pang et al. (2002) is expressed either with stars or a numerical value assigned by the reviewer. For sentiment datasets that do not have ground-truth sentiment labels, we find that most work develops labels using human annotators. For example, sentiment labels in the Stanford Sentiment Treebank (SST) were manually annotated by human annotators on Amazon Mechanical Turk (Socher et al., 2013). There are also some papers that developed sentiment labels using emoticons. For example, Go et al. (2009) used a list of emoticons [:), :-), :), :D, and =)] to query and label positive tweets, and a list of emoticons [:(, :(, :-(, and : ( ] to query and label negative tweets .

2.2.2 NLP for Pre-processing Text

NLP is a field in computer science that is concerned with how to process and analyze a large amount of natural language data using computers. Since most text content in sentiment dataset is created freely by individual users, there are many informal text expressions that require careful cleaning before further analysis.

There are two important NLP processes that are applied in prior work to remove irrelevant information in text:

- According to Python's[4] built-in string punctuation set, all punctuation marks are listed in Table 2.1. In formal English writing, punctuation[5] marks mainly function as a division of the text into sentences and clauses. However, for text in most sentiment

[4]Source: https://docs.python.org/3.5/library/string.html Last visited Mar 5^{th}, 2019.
[5]Source: https://en.wikipedia.org/wiki/Punctuation_of_English Last visited Mar 5^{th}, 2019.

datasets, punctuation marks can be used in many different scenarios. For example, the brackets commonly appear in emoticons, and the pound sign is commonly used in hashtags. Generally, there are no guidelines on how to deal with punctuation marks. In practice, one way of treating punctuation is to keep them as lexical items and include them in bag-of-words features (Pang et al., 2002). Some studies treat punctuation marks similar to white spaces and use them to tokenize text (Pak and Paroubek, 2010).

Punctuation Marks	!	”	#	$	%	&	’	(	)	*	+	,	-	.	/	:
	;	<	=	>	?	@	[	\	]	∧	_	‘	{	|	}	~

Table 2.1: Python's built-in string punctuation set

- In linguistics, negation is used to expresses falsity. As the negation is mainly used to reverse the polarity of expression, many recent papers find that removing negation during text pre-processing can harm model performance (Pang et al., 2002; Pak and Paroubek, 2010; Taboada et al., 2011; Agarwal et al., 2011). For instance, Pang et al. (2002) found in a preliminary experiment that the model accuracy after removing the negation is on average slightly lower, so they add a "NOT_" tag to every word between a negation word (such as "not", "isn't", and "didn't") and the first punctuation mark following the negation word. Similarly, some studies replace all negators by tag "NOT" (Agarwal et al., 2011).

2.2.3 N-grams

The n-gram[6] in computational linguistics is a contiguous sequence of "n" items from a given document, where an item is mostly referred to as a word in sentiment analysis. For most sentiment analysis in computational linguistics that builds machine learning classifiers, commonly used "n" values are 1, 2, and 3. N-grams are called as unigrams, bigrams, and trigrams when "n" equals to 1, 2, and 3 respectively. The generation mechanism for n-grams is shown in Figure 2.1. In terms of the prediction power, some studies find that occurrence statistics of unigrams are more predictive than that of bigrams (Pang et al., 2002; Go et al., 2009), but other studies find bigrams are more predictive than unigrams and trigrams (Pak and Paroubek, 2010).

[6]Source: https://en.wikipedia.org/wiki/N-gram Last visited Mar 5^{th}, 2019.

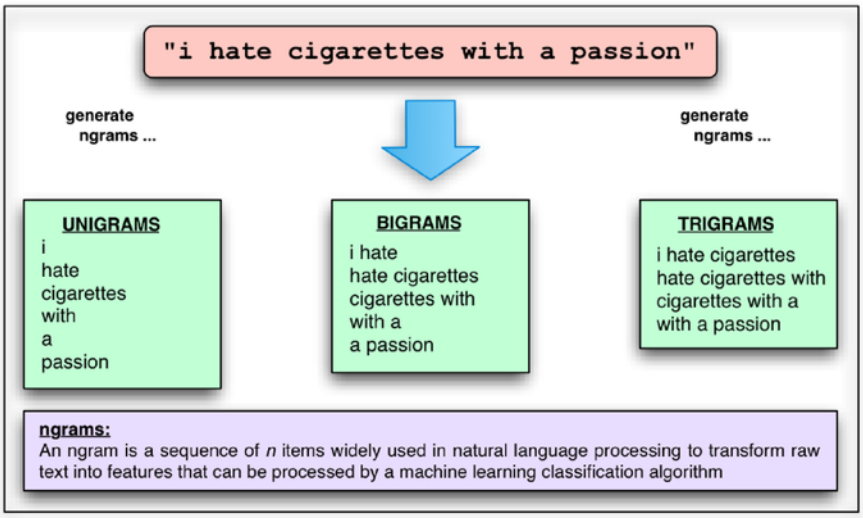

Figure 2.1: Illustration of n-grams (n=1,2, and 3) generated through text (Mysln et al., 2013).

2.2.4 Term frequency-inverse document frequency (Tf-idf)

The Tf-idf[7] is a statistic defined in information retrieval, and it consists of two weighting components of n-grams: Tf and Idf. The Tf assigns higher scores to terms that appear frequently in a document. The Idf is introduced to penalize terms that occur too often in all documents, and it is calculated as:

$$idf_t = \log \frac{N}{df_t} \tag{2.1}$$

where N is the total number of documents in a corpus, and df_t is the Df of term t — the number of documents that contain term t. Then Tf-idf of term t is computed as:

$$tf - idf_t = tf_{t,d} \times idf_t \tag{2.2}$$

where $tf_{t,d}$ denotes the Tf value of term t in document d (number of occurrences of term t in document d), and idf_t is the Idf value of term t. Let $t_1, t_2, t_3, ..., t_m$ represent

[7]Source:https://nlp.stanford.edu/IR-book/html/htmledition/term-frequency-and-weighting-1.html Last visited Mar 5^{th}, 2019.

all n-grams in the corpus, then a single document d can be represented as (or vectorized into) a document vector $\mathbf{d} = (tf - idf_1, tf - idf_2, ..., tf - idf_m)$, where $tf - idf_i$ is the Tf-idf value of t_i in document d. Because the exact order of terms in a document does not affect its document vector, models using Tf-idf vector representation of documents are usually regarded as bag-of-words models. Models using Tf or binary presence vector representation of documents are also regarded as bag of word models due to the same reason.

There is some sentiment analysis that compares the predictive power between Tf-idf, Tf, and binary presence document vectors, and they all find that machine learning classifiers using Tf-idf document vectors perform slightly better than Tf and binary presence document vectors (Socher et al., 2011; Zhang et al., 2015).

2.2.5 Sentiment Classification Models Using Bag-of-words Features

Sentiment classification models are models trained to predict from a set of discrete sentiment classes based on new input. For binary sentiment classification models, there are two sentiment classes — positive and negative. For all records in data that are used in training sentiment classification models, there must be a set of input features and a sentiment label. Specifically, for sentiment classification models that use bag-of-words features, the input features are document vectors that use frequency-based or presence-based statistics of n-grams.

Common machine learning algorithms used in sentiment classification studies include Naive Bayes, logistic regression, and SVMs. We find that the performance of those algorithms is usually dependent on data and tasks. Pang et al. (2002) tested three text classification models — Naive Bayes, maximum entropy classification, and SVMs, and they concluded that SVMs tend to output the best accuracy. In another sentiment analysis paper by Pak and Paroubek (2010), it was found that the multinomial Naive Bayes performs better than SVMs classifier. Experiments in Thelwall et al. (2010a) demonstrated that a simple logistic regression gives a better accuracy score than Naive Bayes and SVMs classifier.

Despite the simplicity of model design, many studies find that bag-of-words sentiment classification models appear to be a strong baseline in terms of both model accuracy and model interpretability, even compared to deep learning models. For example, the bag-of-words logistic regression model in Murdoch et al. (2018) scored a 5.7% error rate on the Yelp review polarity dataset, while their deep learning model scored an only 1.1% lower

error rate. Moreover, the logistic regression coefficients of terms were treated by them as a gold standard to evaluate the effectiveness of all interpretation techniques.

2.2.6 Fine-grained Sentiment Analysis

Early sentiment classification work typically only uses binary sentiment classes: positive or negative (Pang et al., 2002; Go et al., 2009). Later sentiment classification work increases the binary labels by adding a strength component to each class. Although researchers claim those multinomial classes are fine-grained sentiment classes, we think they are still quite different with fine-grained emotions. The reason is that the difference among fine-grained sentiment labels is mainly concerned with the intensity of positiveness and negativity, while fine-grained emotions usually involve more complex dimensions. For example, the relationship between "strongly negative" and "negative" is very different from that of "angry" and "sad". Pang and Lee (2005) increased their number of sentiment classes from the previous positive and negative to a four-point rating scale (that is 0,1,2,3, increasing in positivity). Thelwall et al. (2010a) classified posts into a binary category - positive and negative sentiment, with each sentiment having an additional 5-point strength scale.

Even though those fine-grained sentiment classes are simpler than real fine-grained emotions, it is shown that the model difficulty increases when there are more sentiment classes. Pang and Lee (2005) conducted a manual pilot study and found that the human accuracy decreases from 100% to 69% when the difference in the ratings of the studied pairs changes from 1.5-stars apart to 0.5-star apart (on a 5-star scale).

2.2.7 Word Vectors and Deep Learning Models

In traditional NLP , words are represented as discrete symbols. If we take the example sentence — "i hate cigarettes with a passion" — in Figure 2.1 and assume it is the only sentence in data, words can be represented by one-hot vectors:

$$\text{hate} = [0\ 1\ 0\ 0\ 0\ 0]$$

$$\text{passion} = [0\ 0\ 0\ 0\ 0\ 1]$$

where the dimension of the vector is equal to the number of words in the whole vocabulary (6 in this case). One major problem of representing words by one-hot vectors is that

word similarity could not be captured. Another problem is that a large document could result in a high vector dimension.

The word2vec (Mikolov et al., 2013) framework solves those problems by training a neural network to learn low-dimension and dense word vectors[8] through context words.

Using public pre-trained word vectors to vectorize text, some sentiment analysis builds deep learning models and achieves better model accuracy than n-gram machine learning models (Socher et al., 2011; Zhang et al., 2015). Zhang et al. (2015) developed a character-level convolutional networks model, and found that their model achieved better accuracy on Amazon review dataset than the multinomial logistic regression model using Tf-idf of n-grams. However, their model was outperformed by the logistic regression using bag of n-grams in a binary sentiment classification task on Yelp reviews.

2.2.8 Sequence Models in Sentiment Detection

Because occurrence-based (such as Tf and Tf-idf) document representations are not affected by the exact order of n-grams in documents, models using those n-gram features are usually regarded as bag-of-words models. As the word order is essential in understanding text, many researchers in computational sentiment analysis argue that traditional bag-of-words approach encodes text by term frequency of n-grams, and those vectors often cannot properly capture differences in antonyms as well as word orders (Socher et al., 2011).

Some work includes word sequence information into models by parsing text into trees. It is found that those tree-based models can outperform the bag-of-words machine learning models (Socher et al., 2011) but sometimes do not outperform bag-of-words deep learning models. Socher et al. (2013) introduced the RNTN and discovered that it achieved higher accuracy in sentiment prediction than Naive Bayes with bag-of-bigram features. However, the accuracy was outperformed by the simple one-layer CNN model of Kim (2014), which did not consider word order information.

Besides parsing text into trees, another way to include word sequence information is through RNN or its gated variants (LSTM and GRU). Using RNN in sentiment analysis is found to improve both model accuracy and visualization capability. On recent paper from Bradbury et al. (2016) developed the QRNN model, which leveraged the parallelism

[8]Contents regarding word vectors is based on lecture notes and slides of CS224 at Stanford Source: https://cs224d.stanford.edu/lecture_notes/notes1.pdf and http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture01-wordvecs1.pdf and http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture02-wordvecs2.pdf Last visited Mar 5^{th}, 2019.

feature of CNN and reduced the training time of standard RNN. Their experiment showed that the QRNN achieves slightly higher accuracy than LSTM, and it is three times faster. Moreover, because of the elementwise nature of the recurrent pooling function, they showed that the sentiment changes at each time step could be visualized by the hidden state vectors of the final QRNN layer. The visualization capability breaks the traditional limitation of deep learning models where the prediction process is kept in a "black box".

2.2.9 The DMN Model

Initially developed by Kumar et al. (2016), the DMN model is a Question-Answering model with a special attention and memory module, which simulates the human reasoning process of retrieving relevant information from inputs. After model adjustment, the DMN model can be applied in different types of linguistic tasks. For example, by replacing the question input with "What is the sentiment?", Kumar et al. (2016) trained the DMN model on a fine-grained sentence-level sentiment classification task, with the attention trained in an unsupervised way. The DMN achieved a 4% higher accuracy than that of the previously state-of-the-art model in Kim (2014). They also found that two iterations of memory updates are sufficient for their sentiment classification task, as one iteration is insufficient for reasoning and three iterations tend to overfit.

On top of its high sentiment classification accuracy, the DMN model also demonstrates high interpretability because it utilizes sequential encoders — RNN and its variants — for input representation. Unlike bag-of-words models, DMN keeps the word order information after text encoding, so attention gates can directly relate to the position of original sentences in text. As words with high attention gates have higher weights to determine the final answer, important words that are relevant to questions can be identified based on their attention (Figure 1.2).

Given DMN's good performance in sentiment classification, we develop an approach that uses the model to extract emotion triggers from our data: first, we train a sentiment classification model on our data; then, we re-use the trained model to identify trigger sentences by using new questions that ask for the reason of feeling a particular emotion.

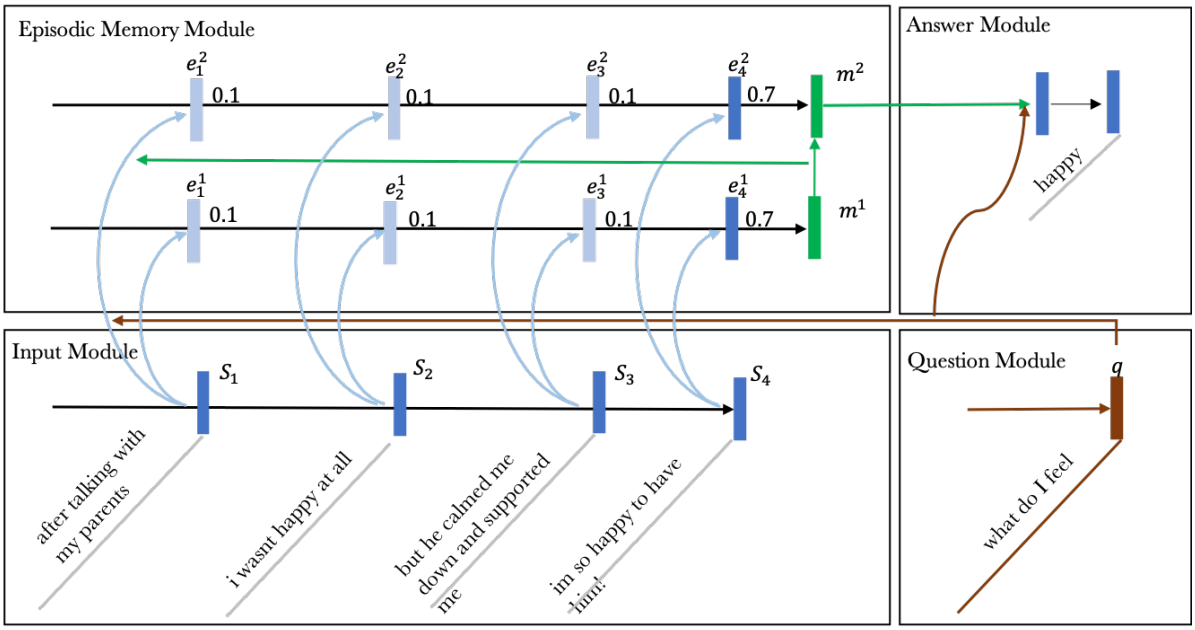

Figure 2.2: An adapted example of the DMN model according to Kumar et al. (2016). For a synthetic mood journal with 4 sentences, attention gates (g_t^i) are triggered by the question — "what do i feel". Gate values are shown above the corresponding vectors. The final state of the episodic memory is the input to the answer module, which generates the predicted emotion.

Figure 2.2 shows that there are four main modules in DMN: 1) the input module, 2) the question module, 3) the episodic memory module, and 4) the answer module. The models are defined as follows:

1. **Input Module**: Given a text input that consists of a sequence of T_C sentences and T_I words, the input module encodes each word by sequential encoders — GRU. The word vector of word t is calculated by: $x_t = L[w_t]$, where L is the word embedding matrix and w_t is the corresponding word index of word t. The hidden state at each time step t (h_t) is:

$$h_t = GRU(x_t, h_{t-1}), t = 1, 2, 3, ..., T_I \tag{2.3}$$

   In the sentence mode, the final outputs of the input module are hidden states at the end of all sentences: C_t ($t = 1, 2, 3, ..., T_C$).

2. **Question Module**: The question module also encodes a sequence of T_Q words in a question using GRU:

$$q_t = GRU(L[w_t^Q], q_{t-1}) \tag{2.4}$$

Each word is represented by q_t, and the final output of the question module is q_{T_Q}, which is the hidden state at the end of the question.

3. **Episodic Memory Module**: The initial memory m^0 is first set to q_{T_Q}, which is the final output of the question module. During each memory episode i, the attention mechanism generates a score Z_t^i for sentence t based on the outputs of the input module — C_t, the previous memory m^{i-1}, and the question module — q_{T_Q}:

$$Z_t^i = G(C_t, m^{i-1}, q_{T_Q}) \tag{2.5}$$

The function $G(C_t, m^{i-1}, q_{T_Q})$ is defined as:

$$W^{(2)} tanh(W^{(1)} z(C_t, m^{i-1}, q_{T_Q}) + b^{(1)}) + b^{(2)} \tag{2.6}$$

where tanh is the Tanh activation function. $W^{(2)}$ and $W^{(1)}$ are weight parameters (W). $b^{(1)}$ and $b^{(2)}$ are bias parameters (b). The $z(C_t, m^{i-1}, q_{T_Q})$ is defined as:

$$[C_t, m^{i-1}, q_{T_Q}, C_t \circ q_{T_Q}, C_t \circ m^{i-1}, |C_t - q_{T_Q}|, |C_t - m^{i-1}|, C_t^T W^{(b)} q_{T_Q}, C_t^T W^{(b)} m^{i-1}] \tag{2.7}$$

Fed into a $softmax$ function, all Z_t^i scores during the episode i generate the attention gates (g_t^i) of all sentences by :

$$g_t^i = softmax(Z_t^i) = \frac{\exp Z_t^i}{\sum_{j=1}^{T} \exp(Z_j^i)} \tag{2.8}$$

The new input representation of episode i (e^i) is the weighted sum of sentence inputs:

$$e_i = \sum_{t=1}^{T} g_t^i C_t \tag{2.9}$$

Then, the new memory at the episode i (m^i) is computed as:

$$m^i = GRU(e^i, m^{i-1}) \tag{2.10}$$

4. **Answer Module**: For sentiment tasks that only require a final output, the answer module is triggered at the end of the memory iterations. The input to the answer module (a_0) is the latest (T_M) memory output in the episodic memory module. Then a simple one layer dense network and a $softmax$ layer are used to determine a "probability" score of each emotion class (y_0).

$$a_0 = m^{T_M} \tag{2.11}$$

$$y_0 = softmax(W^{(a)} a_0) \tag{2.12}$$

Two types of training are used to train the DMN model:

- Training both attention and answer in a supervised way: This method leverages two types of labels in training data: the location of sentences that contain supporting facts and the actual answer. The loss function is defined as follows:

$$J = \alpha E_{CE}(Gates) + \beta E_{CE}(Answers) \tag{2.13}$$

- Training answer in a supervised way but attention in an unsupervised way: This method only uses one type of labels in training data — the actual answer. The loss function is defined as follows:

$$J = E_{CE}(Answers) \tag{2.14}$$

In Equation 2.17, α and β are hyperparameters that determine the relative importance of two cross-entropy cost terms. E_{CE} in both Equation 2.17 and 2.18 denotes the standard cross-entropy cost.

Taking the example in Figure 2.2, the DMN first encodes the four sentences in text to C_1, C_2, C_3, and C_4. Then, it encodes the question to q_4 (there are 4 words in the question). Given the outputs of the input module (C_1, C_2, C_3, and C_4) and q_4, the first memory episode assigns an attention gate of 0.7 to the fourth sentence, and an attention gate of 0.1 to the remaining sentences. At the second memory episode, the attention mechanism also assigns an attention gate of 0.7 to the fourth sentence, and an attention gate of 0.1 to the remaining sentences. In the end, an answer of "happy" is generated based on the memory states of the second episode and the output of the question module.

2.3 Emotion Analysis in Computational Linguistics

During the past ten years, many researchers have started to realize that the popularly studied sentiment in computational linguistics is different from emotions in psychology. Thereby, many studies begin to emphasize the importance of analyzing emotions rather than the sentiment of customer reviews. Despite the change of data sources and research focus, many studies still subject to an issue similar to computational sentiment analysis — lack of model interpretation. Therefore, to date, there has been little work in computational linguistics that provide aggregate insights into emotions and emotion triggers.

2.3.1 Emotion Data Sources and Emotion Labels

While the sentiment datasets are collected from user review sites, such as IMDB, Yelp and Amazon, the data used in computational emotion analysis are collected from mood-specific online sites. For example, Mishne (2005) collected research data from Livejournal, which is an online blogging platform for users to express mood status. For general purpose sites like Twitter, where users can post any topics, most research focuses on specific clinical conditions or emotions by using a list of keywords to extract text that contains those words (Mohammad, 2012; Wang et al., 2012).

Another difference between the sentiment datasets and emotion datasets is in the dimensions of labels. Most emotion studies use fine-grained emotions, while sentiment analysis only uses binary sentiment classes. For example, many emotion studies use a set of emotions that are defined in Psychology, such as Anger, Disgust, Fear, Sadness, Surprise, and Happiness (Mohammad, 2012; Wang et al., 2012; Ghazi et al., 2015).

Although some mood labels in emotion datasets are developed by hashtags in original text, we consider those labels to be less accurate than ground-truth labels that are chosen by writers. It is possible that the matched mood hashtag only represents an initial mood status but not a final mood status that can summarize the whole text. For example, in the text " I ... #sad, but ...", the content follows "but" may express a different emotion than sad.

2.3.2 Emotion Analysis in Computational Linguistics

Although most emotion analysis in computational linguistics uses emotion data sources that are more relevant to human emotions than general sentiment, we consider them to be different from our studies in terms of research goals and methodologies.

Many emotion analysis papers in computational linguistics mainly focus on optimizing model accuracy of detecting emotions but do not study emotion triggers. For example, Wang et al. (2012) applied two machine learning models to predict among 7 emotion labels that were developed using hashtags. They compared models using the binary presence features of either unigrams, bigrams, or trigrams, and found unigram features were generally more predictive than bigrams and trigrams. In another emotion classification work by Mohammad (2012), they built n-gram machine learning classifiers to predict emotions among 6 emotion labels, but they did not explore emotion triggers.

Although some papers involve emotion triggers, they do not aim to obtain aggregate insights into emotion triggers for different types of emotions. Some papers developed complex linguistic rules to extract emotion triggers from text (Neviarouskaya and Aono, 2013; Li and Xu, 2014), but they did not provide insights into potential emotions triggers. For example, Li and Xu (2014) only used the extracted emotion triggers as additional word features to classify emotions. Ghazi et al. (2015) employed an external lexical database to label emotion triggers in text and built supervised models to extract emotion triggers, but still, they did not attempt to analyze emotion triggers. Also, the data used in Ghazi et al. (2015) only contains 820 sentences, which is significantly smaller than us.

2.4 Limitations of Previous Work

Overall, we find that most sentiment or emotions studies in the field of computational linguistics have only focused on optimizing computational methods but not attempted to further study emotion triggers. For previous emotion studies in psychology, they are unable to provide a holistic picture of emotion triggers due to limitations of survey data. As a result, we find that we still know little about emotion triggers, although there have been lots of papers analyzing sentiments or emotions leveraging large-scale social media data. Few related studies can show concrete examples of emotion triggers for different emotions. The lack of findings in emotion triggers has resulted from three main limitations of prior work:

1. As many emotion studies focus on specific emotions or clinical mental conditions, the data that are collected from social media sites rely on a small set of predefined keywords. Hence, we consider that their findings are limited to specific emotions or clinical mental conditions. The keyword lists could be a set of mental disorders (Coppersmith et al., 2015; McIver et al., 2015) or a set of hashtags (Mohammad, 2012; Wang et al., 2012). Some studies collected data only from certain Reddit communities

(Park et al., 2018; Bagroy et al., 2017; Wang et al., 2012). Although these studies can conclude triggers for a specific emotion or mental health condition, we consider the extraction process is problematic if the study aims to reach a conclusion that can be generalized to the public — at least to all social media users.

2. The data used in sentiment classification studies usually lack connection with real human emotions, because their sentiment labels are either too coarse or not self-identified by users. We find that a majority of sentiment studies only use binary sentiment classes — positive and negative. Even though many studies claim to use fine-grained sentiment scales, the difference between those scales only lies in the different level of positivity or negativity (whether it is mild or extreme). However, we think that human emotions are much more complicated. For example, while generally considered as negative emotions, "sad" and "angry" are two very different emotions found in our studies. Without fine-grained emotions, even if we know triggers of negative and positive emotions, the results would be too obvious and too hard to apply back to human emotions to generate meaningful emotion management insights.

3. We find that almost all sentiment analysis and emotion analysis work in computational linguistics does not attempt to interpret emotion triggers. Many studies only focus on getting a high accuracy of predicting sentiment classes but ignore the necessity of providing psychologically meaningful insights about emotion triggers. Even though Bagroy et al. (2017) analyzed the predictive terms and top n-grams after modeling, because their model was not about classifying emotions, they did not conclude any emotion triggers.

As most sentiment analysis primarily focuses on prediction accuracy, few studies attempt to increase model interpretability. Therefore, many models that were developed and used in sentiment analysis are uneasy to interpret, especially deep learning models.

Even though we have shown that the DMN model (Kumar et al., 2016) is promising in terms of interpretable learning, some work suggests that there are major issues with unsupervised attention training, which can result in uniform attention gates for all input facts. Lin and Xiong (2016) compared the attention weights obtained from two types of training options — unsupervised attention training and supervised attention training. We list their results in Figure 2.3. It shows that if the memory module of the DMN model is trained in an unsupervised way, the attention does not shift over different episodes and it is not sparse.

Figure 2.3: Visualization of attention mechanism at each episode in Lin and Xiong (2016). The E1, E2, and E3 in the X-axis represent the first, second, and third memory episodes respectively. The Y-axis represents a sequence of input sentences. (Top) Without supervised gate training, attention shift over episodes is not apparent. (Bottom) With supervised gate training, the attention is sparser and show a significant shift over time.

Even though Xiong et al. (2016) proposed an enhanced DMN model (DMN+) that enables the model to achieve better accuracy when the attention is trained in an unsupervised way, they did not address the issue with uniform attention gates. To our best knowledge, no prior work has provided explanations or solutions to prevent uniform attention layers.

In our study, we also observe this problem of attention when training both the DMN and DMN+ with unsupervised attention training. This attention issue poses critical challenges for us to leverage attention weights to identify emotion triggers. If the attention layer is not sparse enough, we would not be able to use the attention to filter out non-emotion-trigger related sentences, so the DMN model would not help us identify emotion triggers.

Chapter 3

Data and Methodology

In this chapter, we describe our data and methods following the order of steps in Figure 3.1. We first describe the data and all the pre-processing steps. Then, we explain the exploratory data analysis. After that, we show two types of models that are implemented to extract emotion triggers from our data — bag-of-words models using machine learning classifiers and a sequence model (the DMN). We choose these two types of models as they are found to give both good prediction accuracy and high interpretability by prior work. All of the above processes are implemented using Python.

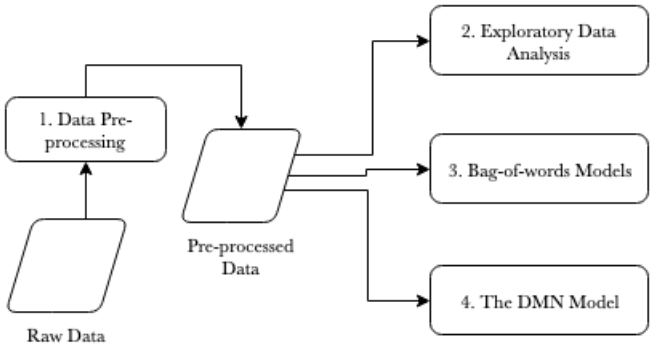

Figure 3.1: Process Flow Diagram

3.1 Data

3.1.1 Data Description

In our study, we analyze a unique journaling dataset containing short pieces of text and associated emotional status self-reported by writers. The dataset was extracted from a mood tracking platform where users can write journals to express their emotions and monitor their emotion trends over time. It covers 18 months — from 01 July 2015 to 31 Dec 2016. For each record in the collected dataset, there are three features: created time, journal text and mood label. The feature description is listed in Table 3.1. The "Journal Text" field stores a short paragraph of text written by users to describe their feelings. The "Mood Label" field contains one of the 17 mood labels that are provided by the platform and is chosen by individual users upon writing journals. The provided mood labels cover a range of basic emotions: *Angry*, *Sad*, *Stressed*, *Frustrated*, *Down*, *Lonely*, *Anxious*, *Overwhelmed*, *Tired*, *Okay*, *Calm*, *Good*, *Productive*, *Accomplished*, *Happy*, *Excited* and *Ecstatic*. By default, the selected mood label is *Okay*.

Feature Name	Feature Description	Data Type
Created Time	The GMT when each journal is created	datetime64[ns]
Journal Text	A non-empty free text input wrote by users.	string
Mood Label	One of the 17 mood labels provided by the journal platform.	string

Table 3.1: Data Description

3.1.2 Data Pre-processing

We first perform data deduplication to rule out the possibility that duplicate data records get overrepresented during analysis. Upon check, we find that several records have identical journal text and mood labels. Since those duplicate records would not provide extra information to our research interest but could cause an over-representation issue, we only keep the first occurrence according to the created time and remove all remaining duplicate records.

After data deduplication, we remove all journals whose mood labels are *Okay*. One primary consideration of this step is that as the mood *Okay* is a default input, we cannot rule out cases where users forgot to change the default mood label after writing journals

to express other emotions. Due to this reason, we decide to remove data with this label all together to reduce potential biases. Another justification of this step is that removing neutral documents or non-subjective documents is a common data cleaning process employed by literature in sentiment analysis (Pang and Lee, 2004; Go et al., 2009). In psychology, Ekman (2003) suggest that the state of no emotion exists, and some emotions could be too slight to be noticeable. We treat the mood label *Okay* as a neutral mood. Upon sampling and manually checking the journals labeled as Okay, we find that most of the time, users are merely logging their daily activities, with no obvious emotions or emotion triggers shown in the text.

Next, we remove journals that contain fewer than two sentences. We define the sentence break as either a comma or a period. Removing journals shorter than a certain threshold is a common way of selecting appropriate text for emotion analysis or sentiment analysis. We choose two sentences as the threshold because we think text contains fewer than two sentences would not be expressive enough for humans or machines to identify mood triggers. More importantly, as we will use a sentence level attention in the later DMN modeling stage to identify emotional triggers, we need at least two sentences for the attention mechanism to be meaningful.

After all the above steps, we have a total of 739,762 English journal records, which are created by 67,522 writers. The average number of words in journals is 28, with the lower quartile as 17 and the upper quartile as 35.

To compare the performance of different models, we randomly split our data into three sets — a training set (80%), a validation set (10%), and a testing set (10%).

To remove irrelevant information in the "Journal Text" and facilitate modeling processes, we apply several text cleaning techniques:

1. We begin the text pre-processing by replacing all whitespace — tab, newline, and extra spaces — with a single space. Since in later modeling processes both of our models do not differentiate paragraphs in documents, and tabs and newlines are only helpful to insert punctuation between paragraphs, we think it is proper to represent them by single spaces.

2. After "normalizing" all whitespace characters, we start to deal with punctuation marks in text. Although some work treats all punctuation marks as single spaces and uses them to tokenize text (Pak and Paroubek, 2010), we find that there are several special punctuation marks function differently in our data. Thereby, we investigate common cases of each punctuation mark and develop a pre-processing rule for different punctuation signs. For example, we find that the pound sign "#" would

typically appear together in a journal, with each hashtag closely followed by a word without spaces. For example, for synthetic terms like "#happy#holiday#party", if we remove all hashtags, all words will be concatenated together. That is, the "#happy#holiday#party" will be converted to "happyholidayparty", which is not an English word. Hence, for "#", it seems reasonable to replace them with a single space, so the term in our example is converted to " happy holiday party". However, it is not the case with the apostrophe mark ('), which is commonly used in possession or contraction. If we replace all apostrophe marks with single spaces, some negations will be lost. For instance, the unigram "don't" will be converted to the bigram "don t"; then, the bigram will be tokenized to "don" and "t" in unigram mode, so the negation becomes harder to interpret than "dont". Therefore, it seems more reasonable to remove apostrophe marks, so certain negated forms are kept. For the comma or the period, we decide to keep them but to pad a single space before and after them, as we use them to identify sentence ends in the later modeling stage (training the DMN). For all remaining punctuation marks, we apply the same treatment as to the comma and the period. Table 3.2 shows different treatments for all punctuation marks.

Treatment	Punctuation Marks
Remove	Apostrophe (') and Backquote (‘)
Replace with a space	Hashtag (#)
Add a space to the right and left	Exclamation mark (!), comma (,), dash (-), period (.), semicolon (;) and question mark (?)

Table 3.2: Treatment for punctuation marks

3. Then, we deal with negations. As shown previously in Section 2.2.2, many recent studies attempt to keep negations during text pre-processing to improve model performance. Also, we observe that the most common way of comprising negations is to add "not" after auxiliary verbs, such as "is not", "shall not", and "do not". Since we remove the apostrophe in the previous step, there are terms like "cant" and "dont" in the text. To aggregate all the remaining negations in the form of "an auxiliary verb + not", we compile a match list and convert all the negations accordingly (See terms before and after the conversion in Appendix A).

4. We convert all letters to lowercase. Converting every letter to lowercase can reduce the variation of words caused by differences in character cases. The lowercase conversion is also necessary as we use publicly trained word embeddings to vectorize text.

As we will use the GloVe by Pennington et al. (2014), where all vocabularies are in lowercase, it is necessary for us to convert our text to lowercase.

There are two standard text cleaning steps that we do not perform in this text cleaning stage. First, we do not perform word stemming and lemmatization. One reason is that we find the stemmed or lemmatized words cannot match publicly trained word embeddings. Another reason is that we find stemmed words are hard to determine their contexts in the later interpretation stage. Second, we do not remove stop words based on curated stop words lists. The work of Nothman et al. (2018) suggests that most curated stop words list are unreliable. They recommend that instead of using any stop words list, it is better to set a threshold of maximum frequency to remove frequent words, so that the removed words would be truly frequent and indiscriminate in the studied data. We follow their recommendation and deal with stop words later in the modeling stage.

3.2 Exploratory Data Analysis

Exploratory data analysis is performed to understand how emotions are distributed and expressed in the data. The results are reported in Section 4.1.

As prior work shows that emotions are subject to circadian cycles and seasonal changes, we want to use our data and further explore the relationships between days of the week and emotions.

We assume that the journals and associated moods in our data represent writers' feeling of that day, as all writers were asked for their current feelings upon creating those journals. Also, as all journals in our data are relatively short — with an average of 28 words, we consider the journals are unlikely to represent writers' long-term mood status.

However, we face a challenge that we only have the GMT time of journals instead of a local time stamp. Also, because our dataset is anonymous, it is impossible to determine the time zone of users. To resolve this issue, we test two ways of local time conversion. We either convert all the GMT time to the Central Time or convert all the GMT time to Eastern Time. We choose these two time zones as we learn that most users in our data live in North America, and the Central Time zone and Eastern Time zone are two zones that have the highest census population size.

Our experiment shows that those two methods of time conversions do not significantly affect the breakdown of emotions by days of the week. Therefore, we choose to convert all the GMT time to Eastern Time.

We also explore the relationship between journal length and expressed emotions as well as the relationship between journal length and days of the week, but find no significant relationships among them.

Additionally, we experiment with three topic modeling techniques to test if topic modeling is an effective approach to extract emotion triggers from our data:

- A simple approach based on n-gram Tf: We study top frequent n-grams for each subset of journals that have the same mood label. For example, for all journals with mood *Sad*, we extract the 5 most frequent unigrams, bigrams, and trigrams that are ranked by the Tf.
- K-means clustering: We first vectorize each text by computing the arithmetic mean of word vectors. Then, we perform k-means clustering on each subset of text journals that have the same mood label.
- Non-negative matrix factorization: We first vectorize all text by computing Tf-idf of unigrams. Then, we use the Non-negative matrix factorization to develop the topic matrix, which is similar to the approach in Toulis and Golab (2017).

However, our preliminary results show that these methods can not help us extract emotion triggers because of two major reasons:

- Although we leverage emotion labels in the modeling process by performing separate topic modeling processes on subsets of text journals that have the same emotion, we do not sufficiently use emotion labels to select relevant information in text that can predict emotions. Without using an emotion classification method before topic modeling, we can only conclude that the topics under a certain emotion are different types of conversations that people talk about, but we can not conclude those topics are emotion triggers for that emotion.
- The topic naming process of those topic modeling methods relies on a manual summary of frequent n-grams in text with the same topic. We find that most frequent n-grams of different topics are quite similar, and do not converge to topics that can be interpreted as emotion triggers. For example, two frequent trigrams for most emotions are "to go to" and "im going to", which are purely syntactic components of sentences.

3.3 Bag-of-words Models

Bag-of-words models in sentiment analysis are models that do not consider the order of words and use mainly bag-of-words features. According to the model performance of previous sentiment classification work, we first experiment with three popular machine learning models to classify emotions: Multinomial Logistic Regression, SVMs, and Random Forest. Then we compare their performance and select the best model for further analysis. Detailed steps are included in Figure 3.2.

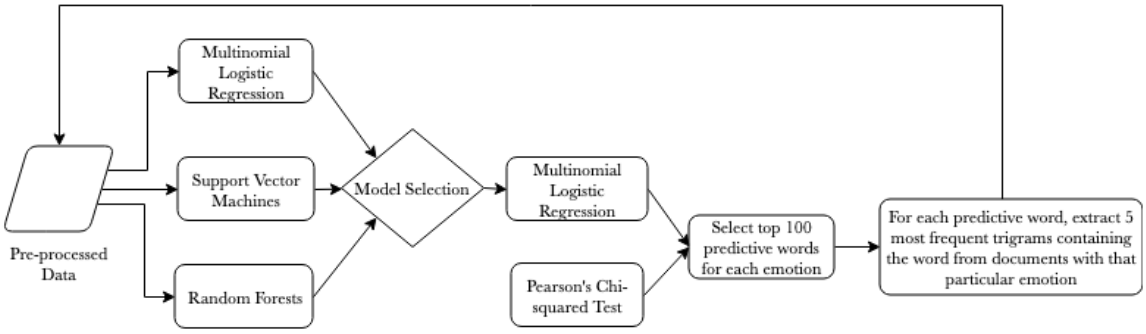

Figure 3.2: Process Flow Diagram of Bag-of-words Models

We use the Tf-idf (Section 2.2.4) value of n-grams (Section 2.2.3) to vectorize all documents, and we choose common n value: 1, 2, and 3. For each document, the input features to machine learning classifiers are represented by the document vector $\mathbf{d} = (tf - idf_1, tf - idf_2, ..., tf - idf_m)$, where $tf - idf_i$ is the Tf-idf value of n-grams in the document. Labels of documents can be one of the 16 emotion labels in our dataset.

3.3.1 Machine Learning Classifiers

Multinomial Logistic Regression: For a classification problem with more than 2 classes, Multinomial Logistic Regression generalizes the binary logistic regression to multiclass problems. The model is optimized trough minimizing the cross-entropy loss or maximizing the posterior probabilities of documents predicted to the correct class.

Linear Support Vector Machines Classification: When there are only two classes and data points are linearly separable in the feature space, the SVMs try to find a hyper-plane that creates the biggest margin between training points of the positive class and

negative class. In our experiments, we employ LinearSVC[9], which uses one vs the rest method in multi-class classification problems. Therefore, our modeling process could be viewed as 16 binary SVMs models, which are trained by the above optimization function.

Random Forests: We also experiment Random Forests as it is a popular method for classification. According to (Friedman et al., 2001), Random Forests can enhance model performance by building a large collection of de-correlated classification trees for classification problems. The final prediction is made by the average vote from those de-correlated trees. Through the process, the model variance is generally reduced.

3.3.2 Using Randomized Search to Optimize Hyperparameter Setting

For each of the above three classifiers, we run RandomizedSearchCV[10] on all of our data to find its optimal parameter settings and best performance.

Besides parameters that are specific to each classifier, we also use the RandomizedSearchCV process to test the best settings to vectorize text features. As suggested by Nothman et al. (2018), setting a threshold of maximum frequency to remove frequent words is a better way of removing stop words than using curated stop words lists. Also, we conduct a preliminary inspection of rarely occurred words and find that those words are mostly misspelled. In order to test the optimal thresholds to remove stop words and misspellings, we include those parameters in the process. Furthermore, we compare prediction power among unigrams, bigrams, and trigrams.

We choose mean 10-fold cross-validation accuracy to represent model's prediction power. The best performance of each classifier is reported in Table 3.3. It is shown that the multinomial logistic regression model generates the best accuracy on our dataset. Therefore, we choose the multinomial logistic regression model for further analysis.

Classifiers	Mean Cross-validation Accuracy
Baseline_1: Random Guess	6.3%
Baseline_1: Vote by the most frequent class	16.0%
Multinomial Logistic Regression	32.8%
LinearSVC	32.4%
Random Forest	29.4%

[9]Source:https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html Last visited Mar 5^{th}, 2019.

[10]Source: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html Last visited Mar 5^{th}, 2019.

Table 3.3: Model Performance for bag-of-words models.

Among the list of hyperparameter values that we pre-defined for multinomial logistic regression:

- N-gram: [using unigrams only, using bigrams only, using trigrams only].
- Maximum document frequency (ignore terms that have a document frequency strictly higher than the given threshold; for example, ignore terms that appear in over 70% of documents): [70%, 80%, 90%, 100%].
- Minimum document frequency (ignore terms that have a document frequency strictly lower than the given threshold; for example, ignore terms that appear in less than 0.001% of documents): [0.001%, 0.006%, 0.010%, 0.014%, 0.019%].

the randomized hyper-parameters search process finds that using the following combination of hyperparameters, multinomial logistic regression can produce the best testing accuracy:

- N-gram: using unigrams only.
- Maximum document frequency: 90%.
- Minimum document frequency: 0.006%.

We do not test hyperparameter combinations of unigrams, bigrams, and trigrams in the above randomized search process because we find that combining different types of n-grams will cause overfitting. For example, in one experiment of multinomial logistic regression, using a combination of unigrams, bigrams, and trigrams altogether generated a training score of 44.8% and a testing score of 33.8%, while using only unigrams generated a training score of 36.2% and a testing score of 32.8%. Although the former setting achieved a 1% higher testing score, it demonstrated a larger extent of overfitting.

Using the above optimal setting for multinomial logistic regression, we fit the model on the training set and the validation set (in total 90% of data), and test the model performance on the testing set. The model achieves a training accuracy of 35.4% and a testing accuracy of 33.5%.

3.3.3 Extracting Emotion Triggers by Interpreting Multinomial Logistic Regression

After fitting a multinomial logistic regression model using the optimal hyperparameter setting, we further interpret the model to understand what types of emotion triggers predict a particular emotion. We perform the following steps to prepare data for manual inspection:

1. We first store logistic regression coefficients in matrix C: $\begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & x_{1n} \\ c_{21} & x_{22} & c_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ c_{K1} & c_{K2} & c_{K3} & \dots & x_{Kn} \end{bmatrix}$, where K (number of emotion classes) is equal to 16 and n is the vocabulary size of our corpus. Using each word in the vocabulary as the vector dimension and the coefficients of that word corresponding to a particular emotion, we can represent all emotions in the column space of C.

2. Next, we perform the Pearson's chi-squared test — a feature selection[11] technique to remove features that are most likely to be independent of class and therefore irrelevant for classification — to check whether each word in our vocabulary is statistically independent of the emotion class. The matrix of observed frequencies is O: $\begin{bmatrix} o_{11} & o_{12} & o_{13} & \dots & o_{1n} \\ o_{21} & o_{22} & o_{23} & \dots & o_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ o_{K1} & o_{K2} & o_{K3} & \dots & o_{Kn} \end{bmatrix}$, where each column represents the observed frequencies of word j. Matrix element o_{ij} denotes sum of Tf-idf value of word j among all documents with emotion i. The corresponding matrix of expected frequencies is E: $\begin{bmatrix} e_{11} & e_{12} & e_{13} & \dots & e_{1n} \\ e_{21} & e_{22} & e_{23} & \dots & e_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ e_{K1} & e_{K2} & e_{K3} & \dots & e_{Kn} \end{bmatrix}$, where each column represents the observed frequencies of word j. Matrix element e_{ij} is the product of percent frequency of emotion i and sum of Tf-idf value of word j among all documents. For each word, we calculate its chi-squared statistics and p value based on its observed frequencies and expected frequencies. A p value larger than a critical point (for example, 0.05) is commonly interpreted as there is not sufficient evidence to reject the null hypothesis that the

[11] Source: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html#sklearn.feature_selection.chi2 Last visited Mar 5^{th}, 2019.

feature is independent of the class variable. In other words, a "large" p value means that we cannot conclude the word is statistically related to our emotion class.

3. We select words that are dependent on our emotion class by setting a significance level at the 0.005: we select words whose p value from the above chi-squared test is smaller than 0.005. This step removes words that are unlikely to be dependent on our emotion class and therefore irrelevant for emotion classification. The coefficient matrix C is reduced to C_p: $\begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & x_{1p} \\ c_{21} & x_{22} & c_{23} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ c_{K1} & c_{K2} & c_{K3} & \dots & x_{Kp} \end{bmatrix}$, where K (number of emotion classes) is equal to 16 and p is the number of significant words (2,131). Thereby, each emotion is represented as a vector with a dimension of 2,131.

4. Then, for each emotion, we select the top 100 most predictive words based on their logistic regression coefficients. For example, for the 1^{st} emotion, we sort the list $[c_{11}, c_{12}, c_{13}, \dots, c_{1p}]$, take the indices of the top 100 largest elements, and then extract the corresponding words from the vocabulary dictionary using the indices.

5. As words generally do not reveal their contexts, we further employ an automated process to study commonly used contexts of a particular word. For example, for the word "hate" that appears in the selected word list of emotion *Angry*, we concatenate all journals with mood *Angry* to one single document, then from that document, we extract 5 most frequent trigrams that contains the word "hate". We perform this automatic process for all the 1,600 predictive terms and store the extracted trigrams to help us understand the contexts of each predictive term.

After the above steps, manual inspection is used to inspect the above predictive terms. Leveraging the trigrams, we summarize all predictive terms into meaningful categories of emotion triggers. We report all related findings in Section 4.2.3.

3.4 The DMN model

As it is discussed in Section 2.2.9, we choose the DMN model to extract emotion triggers because of its high accuracy and high interpretability. Figure 1.3 demonstrates that the DMN model can focus on relevant words in predicting the correct sentiment.

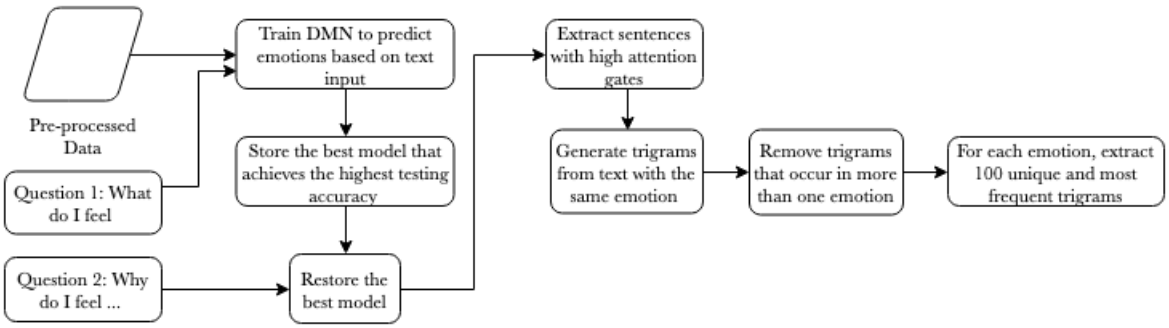

Figure 3.3: Process Flow Diagram of the DMN Model

We first train the DMN model on our dataset to classify emotions using the same question for all input text — "What do I feel", and then re-use the trained attention layer by feeding the trained network with journal inputs and associated questions that ask why the writer feel the associated emotion. For example, for a journal that is labelled as "Sad", we use the question — "why do I feel sad". After that, attention gates are stored for further analysis. Main steps in our DMN modeling processes are illustrated in Figure 3.3.

As our data do not have location labels of sentences that contain emotion triggers, we train the attention module in an unsupervised way. Similar to what was observed in work from Lin and Xiong (2016), we find that our trained model assigns uniform attention gates to all inputs. This problem greatly reduces the interpretability advantage of DMN as the attention weights are uniformly distributed across all sentences in a journal. To our best knowledge, there is no existing solution to the issue.

In our study, we investigate the cause of uniform attention gates in unsupervised attention training, and we present the results in Section 3.4.1.

3.4.1 The Fallout of the L2 Regularization

During the process of training the DMN on our dataset to predict emotions, we observe that the attention layer tends to reach a status where all the attention gates are equal for every sentence, long before the model is fully trained and reaches the best testing accuracy. Also, we find that the attention barely shifts between different episode iteration. The differences between attention gates are compared in Figure 3.4.

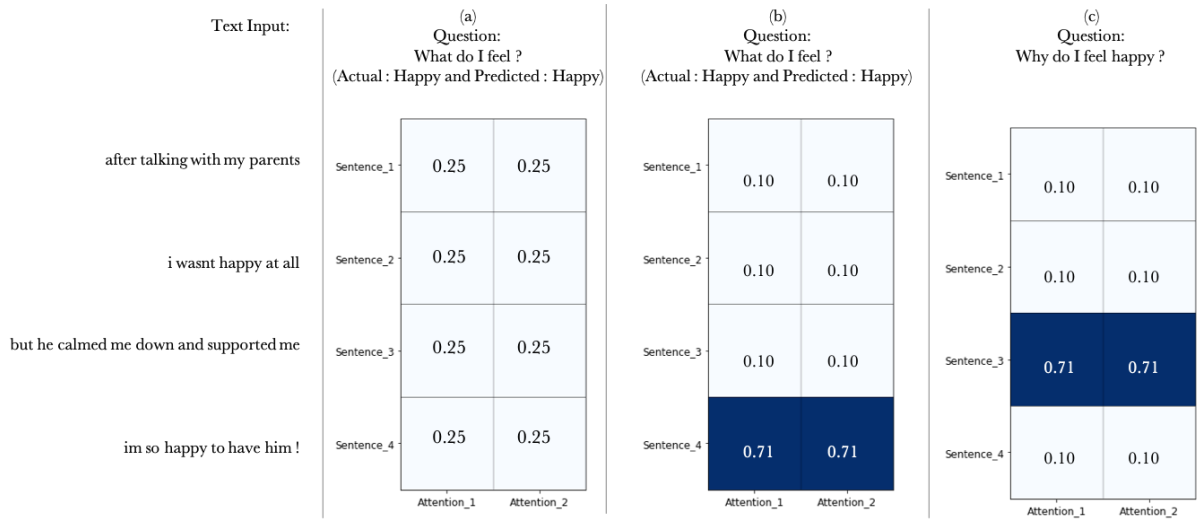

Figure 3.4: Visualization of unsupervised attention gates for one synthetic journal example. (a) The DMN model generates uniform attention gates for all input sentences when using the L2 regularization on parameters in the attention module. (b) Attention gates in sentiment classification setting. The model is trained using our proposed regularization method — maximizing the variance of attention gates. (c) Reuse model trained in (b) and replace the question with "why do I feel happy".

Part (a) of Figure 3.4 illustrates the issue with equal attention. We demonstrate that when using the current L2 regularization on parameters in the attention module, the DMN assigns an equal attention gate (0.25) for all the four input sentences. The same attention issue is observed for all journals in our data, starting from the 2^{nd} epoch during training.

We suspect that training unsupervised attention layers rather than supervised attention layers is a major reason for uniform attention gates. Upon investigation, we find another reason that reduces attention sparsity — the L2 regularization that is used in unsupervised attention training. We discover that the L2 regularization on parameters in the attention model actually "encourages" attention layers to assign equal attention to every sentence.

Although the authors of the DMN model do not publish their code, we find that there are three most popular online implementation methods[12] of the DMN model that all employ the common L2 regularization in unsupervised attention training. Additionally, the

[12]Source: https://github.com/YerevaNN/Dynamic-memory-networks-in-Theano and https://github.com/Steven-Hewitt/QA-with-Tensorflow/blob/master/QA%20with%20Tensorflow.ipynband and https://github.com/barronalex/Dynamic-Memory-Networks-in-TensorFlow Last visited Mar 5^{th}, 2019.

enhanced DMN model by Xiong et al. (2016) specified that L2 regularization was applied to all weights except bias during training. Therefore, we think that all the current DMN training uses the L2 regularization when attention is trained in an unsupervised way, so the training process of the DMN is to minimize the loss function:

$$J = E_{CE}(Answers) + \lambda \times ||w||^2 \tag{3.1}$$

where E_{CE} denotes the standard cross-entropy cost. λ is a hyperparameter that takes a positive value, and it determines the penalizing strength of L2 regularization. $||w||^2$ is the L2 regularization term of all parameters (w) in the network, which include the weight parameters in the memory module.

Recall the definition of attention gates in Section 2.2.9, we know that all the attention gates are the output of a softmax function, so the sum of attention gates (g_t^i) for all sentences in one journal is always equal to 1 for every memory iteration:

$$\sum_{t=1}^{T_C} g_t^i = 1 \tag{3.2}$$

where T_C represents the total number of sentences and i represents the i^{th} memory episode.

The L2 norm of all attention gates for one journal in the i^{th} memory episode (g^i) is:

$$||g^i||^2 = \sum_{t=1}^{T_C} (g_t^i)^2 \tag{3.3}$$

The variance of g^i is:

$$Var(g^i) = \frac{1}{T_C} \times \sum_{t=1}^{T_C} (g_t^i - \mu)^2 \tag{3.4}$$

where μ denotes the mean of g^i.

Because equation 3.2, we have:

$$\mu = 1/T_C \tag{3.5}$$

Thereby, we have:

$$Var(g^i) = \frac{1}{T_C} \times ||g^i||^2 - \frac{1}{{T_C}^2} \tag{3.6}$$

Given a specific input, T_C — the number of sentences — is fixed. Based on equation 3.6, we find that when we minimize the L2 norm of g^i, the $Var(g^i)$ also reduces.

Based on the above discovery, we find that if we want the attention gates of a journal to be variant when the attention is trained in an unsupervised way, we must maximize the L2 norm of g^i rather than minimize it. On the contrary, if we minimize the L2 norm of g^i when the attention is trained in an unsupervised way, the attention module will assign all sentences an equal weight rather than only assign high weights to important sentences. This finding is contradictory to the current practice of minimizing the L2 norm of g^i.

3.4.2 Model Adjustment

In terms of the network structure, we use the original input module, question module, and episodic memory module of the DMN model (Section 2.2.9), and only adjust the answer module. Upon obtaining the memory of the last episode T_M, we concatenate it with the final hidden state of the question q_{T_Q}. The input to the answer module is:

$$a_0 = [m^{T_M}, q_{T_Q}] \tag{3.7}$$

Next, the input is fed into a dense neural network layer. Then, the output of that layer is used to calculate a Euclidean distance score d_k (k = 1,2,3..., 16) for each of the mood classes:

$$d_k = -|W^{(a)}a_0 - L[w_k^E]|^2 \tag{3.8}$$

w_k^E stands for the word index of emotion class k. In the end, we pick the mood label whose distance score is the smallest as the answer output.

3.4.3 Training Adjustment

Similar to the training settings in the original DMN paper, we train the network via backpropagation and Adam Optimizer, and we use pre-trained word vectors — GloVe. We set the word dimension and GRU's hidden layer size to 100. We do not choose a higher word dimension as we find in our experiments that a higher word dimension result in a much longer training time but a similar accuracy. We set the number of memory passes to 2, as suggested by Kumar et al. (2016) that 2 passes outperform a single pass and zero passes.

To address the issue with L2 regularization (Section 3.4.1), we tested three methods to encourage a sparse attention layer through changing the loss function, with the goal of choosing the most effective method:

- Maximizing the L2-norm of all parameters in the memory module by minimizing the total loss function:

$$J = E_{CE}(Answers) - \lambda \times ||w||^2 \tag{3.9}$$

- Maximizing the L2-norm of attention gates in all memory episodes by minimizing the total loss function:

$$J = E_{CE}(Answers) - \lambda \times ||g^i||^2 \tag{3.10}$$

- Maximizing the mean variance of attention gates in all memory episodes by minimizing the total loss function:

$$J = E_{CE}(Answers) - \lambda \times \frac{1}{T_M} \times \sum_{i=1}^{T_M} var(g^i) \tag{3.11}$$

  The λ in Equation 3.9, 3.10, 3.11 is a hyperparameter that takes a positive value and it determines the penalizing strength. E_{CE} denotes the standard cross-entropy cost. T_M is the maximum number of memory iterations in the episodic memory module.

To compare the effectiveness of the above new loss functions, we experiment with several λ values for each method. For each λ value, we train a DMN model with a batch size of 52 and a maximum of 10 epochs. We set the learning rate at 0.001 for all runs. We implement and train the adjusted DMN model using TensorFlow[13] framework. Our code is adapted from several popular online implementation methods (See footnote 16).

During training, the training set is used to train the model, and the validation set is used to validate the model. After the model is fully trained, a final testing score is calculated on the testing set, using the model that achieves the best validation accuracy during training.

For each experiment on regularization methods, we list out the testing accuracy as well as the mean variance of the attention in the second attention episode ($\frac{\sum_1^N Var(g_{n,2})}{N}$, where N is the total number of journals and $g_{n,2}$ denotes the attention gates in the second attention episode of the n^{th} journal) in Table 3.4.

[13]Source: https://www.tensorflow.org/ Last visited Mar 5^{th}, 2019.

Regularization methods	λ	Best Testing Accuracy	Mean variance of attention in the last memory episode	Percentage of journals with variance of attention in last memory episode $\geq$ 0.01
Min L2-norm of all parameters	0.01	34.56%	2.07E-10	0.00%
Max L2-norm of attention gates (2)	0.01	34.55%	1.38E-02	20.53%
Max Variance of attention gates (3)	**1**	**34.54%**	**2.18E-02**	**31.07%**
Max L2-norm of attention gates (2)	1	34.46%	1.89E-02	26.83%
Max Variance of attention gates (3)	100	34.34%	5.35E-03	30.99%
Min L2-norm of all parameters	1	34.29%	4.27E-10	0.00%
Max L2-norm of attention gates (2)	100	34.25%	2.42E-02	34.51%
Max Variance of attention gates (3)	0.01	34.18%	2.19E-02	7.83%
No regularization	0	34.10%	2.05E-02	28.48%
Min L2-norm of all parameters	100	33.72%	1.93E-12	0.00%
Max L2-norm of all parameters (1)	0.01	33.41%	2.63E-02	37.46%
Max L2-norm of all parameters (1)	1	33.00%	3.09E-02	43.41%
Max L2-norm of all parameters (1)	100	32.14%	2.64E-02	37.90%

Table 3.4: Performance of the DMN model when using our three proposed regularization methods. Results are sorted by the best testing accuracy.

Based on Table 3.4, we demonstrate that our proposed regularization methods significantly improve the variance of attention gates even when we also train the attention in an unsupervised way. Using the third proposed method — maximizing the mean variance of attention gates in all memory episodes and setting λ to 1, the model can achieve the same accuracy (only 0.01% lower) but significantly more variant attention gates (See the row in bold in Table 3.4). We find that under that setting, 31.07% of journals can pass the 0.01 variance threshold. However, there is no journal whose variance of attention is greater or equal than 0.01 when using the L2 regularization in prior implementation — minimizing the L2-norm of all parameters in the memory module.

Part (b) in Figure 3.4 further shows the difference in attention layer through the synthetic example: compared to the Part (a), the attention mechanism assigns higher attention gates for the fourth sentence, which helps the model to give a correct prediction.

3.4.4 Extracting Emotion Triggers by interpreting the DMN Model

Based on the performance comparison in Section 3.4.3, we choose to employ the regularization method — maximizing the mean variance of attention gates in all memory episodes.

During training, we use the same question "What do I feel" for all inputs, and we store all model variables when a batch produces a better testing accuracy than all previous batches. After the training is completed, we restore the model with the best testing accuracy. We use TensorFlow to store and restore models.

To extract emotion triggers, we feed all journals and associated questions ("Why do I feel X", where "X" is replaced with the associated mood labels) into the restored model. Taking the synthetic example in Figure 3.4, we use the question "Why do I feel happy", and the attention mechanism assigns corresponding attention gates to all input sentences (See Part (c) in Figure 3.4).

Based on the attention gates of all journals, we set a variance threshold at 0.01 to filter out journals whose attention gates are not sparse. That is, we do not extract emotion triggers from journals whose variance of attention gates at the last memory episode is less than 0.01. To verify whether those journals truly do not have comparatively important sentences that contain emotion triggers, we randomly select 5 journals for each emotion (80 journals in total), and two academic assessors manually verify those journals. Due to the confidentiality of our data, we do not employ more assessors. The assessors are provided with the journals and associated emotion labels, and they manually check each journal and decide whether there is a sentence that contains obvious emotion triggers for the associated emotion. We report the assessment results in Table 3.5. It shows that the conservative accuracy score — percentage of journals labelled as correct by both assessors — are 81%. Because the percentage of journals that do not contain obvious emotion-trigger-related sentences is relatively high, we think that removing those journals helps remove irrelevant information from text.

Samples that the DMN assigns equal attention gates	Assessor_1	Assessor_2	Assessor_1 or Assessor_2
Percentages of samples that are labelled as wrong	11%	14%	19%

Table 3.5: Assessment result of journals whose attention gates are not sparse

After this filtering step, we have about 230,000 journals that have variant attention gates. Then, for each journal, we extract sentences whose attention gates are higher than

the average attention gates in that journal. For example, we will extract the third sentence of the synthetic case in Figure 3.4, as its attention gate is higher than the average (0.25).

To evaluate whether the DMN model can assign high attention scores to sentences that actually contain emotion triggers, we randomly select 5 journals for each emotion (80 journals in total) from the 230,000 journals. The same two assessors are provided with the journals, associated emotion labels, and the extracted sentences. For each journal, the assessors manually check whether the extracted sentences contain the actual emotion triggers. As shown in Table 3.6, the extracted emotion triggers in 85% of journals are considered as correct by both assessors. As the percentage of correct samples is relatively high, we think our approach is effective in extracting sentences that contain emotion triggers.

Samples that the DMN assigns equal attention gates	Assessor_1	Assessor_2	Assessor_1 or Assessor_2
Percentages of samples that are labelled as wrong	10%	11%	15%

Table 3.6: Assessment result of journals whose attention gates are sparse

We perform the following steps to further interpret those extracted sentences:

1. We first concatenate all sentences with the same emotion to a single document.
2. Next, we extract all trigrams from each document and rank those trigrams by their term frequency in descending order. We choose to use trigrams rather than unigrams or bigrams because we find that trigrams can provide more context information and are easier to interpret.
3. After that, we remove trigrams that appear in more than one emotion. The purpose of removing trigrams that appear in multiple emotions is to remove phrases that do not contain emotion triggers. For example, "feel like im" is a frequent trigram for all emotions, but it is merely a syntactic phrase and does not indicate any emotion trigger.
4. For each emotion, we extract 100 most frequent trigrams.

After all the above steps, manual inspection is performed to study those frequent and unique trigrams. Similar to the manual inspection method in Section 3.3.3, we summarize

all trigrams into categories of emotion triggers, and we report all related findings in Section 4.3.2.

We also tried several topic modeling methods (listed in Section 3.2) over these extracted sentences, but the results are not as good as the above approach. The reason is similar to what is described in Section 3.2: we find that for most topics, the frequent n-grams of different emotions seem to be quite similar and contains much non-emotion-trigger-related content.

Chapter 4

Research Findings

In this chapter, we first report findings obtained from three lines of analysis that we describe in Chapter 3: 1) exploratory data analysis, 2) the selected bag-of-words model — Multinomial Logistic Regression, and 3) the DMN model. These three lines of analysis are highlighted in Figure 4.1. At the end of this chapter, we compare the Multinomial Logistic Regression and the DMN model in terms of their ability to extract emotion triggers from text.

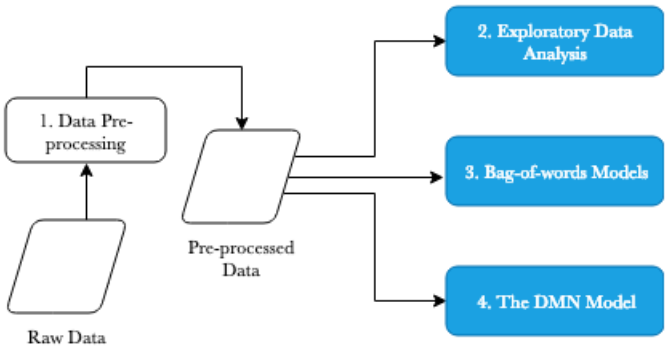

Figure 4.1: Process Flow Diagram with Analysis Processes Highlighted

4.1 Exploratory Data Analysis

4.1.1 Good, Calm, Tired, and Happy are the most frequent emotions in our data.

To explore the distribution of emotions in our pre-processed data, we calculate the percentage based on the number of journals that are labeled by the corresponding emotions. The data is presented in Figure 4.2. Based on their percentage, we group emotions into three tiers:

- The most frequently expressed motions are *Good*, *Calm*, *Tired*, and *Happy*, which constitute about 49% of all emotions.
- The less frequent emotions includes *Sad*, *Down*, *Anxious*, *Frustrated*, *Overwhelmed*, and *Productive*, whose percentage ranges from 4% to 6%.
- Th least frequent emotions are emotions starting from the *Accomplished* to the *Ecstatic*.

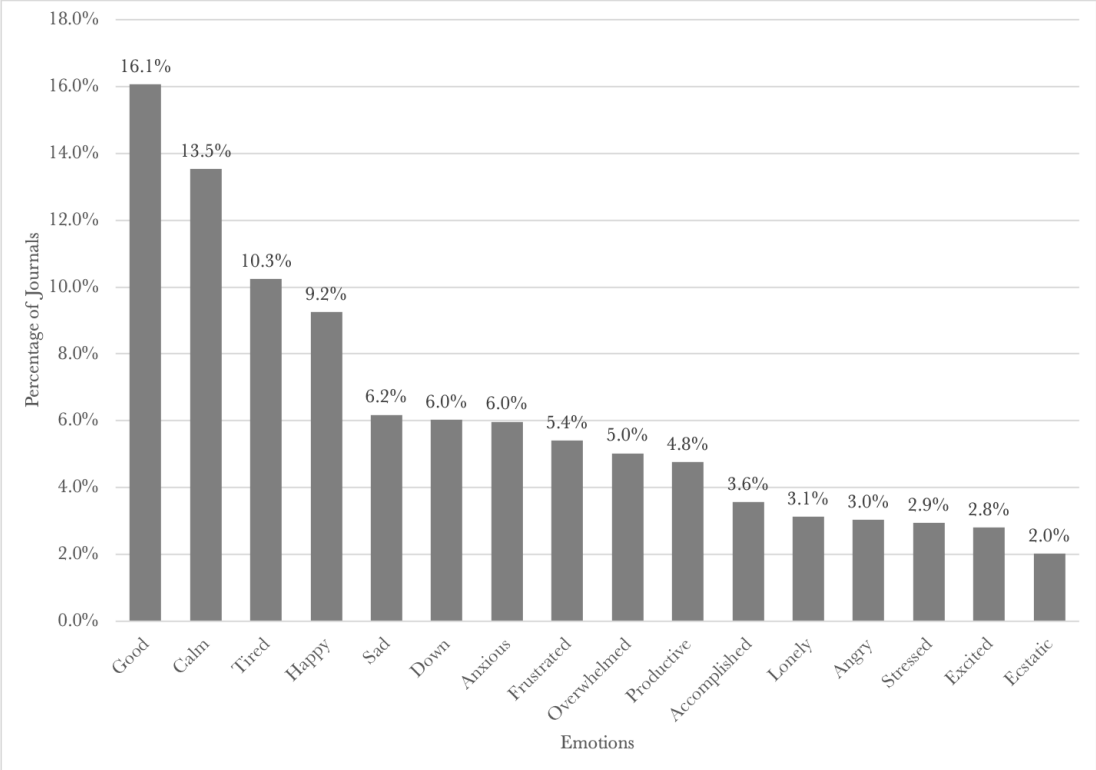

Figure 4.2: Distribution of emotion labels based on number of journals.

4.1.2 Percentages of Happy, Sad and Lonely journals peak on Weekends.

We calculate the daily percentages of emotions by computing the proportion of journals that are labelled with a certain emotion on a day of the week to the total number of journals on that day, and evaluate the percentage trend for each emotion. We find that while the percentages of most emotions do not show obvious change during the week (See Appendix B2), there are five emotions — *Happy*, *Tired*, *Overwhelmed*, *Sad*, and *Lonely* — that fluctuate significantly across days of the week(Figure 4.3).

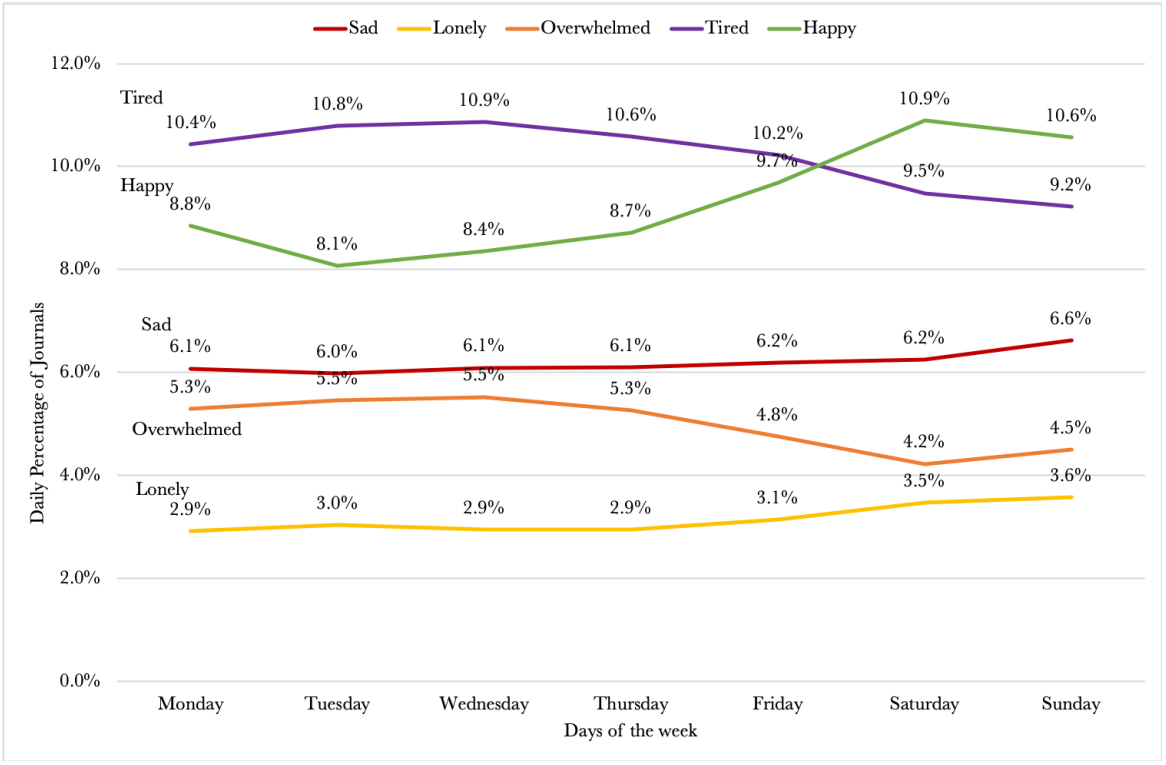

Figure 4.3: Daily distribution of emotions across weekdays and weekends.

We observe that people generally post higher percentages of *Happy* journals and lower percentages of *Tired* and *Overwhelmed* journals on Fridays, Saturdays, and Sundays than other days.

Although people post a higher percentage of *Happy* journals on weekends than weekdays, we see that the percentages of *Sad* and *Lonely* peak on Sundays.

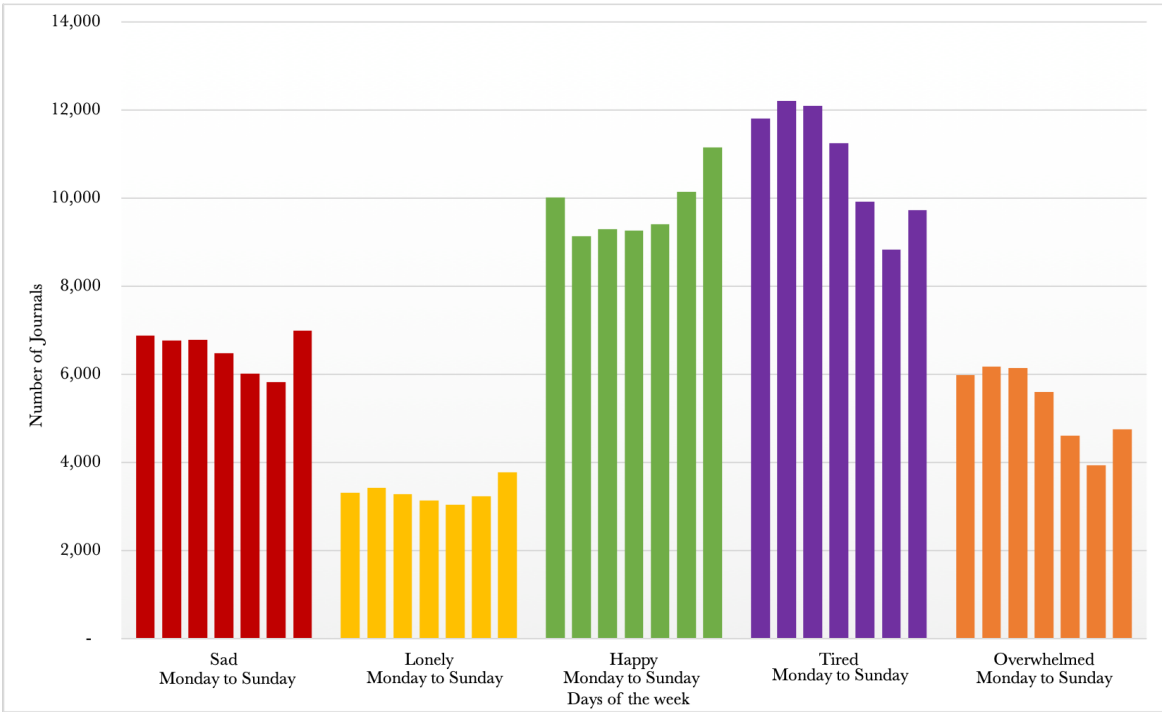

Figure 4.4: Number of journals across weekdays and weekends.

To further investigate whether the actual number of Sad and Lonely journals also peak on Sunday, we plot out the trends of emotions based on the number of journals expressed on a day of the week (Figure 4.4). Although there are fewer journals on Sundays (Figure B1), we find that the actual number of journals to express Sad, Lonely, and Happy also peak on Sundays.

The increase in sadness, loneliness, and happiness on weekends could be partly explained by findings of Dzogang et al. (2017), as they observe that the circadian cycle of these emotions are different between weekends and weekdays. Therefore, we suspect that emotion triggers of sadness, loneliness, and happiness may contain activities that are dependent on the change of time schedules between weekdays and weekends. However, the large decrease in *Tired* on weekends that we observe cannot be explained by their findings, as Dzogang et al. (2017) conclude that the fatigue is resistant to the weekend and weekday changes.

4.2 The Bag-of-words Model — Multinomial Logistic Regression with Unigram Tf-idf Features

In section 3.3.2, we compare testing accuracy — the proportion of true predictions among the total number of testing data — among three bag-of-words models, and find the Multinomial Logistic Regression achieve the best testing accuracy. Thereby, we proceed with Multinomial Logistic Regression for analysis.

4.2.1 Model Evaluation

To further evaluate our Multinomial Logistic Regression model, we compute the confusion matrix to investigate prediction performance for each emotion class (See Figure 4.5).

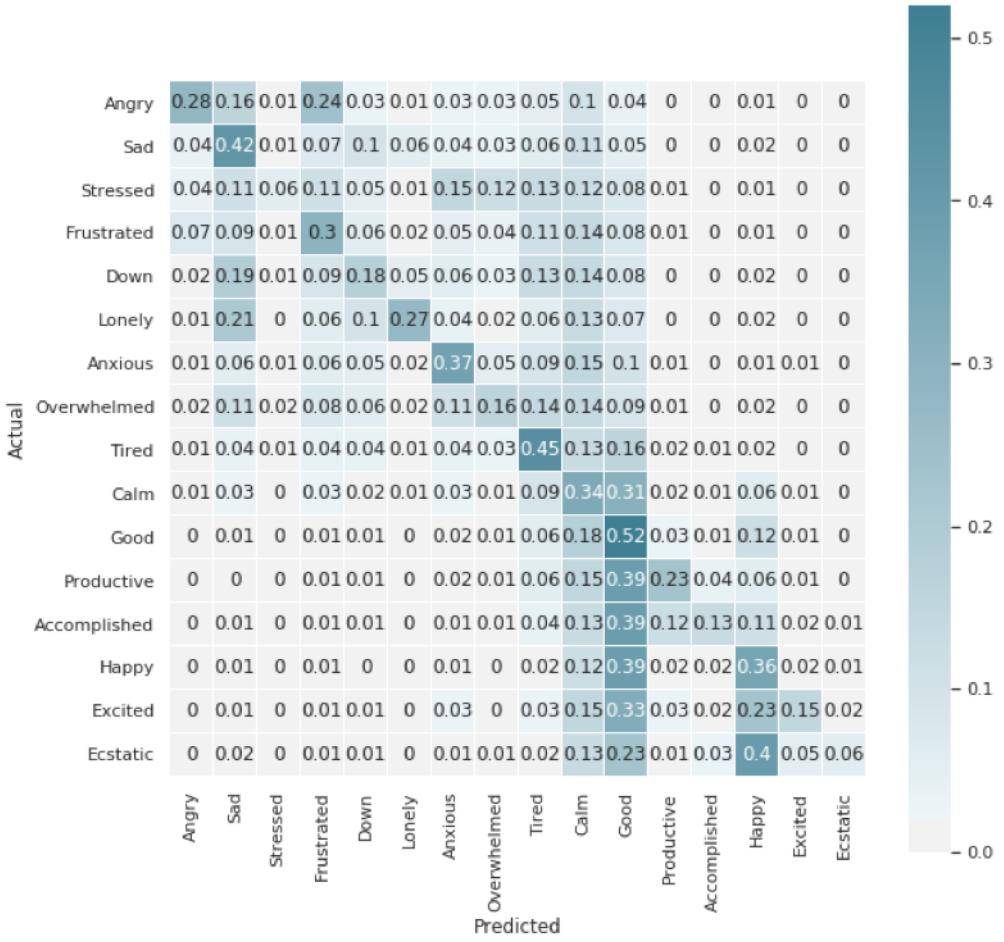

Figure 4.5: Confusion Matrix of Multinomial Logistic Regression Model

It is shown that only *Good* has a accuracy over 50%, and it is even more challenging to predict *Stressed*, *Down*, *Overwhelmed*, *Accomplished*, *Excited* and *Ecstatic*. For example, only 6% of *Stressed* journals are predicted correctly, with 11%, 11%, 15%, 12%, 13%, and 12% of them predicted as *Sad, Frustrated, Lonely, Anxious, Overwhelmed, Tired, and Calm* respectively. We think that there are possibly three reasons that result in low accuracy scores:

- We have a large number of emotions to predict. Previous papers (Pang and Lee, 2005; Kumar et al., 2016) have shown that the model accuracy decreases when there

are more sentiment classes. For example, the DMN model's accuracy in five-class sentiment prediction is 52.1%. Although the datasets we are comparing are different, we think our accuracy is reasonable as we have three times their number of classes.

- Class imbalance is another reason for the low performance of less frequent emotions. As there are more data for *Good*, *Calm*, and *Tired*, Multinomial Logistic Regression will put more weights on classifying those frequent emotions correctly during training.

- High correlation among certain emotions also leads to the misclassification among emotions. For instance, 11% of *Stressed* journals are predicted as *Frustrated*. If these two emotions are indeed correlated, their journals may contain similar content that is hard for classifiers to predict correctly.

4.2.2 Correlation Among Emotion Labels

To further investigate whether there are some emotions that are highly correlated, we explore the intercorrelations among the 16 emotions. We calculate the pairwise correlation of emotions by calculating column-wise Pearson correlation coefficient of the matrix C^T (See definition of C in section 3.3.3): $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, where X and Y represent any two columns of C^T.

We visualize the above calculated correlation data using a heatmap (See Figure 4.6). It is shown that there are two major categories of emotions: a group of positive emotions and a group of negative emotions. We see that emotion *Angry*, *Sad*, *Stressed*, *Frustrated*, *Down*, *Lonely*, *Anxious*, *Overwhelmed*, and *Tired* are all negatively correlated with emotion *Calm*, *Good*, *Productive*, *Accomplished*, *Happy*, *Excited*, and *Ecstatic*. We define negative emotions and positive emotions based on the above two groups.

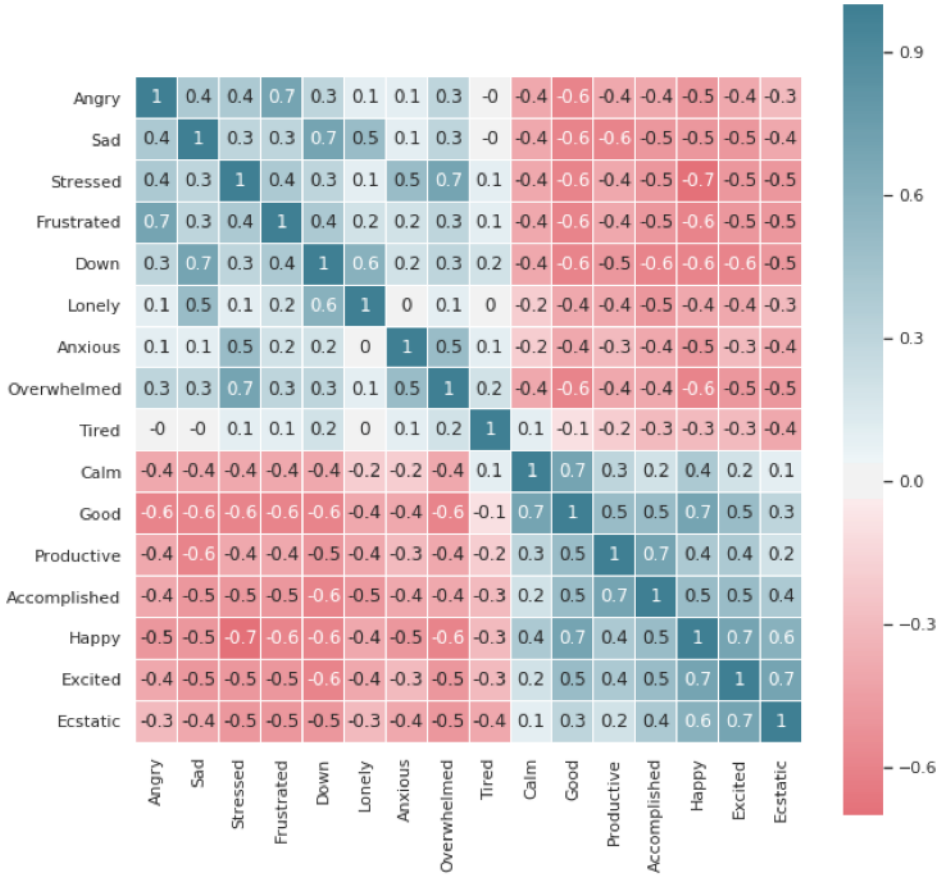

Figure 4.6: Heatmap of Mood Correlation

Additionally, we observe that different moods are correlated with each other in an intricate pattern beyond a simple scale of positiveness and negativity. For example, *Angry* is more correlated with *Frustrated* than *Stressed*. Setting a correlation cut-off at 0.5, we group emotions with a pairwise correlation greater than or equal to the cut-off. In the end, we have six categories of emotions (Table 4.1). Interestingly, we find that *Tired* is not highly correlated with any other emotions; thereby, itself constitutes a category. For the rest of the emotions, we find that: 1) *Angry* and *Frustrated* are highly correlated; 2) *Sad*, *Down*, and *Lonely* are shown to be highly correlated; 3) *Stressed*, *Anxious*, and *Overwhelmed* are highly correlated; 4) *Productive* and *Accomplished* are highly correlated; 5) *Calm*, *Good*, *Happy*, *Excited*, and *Ecstatic* are grouped together.

Categories	Groups (setting correlation cutoff as ≥ 0.5)
1	Angry, Frustrated
2	Sad, Down, Lonely
3	Stressed, Anxious, Overwhelmed
4	Tired
5	Productive, Accomplished
6	Calm, Good, Happy, Excited, Ecstatic

Table 4.1: Correlated Emotions

4.2.3 Emotion Triggers Summarized by Interpreting Multinomial Logistic Regression

As described in Section 3.3.3, we perform model interpretation to extract emotion triggers from text. Specifically, we manually inspect a total of 1,600 predictive terms for all emotions, and group those terms into different types of emotion triggers.

One interesting result is that several predictive terms can predict vastly different emotions. Table 4.2 summarizes those terms. For example, the scenario "job interviews" is highly predictive for emotion *Anxious* and *Excited*, which suggests that while some people feel *Excited* at job interviews, some people feel *Anxious*. The difference between the triggered emotions provides a possibility of emotion regulation according to Gross and Muñoz (1995). If we change our mental environment and focus on the potential opportunities brought by a job interview, we can regulate anxiety towards a positive emotion status.

Categories	Predictive Terms	Scenarios	Predicted Moods
General Work-related	busy	be left out	Lonely
		had a busy day	Tired, Productive
	finish	facing deadlines	Stressed, Overwhelmed
		going to finish	Productive
	interview	job interviews	Anxious, Excited
	presentation	going to give a presentation	Anxious
		gave a good presentation	Accomplished
Sleep-related	slept	not slept well	Tired
		well rested	Calm, Good
	nap	need to take naps	Tired
		well rested	Calm, Good
School-related	essay	facing deadlines	Stressed
		school life, essay done	Productive
	homework	facing deadlines	Stressed, Overwhelmed
		school life, homework done	Productive
	paper	facing deadlines, have a paper due	Stressed, Overwhelmed
		school life, paper done	Productive
Others	passed	death of loved ones	Sad
		school life, passed tests	Accomplished
	packing	packing up	Overwhelmed, Productive

Table 4.2: Predictive terms that could result in vastly different emotions

During the manual inspection, we find that although many predictive terms can reveal certain events or experiences of authors and help us conclude emotion triggers, there are also many terms that do not show specific emotion triggers but purely function as direct emotional expressions. We find that it is due to the existence of three possible types of mood journals (synthetic examples):

- Pure emotional expression: "I am so mad right now."
- Mood triggers: "Today I was made fun of at school again."
- A Mix of emotional expression and Mood triggers: "I am so mad right now. Today I was made fun of at school again."

In terms of emotion prediction, emotional terms make intuitive sense in that their occurrence is a strong signal for specific emotion. Lots of sentiment lexicon libraries contain mostly emotional words to make a prediction (Taboada et al., 2011).

However, those emotional terms do not help us to understand what can actually lead to the emotion. For example, we find that the word "mad" is a significant and highly

predictive term for emotion — *Angry*. A prediction of *Angry* upon the occurrence of "mad" is fairly reasonable, but we cannot know what the author is mad at.

Emotions	Direct expressions of the emotion states	Swearing words	Total
Productive	24%		24%
Accomplished	36%		36%
Sad	40%	1%	41%
Overwhelmed	43%		43%
Happy	47%		47%
Lonely	47%	1%	48%
Tired	50%		50%
Stressed	44%	6%	50%
Excited	52%		52%
Angry	32%	20%	52%
Ecstatic	53%		53%
Anxious	56%		56%
Frustrated	39%	17%	56%
Calm	56%	2%	58%
Good	59%		59%
Down	61%	1%	62%

Table 4.3: Percentage of top 100 predictive terms that only reflect emotion states

Therefore, we do not consider emotional words as emotion triggers and group all emotional words into a category — "direct expressions of emotion states". Another category of words that we do not consider as emotion triggers is swearing words.

As shown in Table 4.3, the top 100 predictive terms for most emotions contain over 50% of words that are not emotion triggers. Emotion *Productive* and *Accomplished* include the least number of emotional words in their predictors. Emotion *Angry* and *Frustrated* both contain a large number of swearing words. The similarity of these emotions is in line with our earlier finding on emotion correlations in Section 4.2.2.

For the remaining predictive terms, we consider them as emotion triggers and further summarize them into meaningful themes. We create a matrix-like map for a high-level overview of emotion triggers we identify from our data (Table 4.4). A cell with a check mark means that the corresponding emotion trigger (row name) can trigger the emotion

(column name). A cell without a check mark indicates that the emotion trigger is not observed to trigger the emotion (column name).

According to Table 4.4, it is found that "Exercise", "Food and meals", "Get things done", "Leisure activities", "Progressing", "Self-recognition", and "Weather" are emotion triggers that only trigger positive emotions.

Grouping emotion triggers based on the categories in Table 4.1, we examine their emotion triggers below. Detailed scenarios for emotion triggers are shown in corresponding cells in the below tables.

Emotion Triggers for Angry and Frustrated

Emotion Triggers	Angry	Frustrated
Had an argument	Arguing with someone, being yelled at	Arguing with someone
Being mistreated	Being lied to, being ignored, being treated disrespectfully/unfairly/badly/like a child/rudely	Being treated inconsiderately/rudely/unfairly
Sickness	Blood pressure problem, pain	Headaches, migraines, pain
Body image	Concerns about body weight	Concerns about body weight, lots of calories intake
Failure	Failed, being unable to handle things	Failed
Self-hatred	Hate oneself	Hate oneself
Personal relationships	Parents, siblings	Parents
Self-disappointment		Being expected to do something, disappoint in oneself, lack of patience
Work		Feeling difficulty, get stuck, have an issue
Infidelity	Being betrayed, being cheated, cheating on partners	
Politics	Concerns about the presidential election	
Self-harm	Hurt oneself, thinking about suicide	

Table 4.5: Emotion Triggers for Emotion Angry and Frustrated Concluded from the Multinomial Logistic Regression

As it is reported in Section 4.2.2, there is a high correlation (0.7) between *Angry* and *Frustrated*. The high correlation among them is demonstrated in Table 4.5, where we summarize their emotion triggers based on the coefficients of the Multinomial Logistic Regression.

We find that "concerns about body weight", "sickness", "having an argument with others", "past failure", "self-hatred", "personal relationships", and "being mistreated" can all trigger *Anger* or *Frustration*, where issues regarding personal relationships mostly focus on parental relationships and sibling relationships.

Emotion Triggers	Angry	Frustrated	Sad	Down	Lonely	Stressed	Anxious	Overwhelmed	Tired	Productive	Accomplished	Calm	Good	Happy	Excited	Ecstatic
Sickness	✓	✓	✓	✓	✓	✓	✓	✓	✓							
Personal relationships	✓	✓	✓	✓	✓		✓	✓				✓	✓	✓	✓	✓
Self-disappointment		✓	✓	✓	✓		✓									
Work		✓				✓	✓	✓	✓	✓						
Politics	✓						✓									
Medications and Treatment							✓									
School life						✓	✓	✓		✓	✓					✓
Sleep				✓			✓		✓			✓	✓	✓		
Had an argument	✓	✓	✓	✓		✓		✓								
Being mistreated	✓	✓			✓											
Body image	✓	✓	✓	✓						✓	✓				✓	✓
Failure	✓	✓	✓	✓		✓		✓								
Self-hatred	✓	✓	✓	✓	✓	✓		✓								
Infidelity	✓															
Self-harm	✓		✓	✓		✓		✓								
Cries			✓	✓	✓			✓								
Financial situation						✓		✓			✓					
Exercises										✓	✓				✓	✓
Food and meals											✓	✓		✓		
Get things done										✓	✓					✓
Leisure activities												✓	✓	✓	✓	✓
Progressing										✓	✓	✓				
Self-recognition										✓	✓		✓		✓	✓
Weather												✓	✓	✓		
Therapy												✓				

Table 4.4: Emotion Trigger Map for all emotions Concluded from the Multinomial Logistic Regression

While sharing common triggers, one unique emotion trigger — "infidelity", which somewhat also belongs to the personal relationship problems — but we list it separately, is found for *Angry* but not for all the other 15 emotions. Moreover, "self-harm" is found for *Angry* but not for *Frustrated*. We think the involvement of hurting oneself and thinking about suicides reveals that the emotion *Angry* may bring more harm to ones' health and safety than *Frustrated*.

Challenges in "work" (such as facing difficulties and getting stuck) and "self-disappointment" can trigger frustration but not anger.

Overall, we find that emotion triggers for *Angry* and *Frustrated* mostly associated with certain conflicts with others, especially loved ones.

Emotion Triggers for Sad, Down, and Lonely

Emotion Triggers	Sad	Down	Lonely
Sickness	Cancer, pain, relapses, depression, mental breakdowns	Pain, depression	Pain
Self-hatred	Hate oneself	Hate oneself	Hate oneself
Personal relationships	Parents, being hated by loved ones, death of loved ones, go to funerals, had a breakup, say goodbyes, lack of friends and love, being rejected	Death of loved ones, go to funerals, lack of friends and love, miss someone	Had a breakup, marriage, divorce, homesick, lack of friends and love, valentines day, miss someone, being rejected, lack of belonging, lack of companionship
Cries	Cries	Cries	Cries
Self-disappointment	Disappoint in oneself, feel useless	Disappoint in oneself, feel useless	Disappoint in oneself, being unsure
Had an argument	Arguing with someone, being yelled at	Arguing with someone, being yelled at	
Being mistreated			Being ignored, did not get invited
Body image	Concerns about body weight	Concerns about body weight	
Failure	Failed, being unable to handle things	Failed	
Self-harm	Hurt oneself, thinking about suicide	Hurt oneself, thinking about suicide	
Sleep		Lack of sleep	

Table 4.6: Emotion Triggers for Emotion Sad, Down, and Lonely Concluded from the Multinomial Logistic Regression

Table 4.6 compares three inter-correlated emotions — *Sad*, *Down*, and *Lonely*. Based on the pairwise correlation of those three (Figure 4.6), it can be concluded that *Sad* is more correlated with *Down* than *Lonely*, and *Lonely* is more correlated to *Down* than *Sad*.

This conclusion is consistent with the emotions triggers that we find for each emotion. We see that *Sad* and *Down* shares the majority of the emotion triggers except that *Down* can be triggered by "lack of sleep".

Similar to *Sad* and *Down*, the emotion triggers of *Lonely* also involves "self-hatred", "self-disappointment", "cries" and "lack of friends and love". The difference between *Lonely* and the other two emotions is that it is not triggered by specific kinds of "sickness" or "death of loved ones" . Also, while *Sad* and *Down* are neither triggered by "exclusion" and "ignorance", *Lonely* can take place when those triggers occur.

Emotion Triggers for Stressed, Anxious, and Overwhelmed

Emotion Triggers	Stressed	Anxious	Overwhelmed
Sickness	Anxiety attacks, headaches, heart palpitations, migraines, pain, threw up, mental issue	Anxiety attacks, breathing issues, chest hurt, chest tightness, dizziness, heart palpitations, obsessive compulsive disorder, shaking, stomach hurts	Anxiety attacks, mental issue
School life	Have exams coming up, grades, homework due	Have exams coming up	Have exams coming up, have many assignments due, have many tests upcoming, have a paper due, have projects due, behind in school
Work	Being late for work	Going to give a presentation, job interviews	Too much work, Lack of preparation, messy home, moving home, packing up, procrastination
Had an argument	Arguing with someone, being yelled at		Arguing with someone
Failure	Going to fail, being unable to handle things		Going to fail, being unable to handle things
Self-hatred	Hate oneself		Hate oneself
Self-harm	Thinking about suicide		Thinking about suicide
Personal relationships		Love and romance	Had a breakup
Financial situation	Pay the bills		Pay the bills
Politics		Concerns about the presidential election	
Cries			Cries
Medications and Treatment		Taking medications, doctor appointments	
Self-disappointment		Over-thinking and self-doubt	
Sleep		Nightmares, too much caffeine intake, too much noise	

Table 4.7: Emotion Triggers for Emotion Stressed, Anxious, and Overwhelmed Concluded from the Multinomial Logistic Regression

Figure 4.6 shows that *Stressed* is highly correlated with both *Anxious* and *Overwhelmed*, but it is closer to *Overwhelmed* than *Anxious*. Table 4.7 also demonstrates this interconnection.

Stressed and *Overwhelmed* can both be triggered by stress in "school life", "work" and "bills to pay", physical "sickness", "mental health issues", "had an argument", "failure", "self-hatred", and "self-harm".

While "failure" is found to also trigger *Angry* and *Frustrated*, the failure for those two emotions are more about past failure. "failure" that triggers *Stressed* and *Overwhelmed* is more about worrying about potential failure in future.

Unlike *Anxious* and *Overwhelmed*, which can be triggered by issues regarding "personal relationships", it is found that those issues are less likely to trigger *Stressed*.

The emotion *Anxious* differs from *Stressed* and *Overwhelmed* in that it is mostly triggered by many specific illnesses, "medications" and "doctor appointments". Also, certain "sleep" issues, including nightmares, too much caffeine intake, noisy environment, is found to cause *Anxious*.

Emotion Triggers for Tired

Emotion Triggers	Tired
Work	Too much work
Sleep	Hangover, issue with falling asleep, nightmares, not slept well, feeling sleepy, stayed up late, wake up early
Sickness	Allergies, colds, coughing, cramps, eyes hurt, feet hurt, fevers, flu, headaches, infections, legs hurt, migraines, pain, soreness, stuffy nose, throat pain

Table 4.8: Emotion Triggers for Emotion Tired Concluded from the Multinomial Logistic Regression

Emotion triggers for *Tired* are provided in Table 4.8. There are two main triggers for *Tired*. The first is "sleep". We find that around 24% of predictive terms reflect "sleep" issues: "hangovers", "issues with falling asleep", "nightmares", "not slept well", "feeling sleepy", "stayed up late", and "wake up early". The second main emotion trigger is "sickness". We find that around 21% of predictive terms reflect specific illnesses: "allergies", "colds", "coughing", "cramps", "eyes hurt", "fevers", "flu", "headaches", "infections", "legs hurt", "migraines", "pain", "soreness", "stuffy nose", and "throat pain".

Emotion Triggers for Productive and Accomplished

Emotion Triggers	Productive	Accomplished
Body image	Painted nails and toes	Lost weight
Exercises	Running, workout	Climbing, cycling, hiking, hit 10k steps, running, walking, workout, yoga
Get things done	Get things done, going to finish, did laundry, went grocery shopping, productive meetings, cooked meals	Achieved goals, get things done, get things figured out, got paid off, handled things well, survived from challenges
Progressing	Making progress	Making progress
School life	Finished paperwork, completed assignments/essay/homework/paper/studying, get high scores, reading and writing	Gave a good presentation, had training, finished essays, passed tests, score well on exams
Self-recognition	Proud of oneself	Proud of oneself, reward oneself, self-confidence
Financial situation		Bonus from work
Food and meals		Cooked meals, eat healthily
Work	Had a buy day	

Table 4.9: Emotion Triggers for Emotion Productive, Accomplished Concluded from the Multinomial Logistic Regression

According to Figure 4.6, *Productive* and *Accomplished* are highly correlated, with a correlation of 0.7. Table 4.9 shows that *Productive* and *Accomplished* share majority of emotion triggers except that "had a busy day" can trigger *Productive*, while *Accomplished* can be triggered by bonuses and "eating healthy".

Overall, we conclude that *Productive* and *Accomplished* are both triggered by getting things done either in school or at work.

Accomplished are more likely triggered by not only completing but also doing a good job. Also, although "exercises" can trigger both those two emotions, specific workouts such as "climbing", "cycling", "hiking", "hitting 10k steps", and "yoga" can make people feel accomplished. Together with emotion triggers of "lost weight" and "eat healthily", we think that a successful fulfillment of weight management plans can contribute to the emotion *Accomplished*.

Emotion Triggers for Calm, Good, Happy, Excited, and Ecstatic

Emotion Triggers	Calm	Good	Happy	Excited	Ecstatic
Personal relationships	Family relationships, Hung out with friends	Hung out with friends, had a talk, had an online chat, had lunch with friends, had good laughs	Hung out with friends, had fun, had good laughs	Love and romance, hung out with friends, sleepovers, meet with someone	Love and romance, hung out with friends, have good laughs
Leisure activities	Listening to music, reading and writing, take bath, watch TVs and movies, chilling and relaxing	Play games, went grocery shopping, watch TVs and movies, chilling and relaxing	Play games, go to the beach/lake, watch TVs and movies, chilling and relaxing	Play games, go to events/concerts/party, birthday, Christmas, travels and trips, watch TVs and movies	Play games, go to events/the club/concert, travels and trips, watch TVs and movies, chilling and relaxing
Self-recognition		Proud of oneself, self-acceptance, self-confidence		Proud of oneself, self-confidence	Proud of oneself, self-confidence
Sleep	Rested well	Rested well	Rested well		
Weather	Raining, storms, warm	Sunny	Sunny		
Body image				Get hair done	Get hair done
Exercises				Dancing	Dancing, swimming
Food and meals	Drinking tea		Eat pizza, have yummy food		
Get things done					Got accepted into
Therapy	Getting massages, did some mindfulness or meditation				
Progressing	Making progress				
School life					High school started

Table 4.10: Emotion Triggers for Emotion Calm, Good, Happy, Excited, Ecstatic Concluded from the Multinomial Logistic Regression

Table 4.10 shows the emotion triggers identified for several positive emotions — *Calm*, *Good*, *Happy*, *Excited*, and *Ecstatic*. According to Figure 4.6, those five emotions are highly inter-correlated. Common emotion triggers for all those positive emotions are good personal relationships and variant leisure activities. "hanging out with friends" is a universal trigger for all positive emotions. Especially, good "romantic relationships" can trigger *Excited* and *Ecstatic*.

Several leisure activities are found to trigger *Calm*, *Good*, *Happy*, *Excited*, and *Ecstatic*: "listening to music", "reading and writing", "watching TVs and movies", "playing games", "grocery shopping", "going to events", "travelling", as well as a simple "chilling and relaxing".

"Self-recognition" can contribute to Emotion *Good*, *Excited*, and *Ecstatic*.

Different weather conditions are found to impact emotions. "Raining", "Stormy", and "Warm" weather can trigger *Calm*, while "Sunny" weather can trigger *Good* and *Happy*.

While sleep issues are found to trigger *Down*, *Anxious*, and *Tired*, "rested well" can contribute to *Calm*, *Good*, and *Happy*.

One interesting aspect of body image — "get hair done" — can trigger *Excited* and *Ecstatic*. Dancing can also cause these two emotions.

Particularly for *Calm*, "getting massages" and "doing meditations" is found to contribute to a sense of calmness.

4.3 The DMN model

4.3.1 Model Performance

Testing model performance on the same testing dataset, we find that the DMN achieves a 1% higher accuracy than the multinomial logistic regression model (Section 3.3.2).

4.3.2 Emotion Triggers Summarized by Interpreting the DMN Model

As described in Section 3.4.4, we extract sentences containing emotion triggers based on their attention gates and conclude themes of emotion triggers from those sentences. In Table 4.11, we provide an overview of emotion triggers that we identify from the DMN approach. A cell with a check mark means that the corresponding emotion trigger (row name) can trigger the emotion (column name). A cell without a check mark indicates that the emotion trigger is not observed to trigger the emotion (column name).

Table 4.11 shows that "Leisure activities", "Get things done", "Therapy", "Self-recognition", "Exercise", "Progressing" and "Weather" are emotion triggers that only trigger positive emotions.

Comparing with the top 100 predictive terms of the logistic regression model, we find that the top 100 unique trigrams we extracted here contain a much higher percentage of non-trigger-related content. We observe that around 20% more top terms are either direct expressions of emotions, swearing words, or top words. When concluding emotion triggers, we find that the results from the DMN approach contain fewer emotion triggers as it is summarized in Table 4.11. We conclude that although the sentence-level DMN model can filter out non-relevant events from journals, it still selects whole sentences. Thereby,

Emotion Triggers	Angry	Frustrated	Sad	Down	Lonely	Stressed	Anxious	Overwhelmed	Tired	Productive	Accomplished	Calm	Good	Happy	Excited	Ecstatic
Self-harm	✓		✓					✓								
Cries			✓	✓	✓			✓								
Personal relationships			✓	✓	✓		✓	✓					✓	✓	✓	✓
Sickness			✓	✓	✓	✓	✓	✓	✓							
Self-hatred	✓	✓	✓	✓		✓										
Body image			✓	✓									✓		✓	
Sleep							✓	✓	✓				✓	✓		
Work				✓	✓	✓		✓	✓	✓	✓				✓	
School life						✓	✓	✓		✓						
Food and meals		✓						✓		✓		✓	✓	✓		✓
Being mistreated	✓	✓			✓											
Failure	✓					✓										
Have an argument	✓															
Financial Situation				✓		✓										
Failure						✓										
Leisure activities												✓	✓	✓	✓	✓
Get things done									✓	✓	✓			✓		
Therapy												✓				
Self-recognition											✓					
Exercises											✓		✓		✓	
Progressing											✓					
Weather													✓	✓		

Table 4.11: Emotion Trigger Map for all emotions Concluded from the DMN

extracted sentences would contain more stop words than the top predictive terms extracted in logistic regression, since logistic regression assigns stop words low coefficients.

One important note for the emotion triggers generated from the top unique trigrams is that because common trigrams among different emotions are not included in the manual inspection stage, we may miss certain common emotion triggers among different emotions.

Grouping emotion triggers based on the categories in Table 4.1, we examine their emotion triggers below. Detailed scenarios for emotion triggers are shown in corresponding cells in the below tables.

Emotion Triggers for Angry and Frustrated

Emotion Triggers	Angry	Frustrated
Being mistreated	Being lied to, being treated rudely/unfairly	Being treated rudely
Self-hatred	Hate oneself	Hate oneself
Self-harm	Thinking about suicide	
Failure	Something doesn't work	
Have an argument	Fight with someone	
Food and meals		To stop eating

Table 4.12: Emotion Triggers for Emotion Angry and Frustrated Concluded from the DMN

Table 4.12 shows that *Frustrated* will be triggered by certain eating disorders ("to stop eating") while *Angry* will not. This trigger is also a new emotion trigger that does not appear in the previous result of our logistic regression approach.

Similar to the results of the logistic regression approach, we find that *Angry* and *Frustrated* share common triggers of "being mistreated" and "self-hatred". Additionally, "self-harm" can lead to *Angry* but not *Frustrated*.

Emotion Triggers for Sad, Down, and Lonely

Emotion Triggers	Sad	Down	Lonely
Cries	Cries	Cries	Cries
Sickness	Pain, depression	Depression	Pain
Self-hatred	Hate oneself	Lack of hopes	
Personal relationships	Parents, being hated by loved ones, had a breakup, love and romance		Marriage, siblings
Work		Lack of energy	Done nothing
Body image	Concerns about body weight	Concerns about body weight	
Being mistreated			Being ignored
Self-harm	Thinking about suicide		
Financial Situation		Can't afford to	

Table 4.13: Emotion Triggers for Emotion Sad, Down, and Lonely Concluded from the DMN

There are two findings in table 4.13 that are consistent with the findings from the logistic regression approach: "cries" and "sicknesses" are common triggers among *Sad*, *Down*, and *Lonely*; "concerns about body weight" are found to trigger *Sad* and *Down*.

One new finding is that financial pressure — "can't afford to" — can trigger *Down* but not *Sad* and *Lonely*.

Emotion Triggers for Stressed, Anxious, and Overwhelmed

Emotion Triggers	Stressed	Anxious	Overwhelmed
Sickness	Threw up	Anxiety attacks, heart palpitations	Anxiety attacks, period cramps, mental breakdowns, mentally exhausted
School life	School grades, have exams coming up	Have exams coming up	Have exams coming up
Personal relationships		Love and romance	Love and romance, parents, siblings
Sleep		Take too much caffeine	Need a break
Work	Being late for work		Too busy, moving in or out, packing up
Self-harm			Thinking about suicide
Failure	Being unable to handle things		
Self-hatred	Hate oneself		
Cries			Cries
Financial Situation	Have no money, worried money		
Failure	Being unable to handle things		
Food and meals			Didn't eat

Table 4.14: Emotion Triggers for Emotion Stressed, Anxious, and Overwhelmed Concluded from the DMN

Again, emotion triggers obtained from the DMN model agree partly with that from the logistic regression approach. It can be seen from Table 4.14 that "sickness" and approaching "exams" can trigger *Stressed*, *Anxious*, and *Overwhelmed*. Also, issues in "personal relationships" can trigger *Anxious*, and *Overwhelmed* but not *Stressed*.

One new finding is that "didn't eat" can trigger *Overwhelmed* but not *Stressed* and *Anxious*.

Emotion Triggers for Tired

Emotion Triggers	Tired
Sickness	Eyes hurt
Get things done	Get anything done
Work	Lack of energy, long work shift, too much work
Sleep	Not slept well, issue with falling asleep, stayed up late, wake up early, sleepy, take nap

Table 4.15: Emotion Triggers for Emotion Tired Concluded from the DMN

Emotion triggers for *Tired* are summarized in Table 4.15. Additional to the findings in Section 4.2.3, it is found that "long work shift" can cause tiredness.

Emotion Triggers for Productive and Accomplished

Emotion Triggers	Productive	Accomplished
Work	Working from home, working on tasks	Get an interview
Get things done	To-do lists done, going to finish, finished errands, help someone	Complete independently, good at doing things
School life	Studying done	
Food and meals	Eat breakfast, cooked meals	
Self-recognition		Proud of oneself, self-confidence
Exercises		Running, workout, hit 10k steps
Progressing		Learning and growth

Table 4.16: Emotion Triggers for Emotion Productive, Accomplished Concluded from the DMN

While we did not find eating-related triggers for *Productive* in the previous Section 4.2.3, we show in Table 4.16 that "eat breakfast" and "cooked meals" can contribute to the emotion *Productive*.

Emotion Triggers for Calm, Good, Happy, Excited, and Ecstatic

Emotion Triggers	Calm	Good	Happy	Excited	Ecstatic
Leisure activities	Watch TVs and movies	Watch TVs and movies	Play games, watch TVs and movies, take someone out, go to the beach	New start, going to the party, travels and trips, go to the beach	Event tickets, football game
Personal relationships		Love and romance, had a talk, hung out with friends	Parents, siblings, had a talk, friendship, hung out with friends, home with family	Parents, siblings	Love and romance
Food and meals	pizza for dinner	had good breakfast	Out to lunch		Ate
Sleep		Rested well	Wake up early		
Exercises		Workout, cycling		Workout	
Weather		Sunny	Sunny		
Body image		Get hair done		Painted nails and toes	
Work				Start working	
Get things done			Off work		
Therapy	Went to therapy				

Table 4.17: Emotion Triggers for Emotion Calm, Good, Happy, Excited, Ecstatic Concluded from the DMN

Table 4.17 presents the emotions triggers for *Calm*, *Good*, *Happy*, *Excited*, and *Ecstatic* found from the DMN approach. The results are mostly consistent with what we see from the logistic regression approach.

Additionally, we discover several emotion triggers that can contribute to different types of positive emotions. "Rested well", "workout", "therapy", "pizza", "eating well", and "doing leisure activities" can trigger positive emotions among *Calm*, *Good*, *Happy*, *Excited*, and *Ecstatic*.

4.4 Comparing Multinomial Logistic Regression with the DMN

4.4.1 Ability to Visualize Individual Examples

To compare the bag-of-words multinomial logistic regression and the DMN model, we take the same synthetic example as in Figure 3.4 (b) and use the multinomial logistic regression to make a prediction. We find that the multinomial logistic regression also makes the correct prediction in this synthetic example.

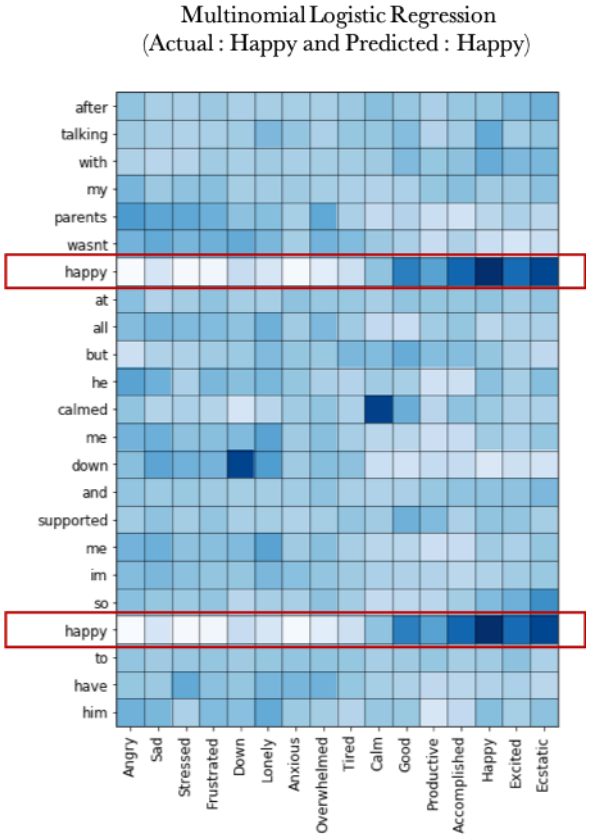

Figure 4.7: Visualization of a correctly classified synthetic example that uses Multinomial Logistic Regression. The X-axis represents emotions. The Y-axis represents a sequence of input words from top to bottom. The value of term coefficients determines the color of cells. Darker cells represent higher term coefficients (higher word importance).

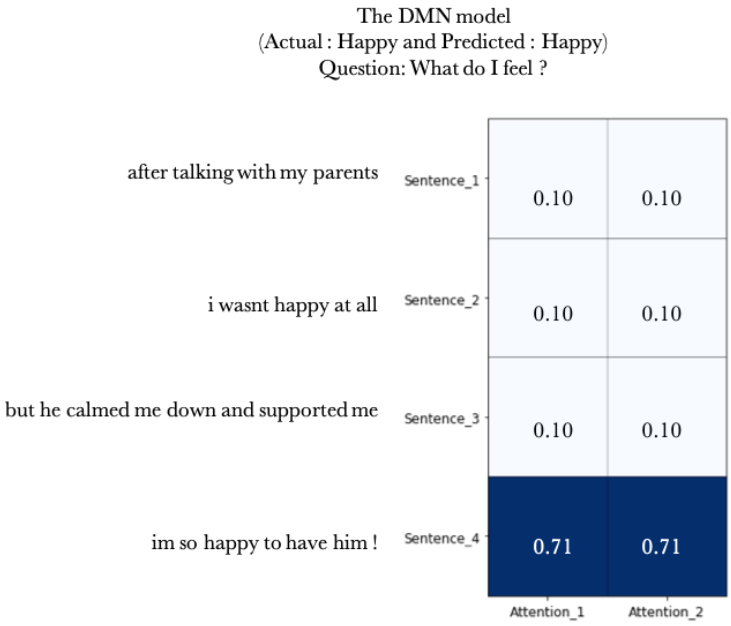

Figure 4.8: Visualization of a correctly classified synthetic example that uses the DMN model. The X-axis represents memory episodes. The Y-axis represents a sequence of input sentences from top to bottom. The value of attention gates determines the color of cells. Darker cells represent larger attention gates (higher sentence importance).

We further visualize the "reasoning" process of the multinomial logistic regression and the DMN model in Figure 4.7 and Figure 4.8 respectively. We conclude that there are two main differences between these two models in terms of model interpretability:

- While the DMN model can be trained and visualized both on the word level and on the sentence level, multinomial logistic regression can only be trained and visualized on the word level. Thereby, we think the DMN can provide an extra option to visualize the importance of sentences.
- The DMN shows a higher ability to adjust word importance based on its context than multinomial logistic regression. In Figure 4.7, the coefficient of the first "happy" is the same as the second "happy", despite that the first one is used in a negated context. On the other side, for the DMN model, the sentence that contains the first "happy" is only given a 10% of attention during the prediction, while the second "happy" is given a 71% of the attention. After investigating more visualization of individual examples, we find that for individual examples, the attention mechanism of DMN model can capture more language context.

4.4.2 Ability to Summarize Emotion Triggers from All Text

Comparing emotion triggers reported in Section 4.2.3 and 4.3.2, we find that the interpretation approach of multinomial logistic regression generates more emotion triggers than the DMN. Thereby, we think multinomial logistic regression is more helpful in obtaining summaries of emotion triggers from the whole data.

The comparable fewer emotion triggers generated from the DMN directly result from the attention issue. Even though our proposed new regularization method has greatly improved attention sparsity as it is shown in Section 3.4.3, there are still about 65% of journals that do not have sufficiently sparse attention gates, so we do not include those journals in the manual inspection. We consider this may potentially remove useful information that indicates emotion triggers.

Chapter 5

Discussions and Conclusions

Motivated by the importance of emotion triggers, we set our research goal to obtain data-driven insights about emotion triggers. We explored a unique journal dataset that contains text and ground-truth emotion labels. To understand emotion triggers for the 16 different emotions in our data, we first built two emotion classification models — n-gram multinomial logistic regression and the DMN — and further performed model interpretation.

For n-gram multinomial logistic regression, we leveraged term coefficients to select predictive terms. For the DMN model, we utilized attention gates to extract sentences with high attention.

5.1 Discussions

5.1.1 Model Performance

Testing model performance on the same testing dataset, we find that our multinomial logistic regression using Tf-idf of unigrams achieves an accuracy of 33.5%, which is only 1% lower than that of the DMN model (Table 3.4). The strong performance of logistic regression is demonstrated in many prior studies (Bagroy et al., 2017; Thelwall et al., 2010a; Murdoch et al., 2018; Socher et al., 2011; Zhang et al., 2015).

Among unigrams, bigrams, and trigrams, our randomized hyperparameter experiments on multinomial logistic regression show that unigrams are more predictive than the others (Section 3.3.2). This finding is consistent with most prior work (Pang et al., 2002; Go et al., 2009).

For mining emotion triggers, we show that both the unigram multinomial logistic regression model and the DMN model generate findings of emotion triggers. However, the attention mechanism in DMN could not be fully utilized to select emotion triggers accurately, as the attention is trained in an unsupervised way. Comparing emotion triggers reported in Section 4.2.3 and 4.3.2, we find that the interpretation approach of multinomial logistic regression generates more emotion triggers than the DMN. As a result, we think that multinomial logistic regression is more effective in obtaining summaries of emotion triggers from data.

The comparable fewer emotion triggers obtained from the DMN directly results from the attention issue. Even though our proposed new regularization method has greatly improved attention sparsity as it is shown in Section 3.4.3, there are still about 65% of journals that do not have sufficiently sparse attention gates, and we do not include those journals in our manual inspection process. We consider this may potentially cause information loss.

Even though multinomial logistic regression performs well in terms of classification accuracy and summarizing emotion triggers, the DMN still has two major advantages. First, the DMN still achieves a 1% higher prediction accuracy than the multinomial logistic regression model. Second, DMN demonstrates a higher ability to differentiate word contexts. As shown in Figure 4.6 and 4.7, while multinomial logistic regression assigns the same coefficients to the two "happy" words that are expressed in opposite emotions, the DMN assigns higher attention gates to the sentences that use the "happy" in a positive context.

5.1.2 Regularization Issue with Unsupervised Attention

We have shown the issue with unsupervised attention modules in Section 3.4.1, which results in uniform attention weights across all input sentences. We address the issue and discover that the common L2 regularization practice on model parameters actually "encourages" the issue.

We further propose three new regularization methods to encourage attention sparsity when attention is trained in an unsupervised way. Our experiments show that by maximizing the mean variance of attention gates in all memory episodes and setting λ to 1, the DMN model produces the same accuracy but significantly more variant attention gates.

5.1.3 Emotion Triggers and Their implications

In Chapter 4, we demonstrate all emotion triggers that are identified from our analysis. We further discuss those emotion triggers and provide their practical implications as follows:

High percentage of Tired calls for attention to sleep issues, illnesses, and work life balance

We find that *Tired* is the third most common emotion in our data, which is more frequent than *Happy*. We also consider *Tired* negative as it is correlated with several negative emotions — *Stressed*, *Frustrated*, *Down*, and *Anxious*, *Overwhelmed*.

Based on our identified emotion triggers for *Tired*, we find that there are two main triggers — sleep problems and sickness. Specifically, the sleep problems include having nightmares, difficulty in falling asleep, poor sleep quality, and staying up late. The sickness includes many types of illnesses and associated physical pains, such as allergies, headaches, migraines and so on. Long work shift and high workload may also lead to tiredness.

In terms of its days of the week distribution, we observe that *Tired* accounts for a lower percentage on Fridays and Weekends than other days. Our finding is contrary to that of Dzogang et al. (2017) who found that the circadian pattern of fatigue is resistant to the weekend and weekday changes.

Given the high frequency of *Tired*, we think that effective treatments for *Tired* will not only improve an individual's emotion well-being but also contribute to a scale of public mental health. We suggest emotion therapists help patients evaluate their sleep quality, workload, and physical health conditions and address potential causes.

Sad and Lonely reach their highest point on Sundays

We find that the percentages of *Sad* and *Lonely* journals are the highest on Sundays. As suggested by Dzogang et al. (2017), the circadian patterns of sadness are subject to weekday and weekend changes. As people's circadian rhythms are commonly associated with sleep and activities, we suspect that emotion triggers of Sad and Lonely may contain activities that are dependent on the change of time schedules between weekdays and weekends.

Among emotion triggers that we discover for *Sad*, we find that several types of emotion triggers may differ between weekdays and weekends: personal relationships and having arguments with someone. For *Lonely*, which is highly correlated with *Sad*, triggers that may differ between weekdays and weekends are personal relationships, being rejected, having nothing to do, and not getting invited. It is shown that all of the above triggers somehow involves the needs of friends and companionship.

As we also find that "hanging out with friends" is a universal trigger for all positive emotions and *Happy* is the highest on weekends, we think that keeping good relationships

with friends and inviting friends for leisure activities will help to reduce *Sad* and *Lonely* on Sundays.

Negative emotions are triggered by either external environments or internal environments

We find that there are several factors of external environments that have brought all sorts of negative emotions to people: intimate relationship problems (including infidelity), work pressures, school stress, financial burden, politics, and being mistreated by others.

Meanwhile, triggers that only involve ones' health and inner status also impact ones' emotion status. We find that those "internal" triggers also generate all types of negative emotions among people: being sick, taking certain medications, being unable to sleep well, overly concerning about body image and weight, focusing too much on failure, self-hatred, self-harm, self-disappointment, and cries.

The existence of emotion triggers from both external environments or internal environments implies that effective emotion treatments must help people identify, rethink, and develop coping mechanisms for both types of triggers. Particularly, we find that in terms of the inner self, focusing more on self-recognition instead of overly criticizing oneself may bring happiness.

Setting aside time for self-care will improve one's emotion status

Based on emotion triggers for all kinds of positive emotions, we find that self-care activities are essential to a positive emotion status:

- eating healthy and comfort food
- doing all kinds of exercises
- getting a massage
- doing meditations
- resting well
- getting hair done

We recommend that these self-care activities should be emphasized in peoples lives. To help people recognize the importance of self-care, we suggest that governments should carry out campaigns to promote those self-care activities and encourage people to integrate those activities in their daily routine.

Paying attention to weather conditions may make people feel more positive

We find that all sorts of weather conditions are highly predictive for positive emotions (not only the "good" weather). For example, "Raining" and "Stormy" can bring *Calm*. 'Sunny" weather can trigger *Good* and *Happy*.

5.2 Limitations and Future work

We consider that our study has three limitations. First, although our data contain journals from a sufficiently large number of writers (67,000), we think that our data have the same limitation as those sentiment studies that use data from social media sites: findings are only applicable to a specific group of people that use the social media sites. However, this issue is difficult to resolve in future work. Second, although both of our emotion classification models achieve higher accuracy than baselines, the accuracy scores are still relatively low and are far from being 100% accurate. As a result, we consider that the emotion triggers that we identified may subject to certain errors. Third, as we train the DMN model using only supervised learning on emotion labels but not attention, we observe that there are about 65% of journals that are not inspected because of their uniform attention gates. This may limit our findings of emotion triggers.

To address the above limitations, we plan the following future work. On the one hand, more research will be focused on improving emotion classification accuracy. On the other hand, we intend to train the DMN model using supervised learning on both emotion labels and attention. That is, we want to test if supervised attention will produce sparser attention gates and help us identify more emotion triggers. To achieve this, we plan to sample a set of journals and manually create labels of sentences that contain emotion triggers. Once we have the labelled dataset, we also plan to explore other Question-Answering models that will directly output emotion triggers given text input and question input.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.

Shrey Bagroy, Ponnurangam Kumaraguru, and Munmun De Choudhury. A social media based index of mental well-being in college campuses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1634–1646. ACM, 2017.

Alexandra Balahur, Jesús M Hermida, Andrés Montoyo, and Rafael Muñoz. Emotinet: A knowledge base for emotion detection in text built on the appraisal theories. In *International Conference on Application of Natural Language to Information Systems*, pages 27–39. Springer, 2011.

Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

Lyndal Bond, John B Carlin, Lyndal Thomas, Kerryn Rubin, and George Patton. Does bullying cause emotional problems? a prospective study of young teenagers. *Bmj*, 323 (7311):480–484, 2001.

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. Quasi-recurrent neural networks. *arXiv preprint arXiv:1611.01576*, 2016.

Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira Jr, and Virgílio Almeida. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2011.

Laura Campbell-Sills, David H Barlow, Timothy A Brown, and Stefan G Hofmann. Acceptability and suppression of negative emotion in anxiety and mood disorders. *Emotion*, 6(4):587, 2006.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, 2015.

Fabon Dzogang, Stafford Lightman, and Nello Cristianini. Circadian mood variations in twitter content. *Brain and neuroscience advances*, 1:2398212817744501, 2017.

Paul Ekman. *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life.* Henry Holt and Co., New York, NY, US: Times Books, 2003.

Barbara L Fredrickson and Thomas Joiner. Positive emotions trigger upward spirals toward emotional well-being. *Psychological science*, 13(2):172–175, 2002.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning, 2001.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer, 2015.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

James J Gross and Ricardo F Muñoz. Emotion regulation and mental health. *Clinical psychology: Science and practice*, 2(2):151–164, 1995.

William James. What is an emotion? mind. *Reprinted in SW Porges, y MGH Coles (eds.): Psychophysiology. Stroudsburg, Pal., Dowden, Hutchinson y Ross*, 9:188–205, 1884.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387, 2016.

Richard S Lazarus. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819, 1991.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Weiyuan Li and Hua Xu. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749, 2014.

Robert M Liebert and Larry W Morris. Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological reports*, 20(3):975–978, 1967.

Qian Lin and Hongyu Xiong. Dynamic memory network on natural language question-answering, 2016.

David J McIver, Jared B Hawkins, Rumi Chunara, Arnaub K Chatterjee, Aman Bhandari, Timothy P Fitzgerald, Sachin H Jain, and John S Brownstein. Characterizing sleep issues using twitter. *Journal of medical Internet research*, 17(6), 2015.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Gilad Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, volume 19, pages 321–327, 2005.

Saif M Mohammad. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics, 2012.

W. James Murdoch, Peter J. Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkRwGg-0Z.

Mark Mysln, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15:e174, 08 2013. doi: 10.2196/jmir.2534.

Alena Neviarouskaya and Masaki Aono. Extracting causes of emotions from text. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 932–936, 2013.

Joel Nothman, Hanmin Qin, and Roman Yurchak. Stop word lists in free open-source software packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 7–12, 2018.

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

Albert Park, Mike Conway, and Annie T Chen. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. *Computers in human behavior*, 78:98–112, 2018.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions.

In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–161. Association for Computational Linguistics, 2011.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010a.

Mike Thelwall, David Wilkinson, and Sukhvinder Uppal. Data mining emotion in social network communication: Gender differences in myspace. *Journal of the American Society for Information Science and Technology*, 61(1):190–199, 2010b.

Andrew Toulis and Lukasz Golab. Social media mining to understand public mental health. In *VLDB Workshop on Data Management and Analytics for Medicine and Healthcare*, pages 55–70. Springer, 2017.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 587–592. IEEE, 2012.

Maggie Watson and Steven Greer. Development of a questionnaire measure of emotional control. *Journal of psychosomatic research*, 27(4):299–305, 1983.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406, 2016.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

APPENDICES

Appendix A

Negation Conversions

Before	After	Before	After	Before	After	Before	After
'Am not'	'amnt'	'am not'	'amnt'	'Haven t'	'havent'	'haven t'	'havent'
'Amn t'	'amnt'	'amn t'	'amnt'	'Im not'	'I amnt'	'im not'	'I amnt'
'Are not'	'arent'	'are not'	'arent'	'Is not'	'isnt'	'is not'	'isnt'
'Aren t'	'arent'	'aren t'	'arent'	'Isn t'	'isnt'	'isn t'	'isnt'
'Can not'	'cant'	'can not'	'cant'	'Might not'	'mightnt'	'might not'	'mightnt'
'Can t'	'cant'	'can t'	'cant'	'Mightn t'	'mightnt'	'mightn t'	'mightnt'
'Cannot'	'cant'	'cannot'	'cant'	'Must not'	'mustnt'	'must not'	'mustnt'
'Could not'	'couldnt'	'could not'	'couldnt'	'Mustn t'	'mustnt'	'mustn t'	'mustnt'
'Couldn t'	'couldnt'	'couldn t'	'couldnt'	'Shall not'	'shant'	'shall not'	'shant'
'Did not'	'didnt'	'did not'	'didnt'	'Shan t'	'shant'	'shan t'	'shant'
'Didn t'	'didnt'	'didn t'	'didnt'	'Should not'	'shouldnt'	'should not'	'shouldnt'
'Do not'	'dont'	'do not'	'dont'	'Shouldn t'	'shouldnt'	'shouldn t'	'shouldnt'
'Does not'	'doesnt'	'does not'	'doesnt'	'Was not'	'wasnt'	'was not'	'wasnt'
'Doesn t'	'doesnt'	'doesn t'	'doesnt'	'Wasn t'	'wasnt'	'wasn t'	'wasnt'
'Don t'	'dont'	'don t'	'dont'	'Were not'	'werent'	'were not'	'werent'
'Had not'	'hadnt'	'had not'	'hadnt'	'Weren t'	'werent'	'weren t'	'werent'
'Hadn t'	'hadnt'	'hadn t'	'hadnt'	'Will not'	'wont'	'will not'	'wont'
'Has not'	'hasnt'	'has not'	'hasnt'	'Won t'	'wont'	'won t'	'wont'
'Hasn t'	'hasnt'	'hasn t'	'hasnt'	'Would not'	'wouldnt'	'would not'	'wouldnt'
'Have not'	'havent'	'have not'	'havent'	'Wouldn t'	'wouldnt'	'wouldn t'	'wouldnt'

Table A: Negation Conversions used in the text pre-processing. Column "Before" and "After" contain terms before and after we converting negated expressions in text.

Appendix B

Distribution of expressed emotions by days of the week

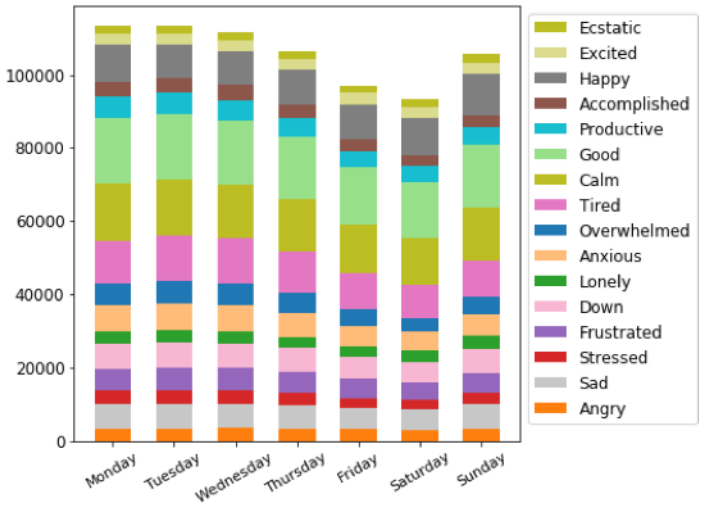

Figure B1: The number of journals by emotions and days of the week.

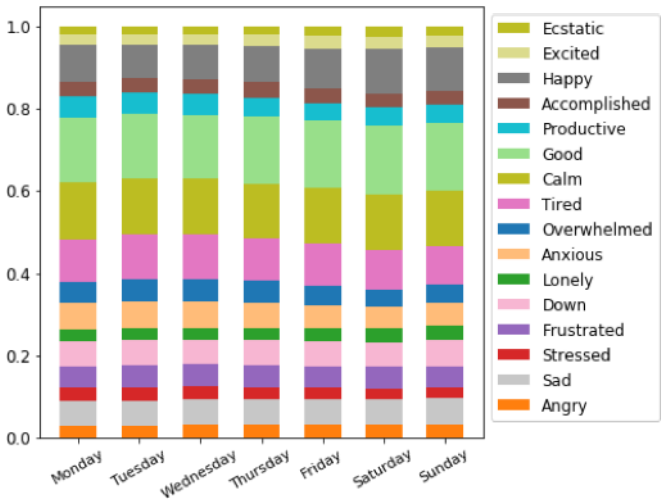

Figure B2: Distribution of emotions by days of the week.