Development of Text Analytics for Debriefing Reflection Essays

by

Md Shadekur Rahman

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science at
The University of Texas at Arlington
May, 2020

Arlington, Texas

Supervising Committee:

Deokgun Park, Supervising Professor

Shirin Nilizadeh

Leonidas Fegaras

Yan Xiao

DEDICATION

I dedicate this thesis to my parents who developed in me a thirst for knowledge and education since my childhood. My father‘s positive encouragement to think big and to always remain connected with educated community made me opt for higher study. I also dedicate this thesis to my loving wife for her untiring support during this whole MS program.

ACKNOWLEDGEMENTS

I would like to thank my professor Dr. Deokgun Park for giving me the opportunity to conduct research in the Human Data Interaction (HDI) Lab of CSE department within the College of Engineering of University of Texas at Arlington. I am also indebted to Dr. Shirin Nilizadeh and Dr. Yan Xiao for their regular collaboration in this particular research project. I am appreciative of all the supports I received from the team members of HDI lab. This work was supported by a Faculty Research Grant Program from Academic Partnerships (AP).

Abstract

Development of Text Analytics for Debriefing Reflection Essays

Md Shadekur RAHMAN, Master of Science

The University of Texas at Arlington, 2020

Department of Computer Science

Supervising Professor: Deokgun Park

Evaluating and providing feedback to hundreds of free text assignments in an online environment is a challenging task for an instructor where he has to scan through essays to identify perspectives that are expected to appear in those essays. Reading large number of essays and then finding themes and providing customized feedback are time-consuming process. We have proposed a text analytics system named *EssayIQ* that aids course instructor in identifying assignment themes, providing theme presence statistics and giving feedback to learners. To the best of our knowledge, this is the first system that analyzes free text assignments in line with the instructor defined themes. Through our experiments on one online course, we have shown that model based on sentence-level semantic embedding outperforms word and phrase based embedding models. We have also shown that *EssayIQ* system can identify themes and can generate overall theme statistics for over hundred submissions within minutes with minimal theme knowledge intake by *EssayIQ*. The theme identification quality of *EssayIQ* system is also comparable with human/coach annotators. The code of this project is publicly available in github (https://github.com/Shadek07/EssayIQ).

keywords: automated concept identification, visual analytics, text analytics, phrase2vec, online learning, word2vec, universal sentence encoder, reflection essay, debriefing essay

Contents

List of Figures

List of Tables

List of Abbreviations

NLP	**N**atural **L**anguage **P**rocessing
USE	**U**niversal **S**entence **E**ncoder
NPMI	**N**ormalized **P**ointwise **M**utual **I**nformation
Phrase2vec	**W**ord2vec for **P**hrases

Chapter 1
INTRODUCTION

Reflective writing (*Reflective writing*) is one of the most important practices in online learning where a writer writes about personal reflection describing a real or imaginary scene or event. Learners can acquire adequate insights and observation and connect with personal experience by writing essays or less structured reflections and journals. These writing assignments are usually guided by reflection and debriefing questions related to new concepts presented in a video or simulation of a real-life experience. Providing evaluation and feedback of debriefing reflection essays is challenging in online environment as it requires a well-trained, well-informed instructor with evidence-based practices to carefully read these essay submissions. Human coaches and instructors have limited time to effectively assess and provide feedback to hundreds of essay submissions in each week. The recent advances in natural language processing (NLP) and deep learning have made it possible to extract important information from natural language texts. These advances aid human instructors by requiring less time in grading and providing structured feedback that aligns with learning objectives.

The goal of this thesis is to develop and pilot test a text analytics system named, *EssayIQ* that can identify themes or concepts which are expected to appear in the submitted debriefing reflection essays and suggest automated feedback to learners . The “research question" that our system will attempt to answer is, how can we combine the power of human and machine to scale up the analyzing process of open-ended reflection essays by automatically spotting themes and by automatically suggesting feedback. Each theme consists of generic answers to one or several debriefing questions presented in the assignment description. Typically, a writing assignment will have 5-6 themes. Another purpose of EssayIQ is to judge relevance to learning objectives from the submitted student‘s essays. Our system is based on two independent models, Phrase2Vec and Universal Sentence Encoder (USE). USE provides models to encode sentence into embedding vectors. The fundamental reason of choosing *USE* over other neural network models such as Word2Vec, Phrase2Vec or Conceptvector is that it is best suited for the task of semantic similarity and our experimental results support this choice.

In chapter two, we will present *EssayIQ* system in an elaborate manner along with literature review. In chapter three, we will present experimental results for two models employed by *EssayIQ*.

Chapter 2
LITERATURE REVIEW

Natural language processing and machine learning have been utilised in the field of automated processing and manipulation of text assignments. Automated essay scoring (AES) is one of the aspects of text processing outcome. The fundamental and modern approaches to AES have been taken in the studies of (Valenti, Neri, and Cucchiarelli, 2003), (Attali and Burstein, 2006), (Foltz, Laham, and Landauer, 1999), and (Shermis and Burstein, 2003). These approaches focus on grading essays as a whole. However, identifying instructor defined themes or concepts and providing meaningful feedback to learners serve a different purposes and bring challenges to completely automated text processing. One particular challenge is - relating recognized themes from open-ended essays to learning objectives. There are several research works that align with this work with regard to providing feedback to learners. (Taghipour and Ng, 2016), (Mittal and Devi, 2016), (Hastings, Hughes, and Britt, 2018), (Li and Sugumaran, 2018), (McNamara et al., 2014) are some of them. However, this work has distinctive nature from others in that it has to deal with the reflective debriefing essays in a graduate level course where instructors and students have to practice evidence based fields such as health care and nursing. The work done by (Altoe and Joyner, 2019) considers example essay summary based rubric generation. However, our work differs from them as we do not aim to evaluate whole essay text rather we want to identify instructor defined themes or concepts. Some essays may have general overview of all themes while some other essays will have deep depiction of one or two themes only.

A visual analytics tool, *CommentIQ* (Park et al., 2016) was developed to aid human moderators in selecting and refining high quality comments by leveraging domain knowledge and automated methods. Topic Modelling can identify latent topics in corpus, where a topic is represented as the probability of terms appearing in the text belonging to topic. Topic modelling requires large corpus and it is hard to invest domain knowledge for each topic by human. These two systems (CommmentIQ, Topic modelling) do not address research question of this project where we need to find semantic similarity between two pair of texts or sentences.

Another tool named *ConceptVector* (Park et al., 2018) has been developed to address the barrier of domain knowledge instillation. It uses word-embedding that maps each term to a vector. Conceptvector is powerful to build custom lexicons, each lexicon consists of terms closer to each other in embedding space. However, the research problem that we are trying to address here has concepts whose terms or lexicons might be completely unrelated to each other. For example, a concept in a reflective essay could be all possible answers of an open-ended debriefing question. Building a lexicon with similar or opposite terms might not serve our purpose. For the same reason Word2Vec (Mikolov et al., 2013) or even Word2Vec for Phrases which we call Phrase2Vec (*Phrase2vec*) model did not prove to be beneficial to address our research problem.

Universal sentence encoder (USE) (Cer et al., 2018) developed by some Google researchers is applicable for various NLP tasks including semantic similarity finding. We have attempted to employ this encoder in our *EssayIQ* system and experimental results using USE is much accurate compared to Phrase2vec.

Chapter 3
ESSAYIQ: A TEXT ANALYTICS SYSTEM

3.1 *EssayIQ* system architecture

In EssayIQ system (Fig. 3.1), there are three main components. First one is, essay submissions; second one is *EssayIQ* model and third one is instructor. EssayIQ takes student's essays and analyzes them with the aim of detecting themes and generating custom feedback. Instructor, by leveraging EssayIQ provides effective feedback to learners. In fig 3.2, an inside look of EssayIQ model is depicted. EssayIQ initially have a knowledge base of theme sentences for every theme in an assignment. These theme sentences are provided to EssayIQ by instructor or coaches. EssayIQ system runs iteratively where in each iteration instructor sees highlighted text content analyzed by EssayIQ and updates knowledge base of theme sentences.

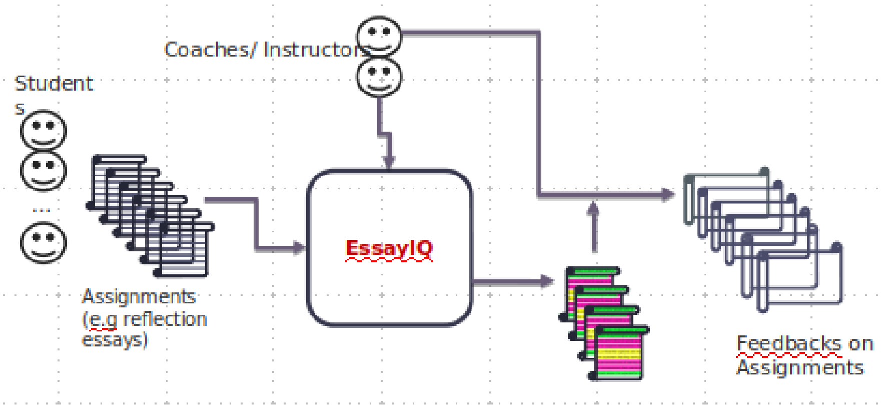


FIGURE 3.1: EssayIQ architecture overview

3.2 Development approach of EssayIQ system

In this project, we will use reflection essays as our development target, however EssayIQ system is applicable other types of essays including simulated debriefing.

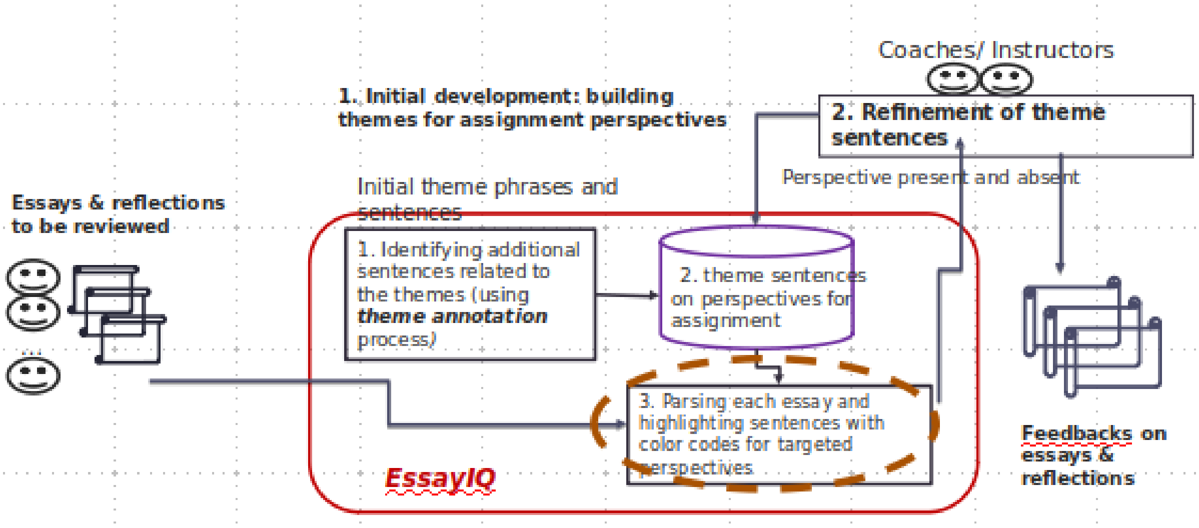
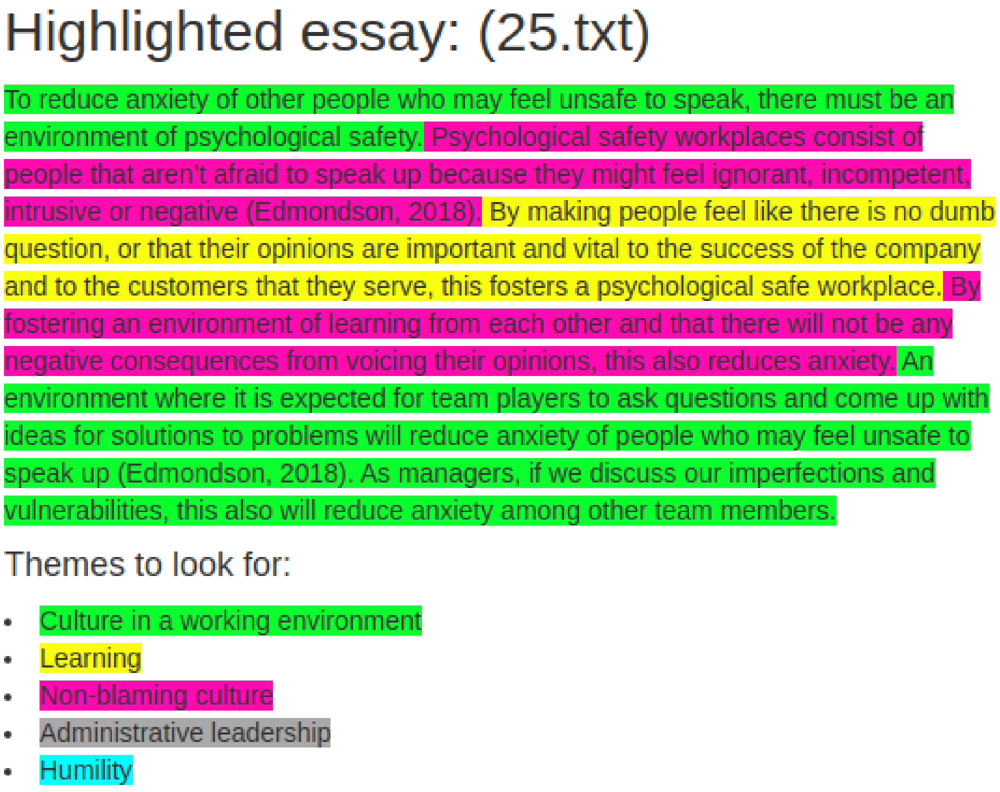

Coaches/ Instructors
1. Initial development: building themes for assignment perspectives
2. Refinement of theme sentences
Perspective present and absent
Initial theme phrases and sentences
Essays & reflections to be reviewed
1. Identifying additional sentences related to the themes (using *theme annotation* process)
2. theme sentences on perspectives for assignment
3. Parsing each essay and highlighting sentences with color codes for targeted perspectives
EssayIQ
Feedbacks on essays & reflections

FIGURE 3.2: EssayIQ model details

In a typical reflection assignment, a student will submit reflection after watching an assignment video where an error is presented in patient treatment. Based on learner‘s experience and background, there would be multiple layers and perspectives in the video. As per illustration, let’s assume there are five important aspects or themes that the instructor wants to explain. Learner‘s submissions will be diverse as some submissions will focus on one or two aspects while others will present a generic overview. The instructors or coaches may want to provide customized feedback for each submission. Instructor can categorize common sentence pattern into groups of themes and build a candidate template for each group during giving feedback. The goal of *EssayIQ* is to scaling up the submission feedback process while meaningfully engaging large number of students.

We will employ both Phrase2vec and Universal sentence encoder (USE) to identify semantically similar sentences with respect to theme sentences. A theme consists of a collection of sentences that will be used as standard to represent that theme. A sentence from a learner‘s submission will be compared against all sentences from each theme. A sentence (from a theme a.k.a theme sentence) that is semantically closer to submitted sentence will be used as marker or identifier for that submitted text. First, two coaches will annotate all submission essays of an assignment. Annotation corresponds to leveling each essay sentence with a theme if sentence belongs to that theme. A certain number of annotated sentences (usually 5-20 sentences) from each theme will be used for EssayIQ system representation. By using this percentage of data, system will identify themes from essay sentences that have not been marked by coaches.

During the assignment analysis phase, the system will show a submitted essay where each sentence is highlighted with a color (i.e theme color) corresponding to a theme. An essay highlighted by EssayIQ is shown in figure 3.3. With this highlighting, instructors can see an overview of the text and build a hypothesis of which reflections exist in the essay. Then they can skip over the essay to verify if the hypothesis is correct. If there are words or sentences that are not helping in theme identification process, coaches can update their annotations. After essay evaluation, the instructors label the essay as containing one or more theme groups and assign appropriate feedback template. Over time, the instructors can see general patterns of the submissions and can examine which perspective is prevalent in the essay. Thus, the instructors can redesign course content including changing or adding reflection and debriefing questions. A summary of assignment highlighting is drawn in a plot (Fig 3.4) showing the counts of submissions for each theme.



Highlighted essay: (25.txt)

To reduce anxiety of other people who may feel unsafe to speak, there must be an environment of psychological safety. Psychological safety workplaces consist of people that aren't afraid to speak up because they might feel ignorant, incompetent, intrusive or negative (Edmondson, 2018). By making people feel like there is no dumb question, or that their opinions are important and vital to the success of the company and to the customers that they serve, this fosters a psychological safe workplace. By fostering an environment of learning from each other and that there will not be any negative consequences from voicing their opinions, this also reduces anxiety. An environment where it is expected for team players to ask questions and come up with ideas for solutions to problems will reduce anxiety of people who may feel unsafe to speak up (Edmondson, 2018). As managers, if we discuss our imperfections and vulnerabilities, this also will reduce anxiety among other team members.

Themes to look for:

- Culture in a working environment
- Learning
- Non-blaming culture
- Administrative leadership
- Humility

FIGURE 3.3: An essay highlighted by EssayIQ system

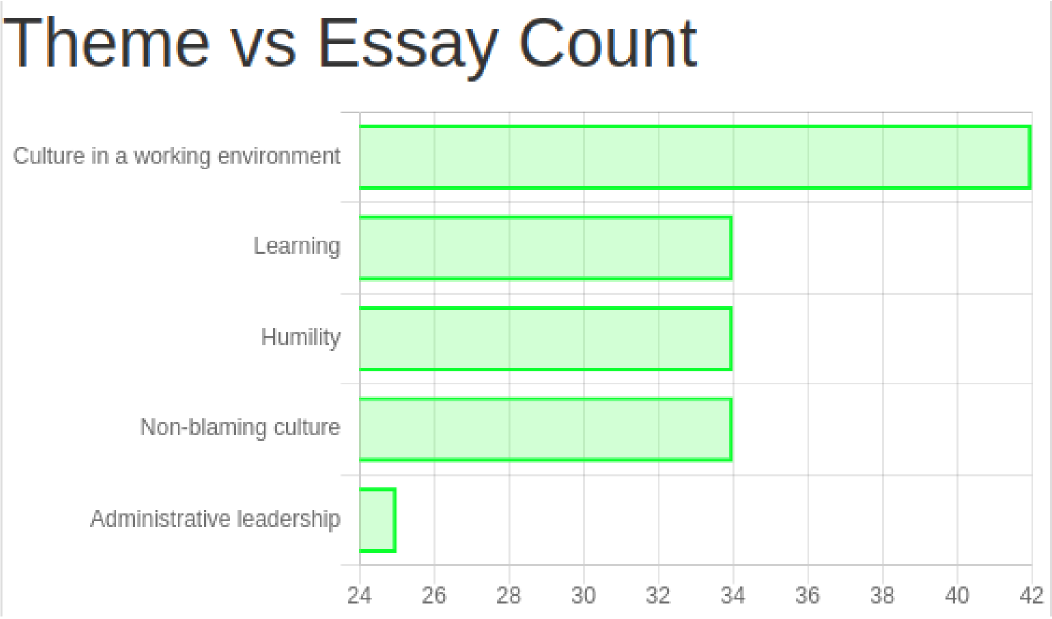


FIGURE 3.4: A plot where x-axis represents essay counts in an assignment and y-axis represents themes belonging to that assignment

3.2.1 Developing EssayIQ using Phrase2Vec model

In this subsection, we will explain the details of building core text processing component of EssayIQ using word2vec for phrases.

Data source

Wiki data(*kaggle reference*) has been used to build Phrase2Vec model. This dataset is a collection of 7.8 million sentences. The range of number of characters in these sentences is 4-255.

Training Phrase2Vec using gensim‘s Phrases

As typical word2vec is not going to be effective for this research task, we have applied gensim‘s Phrases library(*Phrase model*) in multiple stages to eventually build a refined wiki dataset from the original wiki data. The refined wiki dataset can contain phrase tokens upto 4-grams because we applied gensim‘s Phrases model three times in subsequent steps. The first step involves processing original wiki sentence data using ‘Phrases’ model to modify the data to contain potential bi-grams. By following the same principle, the refined data on second stage will contain tokens up

TABLE 3.1: Transformation of a sentence through multiple stages of Phrase Model

Transformation type	**sentence content**
initial wiki sentence	Although Purdue began competing in intercollegiate football in 1887, the school's official record book considers the "modern era" to have begun in 1946.
after 1st phase (bi-gram phase)	purdue began_competing intercollegiate football 1887 school official record book considers modern_era begun 1946
after 2nd phase (tri-gram phase)	purdue began_competing_intercollegiate_football school official record book considers modern_era begun
after 3rd phase (4-gram phase)	purdue begancompetingintercollegiate-football_school official record book considers modernera begun

TABLE 3.2: Parameter settings that have been used to build Phrase Model

PhraseType	**min_count**	**threshold**	**max_vocab_size**	**scoring type**
upto 2-gram	8	0.2	800, 000	npmi
upto 3-gram	8	0.2	800, 000	npmi
upto 4-gram	8	0.2	800, 000	npmi

to 3-grams. We did refinement for one more step to have tokens up to 4-grams in the final refined wiki data. NPMI (Normalized (Pointwise) Mutual Information) scoring method has been used in Phrase model. Min_count param has been set to 8 to ignore all words and bigrams below this value. The formula for NPMI score between two tokens **a** and **b** are defined in equation 3.1 and 3.2.

$$NPMI(a,b) = \frac{\frac{\ln P(a,b)}{P(a)*P(b)}}{-\ln P(a,b)} \tag{3.1}$$

where

$$P(a) = \frac{count\ of\ a}{corpus\_word\_count} \tag{3.2}$$

In table 3.1 we highlighted the changing process of a wikipedia sentence as it goes through "Phrase model" several times.

TABLE 3.3: Word2Vec Model using n-grams wiki data

PhraseType	**min_count**	**vocab_size**	**vocab size increased by**
upto 2-gram	8	256, 372	25,676 (# 2-grams)
upto 3-gram	8	262, 755	6383 (# 3-grams)
upto 4-gram	8	265, 552	2797 (# 4-grams)

Phrase2vec Results

In table 3.4, we presented a word similarity table for assignment M2V3. The terms (including bi-gram) were chosen from assignment themes. These terms also belong to vocabulary list of trained 4-gram Phrase2vec model. For different threshold, we show the number of relevant words closest with a term on first column, we also show the total number of words that appeared within the cosine distance threshold. One thing to note here that, a relevant word does not necessarily mean a closest semantic similar word with respect to candidate term. For example, the word 'good' is considered to be relevant to the word 'bad' although 'bad' may have many others closest semantic similar words such as 'inferior' or 'awful'.

3.2.2 Developing EssayIQ using *USE* model

Introduction of Universal Sentence Encoder (USE)

Universal Sentence Encoder (USE) is a family of models that encode texts to fixed-length vector representation. USE can be applied not only for word-level texts, but also for longer texts such as sentences, phrases and short paragraphs. It targets transfer learning to other NLP tasks such as text classification, semantic similarity, clustering etc. The paper (Cer et al., 2018) presented two models (Transformer and DAN) of *USE* where one model targets for higher accuracy and another model targets for higher efficiency. The study reports that transfer learning using transformer-based encoder performs equally or better than the DAN-based encoder. The embedding tensor produced by the trained encoder model can be directly used or incorporated into other NLP models for specific tasks. In our EssayIQ system, we have leveraged *USE* for semantic similarity purpose. In Figure 3.5, a heatmap has been shown to visualize semantic level sentence similarity scores. The authors of

TABLE 3.4: 20 essays are combined to create this evaluation. results are in the form of x/y where y denotes total number of similar words from essays within threshold distance score and x denotes the number of related words out of y words.

Terms	0.1	0.2	0.3	0.4	0.5	0.6
curiosity	1/1	1/1	2/2	6/10	6/10	6/10
learning	1/1	1/1	5/5	6/10	6/10	6/10
root_cause	0/0	1/1	1/1	6/10	6/10	6/10
encourage	2/2	4/6	7/10	7/10	7/10	7/10
question_asked	0/0	0/0	3/3	5/7	6/10	6/10
afraid	1/1	2/4	2/10	2/10	2/10	2/10
daily	1/1	2/2	2/2	2/3	6/10	6/10
daily_basis	0/0	0/0	1/1	1/1	3/10	3/10
conversation	1/1	1/1	5/6	8/10	8/10	8/10
open	1/1	1/1	1/1	2/3	2/10	2/10
communication	1/1	1/1	1/1	1/2	2/10	2/10
positive	1/1	2/2	3/6	4/10	4/10	4/10
positive_feedback	0/0	0/0	3/3	4/5	4/10	4/10
environment	1/1	1/1	1/1	2/10	2/10	2/10
inviting	1/1	2/2	2/3	2/10	2/10	2/10
participation	1/1	1/1	1/1	2/10	2/10	2/10
confidence	1/1	1/1	1/1	2/10	2/10	2/10
expectation	0/0	0/0	3/8	4/10	4/10	4/10
valuable	1/1	1/1	2/3	3/7	4/10	4/10
mistake	1/1	2/2	3/10	3/10	3/10	3/10
example	1/1	1/1	2/5	3/10	3/10	3/10
share_common	0/0	0/0	0/0	2/4	2/10	2/10
fear	0/0	0/0	2/4	3/10	3/10	3/10
admitting	1/1	2/2	2/5	4/9	4/10	4/10
insecurity	0/0	0/0	2/4	2/10	2/10	2/10
gain	1/1	2/2	2/7	3/10	3/10	3/10
guilt	0/0	0/0	1/5	2/10	2/10	2/10
support	1/1	2/2	4/4	4/10	4/10	4/10
policy	1/1	1/1	2/5	3/10	3/10	3/10
demonstrate	0/0	0/0	0/5	0/10	0/10	0/10

USE have evaluated the universal sentence encoder on STS benchmark (*STSbenchmark*) to identify the alignment of sentence embeddings with human judgements. The pearson correlation coefficient turned out to be 0.804 as a measure of defining quality for machine embeddings of STS dev set against human judgements for same dataset.

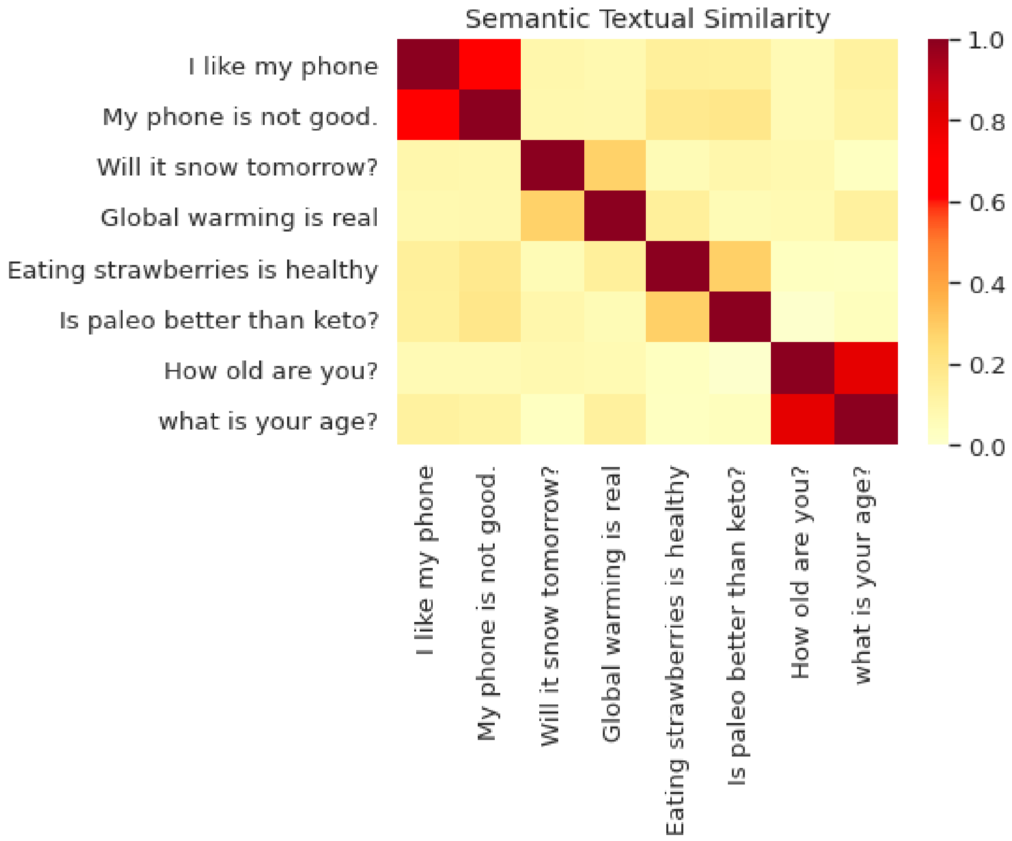


FIGURE 3.5: Heatmap of Sentence Similarity Score using Universal Sentence Encoder embeddings

Chapter 4
EXPERIMENT AND RESULTS

4.1 Data Source

Data for this research experiments comes from Reflection submissions made by students taking the course "Human Factors in Health Care" (NURS 3347 at University of Texas Arlington RN-BSN AO programs). This course is relatively new at UTA and it is highly valued by students as an elective course. This course is ideal for development and testing given that there are weekly multiple reflection assignments. For the purpose of development and testing, we have chosen three reflection assignments from this course. Student submissions to study reflection essays have been downloaded from course platform *Canvass*. Fifty most longest essay submissions have been chosen from each of the three assignment categories to make a total of 150 essays. A total of 300 words or less are expected from each essay submission. A brief summary of these three assignments are sketched in table 4.1.

TABLE 4.1: Testing assignments summary. Each assignment contains 50 essay submissions.

Assignment title	**Reflective questions**
Psychological safety (M2V3)	- How do you reduce anxiety of other people who may feel unsafe to speak up?
Cognitive aids and emergency manuals (M2V2)	- What types of human errors are targeted by emergency checklists? - Are there situations in your work place in which checklists can help to reduce human errors?
Design and human errors (M2V4)	- Have you encountered "tricky doors"? - Any other examples in or outside your workplace that illustrate human errors?

4.2 Data cleaning and storing in database

The essay submissions downloaded from course platform come in html format and contain student‘s identity. Due to privacy concern we removed student‘s identity from all submissions. HTML formatted submissions are parsed to retrieve only essay content. The code snippet of html parsing is provided in listing 4.1:

LISTING 4.1: HTML parsing code

```
    from bs4 import BeautifulSoup
    import glob
    files = glob.glob('./Materials/M2V3/*.html')
    for i, file in enumerate(files):
    f = open(file)
    soup = BeautifulSoup(f, features='lxml')
    tags = soup('p') #all the 'p' tags in a html
    max_size=0
    paragraph=''
    print i, file
    for tag in tags:
        if tag.string is not None:
            paragraph = paragraph + tag.string.
            encode('utf-8').strip()
            paragraph = paragraph + '\n'
    with open('./StudentEssay/M2V3/'+str(i)+'.txt', 'w')
                                                as f:
        f.write(paragraph)
```

Finally, all processed essay contents has been stored to PostgreSQL database using python in server. Code snippet to upload all submissions of an assignment is provided in listing 4.2:

LISTING 4.2: Code for Essay insertion in database

```
import psycopg2
import glob
```

```
connection = psycopg2.connect(
    host="localhost",
    database="————",
    port='5432',
    user="————",
    password='———',
)
connection.autocommit = True
cursor = connection.cursor()
files = glob.glob('./StudentEssay/M2V3/*.txt')
for i, fname in enumerate(files):
    f = open(fname, mode='r')
    all_of_it = f.read()
    username = '———'
    cursor.execute("""INSERT INTO submissions \
    ("userDisplayName", "submissionName", "submissionBody",\
    "userID", "assignmentID") values (%s, %s, %s, %s,\
    %s)""", ('———', fname[fname.rfind('/')+1:],\
    all_of_it, 1, 1))
```

Student‘s essay submissions alone can not make whole dataset for experiment purpose. The annotations of essay submissions are required to be used as gold standard of this research experiment. Two insightful and experienced coaches are appointed to annotate essays on sentence level. These two coaches annotate essays independently for a total of 150 essays from three assignments. Coaches are expected to spend 15 minutes for each essay. A video (*demo*) of coaches doing essay annotation can be found in youtube.

4.3 Experiment Details

This research project serves as a bridge between two sides of the spectrum. While one side covers the area of machine learning, natural language process and deep learning, another side covers visual analytics for online learning, and an interface to interact with machine learning side. The visual interface will be in "serve upon request" communication with backend machine learning side to analyze essay texts, identify themes.

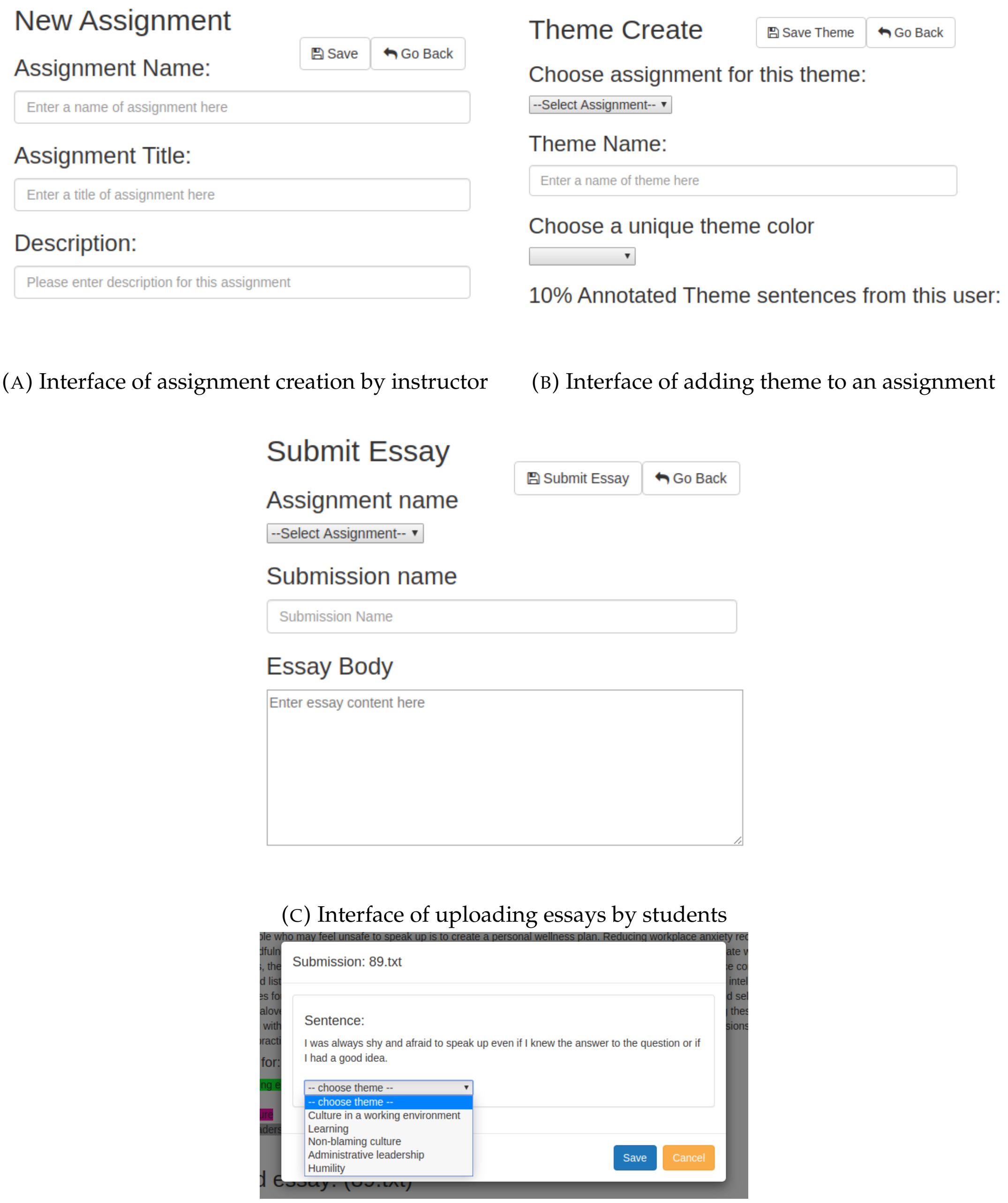
The *EssayIQ* is an end-to-end system in that it offers a communication bridge between two parties: teachers or instructors and students. In the system, an instructor can create assignments, themes pertaining to different assignments. Instructors can also provide customized feedback to students regarding presence of themes in essays by leveraging backend Universal sentence encoder api. In the same system, a student can submit his/her essay text for an assignment, can view the feedback from instructor regarding submission. The interface of creating new assignment, new theme and submitting essay for an assignment has been presented in figure 4.1.

4.4 Results and Discussion

4.4.1 Comparison between Phrase2Vec and USE

In Table 4.2, a comparison chart is sketched to show how Phrase2Vec and Universal sentence Encoder (USE) is performing in *EssayIQ* system. We see that, USE finds out much better semantically similar sentence from a certain essay compared to *Phrase2Vec*. We can also notice that cosine distance produced by *Phrase2Vec* model is relatively bigger than the cosine distance produced by *USE*.

Before we use human coach's annotation as gold standard for this experiment, we need to ensure that two coaches agree with each other regarding their sentence annotations. Experimentation on inter-coach agreement regarding annotation of assignment sentences has been shown in table 4.4. Our two coaches sentence annotation statistics is also reported in table 4.3. The range of kappa coefficient value that constitutes a good range is sketched in fig 4.2 from article (Viera and Garrett, 2005). In this experiment, first we choose an assignment and find the list of sentences from all submissions that have been annotated by two coaches. After that, we randomly choose fixed number (50, 75, 100) of sentences from the list of sentences to find kappa



I was always shy and afraid to speak up even if I knew the answer to the question or if I had a good idea.

(A) Interface of assignment creation by instructor

(B) Interface of adding theme to an assignment

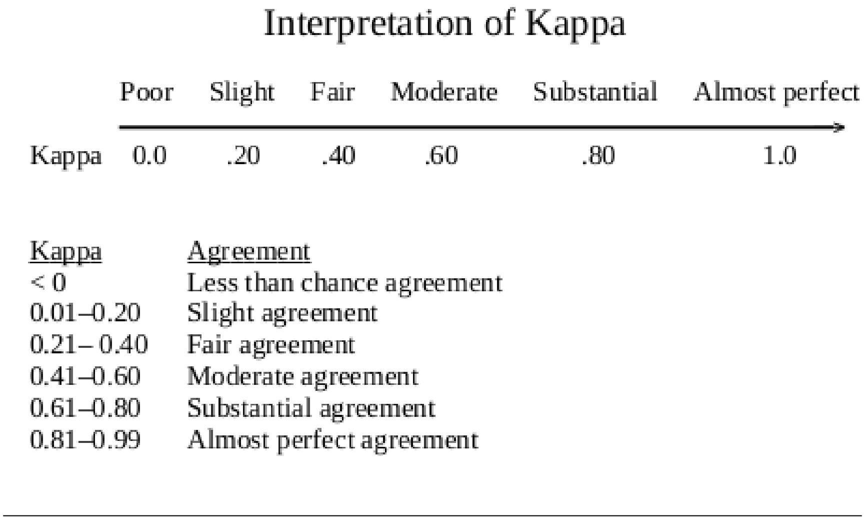
(C) Interface of uploading essays by students

(D) Interface of annotating/labeling sentence by coaches

FIGURE 4.1: Different components of *EssayIQ* web interface

Theme Sentence	**model**	**c_distance**	**Essay Sentence**
In step 2, I would admit to my own fallibility.	Phrase2Vec	0.29	Using situations I've been in and how I felt in those situations to explain that I really do understand how her day is going and I want to help. (**situation ive felt situation explain understand day going want help**)
	USE	0.58	1) Frame the work as a learning problem 2) Acknowledge your own fallibility.
Do they view me as approachable and open to discussion?	Phrase2Vec	0.31	If others that are afraid to ask questions or speak up, see that you are also vulnerable, it will help them to open up and model your actions. (**afraid ask question speak vulnerable help open model action**)
	USE	0.56	I am trying to be as open as possible with my co-workers and let them know that they always can talk to me.

TABLE 4.2: A theme sentence is chosen from annotated dataset. **Essay Sentence** column denotes the sentence content that has been chosen from a certain essay by two different models. In Phrase2Vec, sentence in bracket with bold text shows the actual 4gram sentence that were used. c_distance corresponds to cosine distance between theme sentence and essay sentence.


Interpretation of Kappa
Poor
Slight
Fair
Moderate
Substantial
Almost perfect
Kappa
0.0
.20
.40
.60
.80
1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

FIGURE 4.2: Table to visualize the meaning behind a kappa value

Assignment	total sentences	coach 1 count	coach 2 count	by both
M2V3	572	417	192	174
M2V2	685	360	253	198
M2V4	625	432	240	208

TABLE 4.3: Annotation details of two coaches for three testing assignments

coefficient score. This process is done for 100 times to calculate mean and standard deviation of that fixed sentence set.

There are two different ways to present kappa agreement between *EssayIQ* system and human annotator. One is "sentence level" kappa agreement. In this approach, all essay sentences are separately scanned to see if both *EssayIQ* and human attach a theme label to each sentence. Another approach is to apply essay or submission level agreement. In this version, each submission text is separately scanned to find out the sentences that were labeled by both *EssayIQ* and human and then a boolean vector having size of the number of assignment themes are formed to specify the themes that exist in submission text. In weighted quadratic version of essay level agreement, count vector is used instead of boolean vector to store the sentence counts belonging to each theme. "Sentence level" experiments have been plotted in table 4.5, table 4.6, and table 4.7 for three assignments. It is evident from these plots and human-human agreement plot that the more human coaches agree with their annotations in an assignment, the more kappa score from Universal Sentence

Assignment	num of sentences	num of runs	kappa score
M2V3	50	100	0.26 ± 0.08
	75	100	0.27 ± 0.06
	100	100	0.26 ± 0.04
	125	100	0.26 ± 0.03
M2V2	50	100	0.77 ± 0.04
	75	100	0.77 ± 0.04
	100	100	0.77 ± 0.04
	125	100	0.77 ± 0.03
M2V4	50	100	0.28 ± 0.06
	75	100	0.29 ± 0.06
	100	100	0.29 ± 0.04
	125	100	0.29 ± 0.03

TABLE 4.4: Inter-annotator agreement is represented by kappa score for sentences from three different assignments.

Encoder (*USE*_kappa column in the plot) we get in that assignment.

For the assignment M2V4, we were not able to do experiment with 15 or 20 theme sentences as there are two themes for which only 6 and 19 essay sentences have been annotated respectively. It might happen due to not many essay submissions touched on these two themes.

In Phrase2vec, the model trained to contain upto 4-grams has been used. We made an attempt to see if model built on upto 3-gram tokens performs better that 4-gram model. We did not find significant difference. In table 4.8, kappa scores are shown for 3-gram and 4-gram version of Phrase2vec model.

Another dimension of experiment is to do "essay level" theme existence experiment. In this case, we will not match themes between *EssayIQ* and *coach* on each and every single essay sentence, rather we will match theme existence on essay basis. If both *EssayIQ* and coach find a certain theme in an essay submission, then they agrees with each other regarding that theme. Table 4.10, 4.11, and 4.12 show kappa coefficient scores for both Universal Sentence Encoder and Phrase2vec model. Inter-coach agreement on essay level is depicted in table 4.9. From the different sentence level and essay level experiments, it can be seen that *EssayIQ* built with *USE* performs in proportionate with human. The agreement between *EssayIQ* and human annotator follows one kappa level behind the agreement between human-human in all three

num. of theme sentences	distance thresold	Phrase_kappa	USE_kappa
5	0.5	0.05	**0.19**
	0.6	0.06	0.11
	0.7	0.04	0.13
	0.8	0.08	0.12
10	0.5	0.09	0.07
	0.6	0.04	0.10
	0.7	0.10	**0.17**
	0.8	0.07	0.15
15	0.5	0.04	0.17
	0.6	0.03	0.18
	0.7	0.07	**0.19**
	0.8	-0.02	0.15
20	0.5	0.02	0.15
	0.6	0.0	**0.18**
	0.7	0.04	0.17
	0.8	0.05	0.13

TABLE 4.5: Assignment: **M2V3**. Human annotator 1 is chosen to be gold standard here. **Sentence level** theme prediction or annotation is considered for this part of experiment. Two types of kappa: one when using 4-gram **Phrase2vec** model and another when using **USE** for *EssayIQ* system. First column indicates number of candidate sentences used from each theme.

num. of theme sentences	distance thresold	Phrase_kappa	USE_kappa
5	0.5	0.10	0.21
	0.6	0.14	**0.38**
	0.7	0.14	0.32
	0.8	0.11	0.31
10	0.5	0.18	**0.45**
	0.6	0.17	0.34
	0.7	0.07	0.42
	0.8	0.14	0.33
15	0.5	0.12	0.44
	0.6	0.14	**0.45**
	0.7	0.21	**0.45**
	0.8	0.06	0.34
20	0.5	0.13	**0.42**
	0.6	0.19	0.38
	0.7	0.23	0.40
	0.8	0.12	0.34

TABLE 4.6: Assignment: **M2V2**. Human annotator 2 is chosen to be gold standard here. Same settings as like table 4.5

num. of theme sentences	distance thresold	Phrase_kappa	USE_kappa
5	0.5	-0.010	0.18
	0.6	0.04	0.17
	0.7	0.13	**0.26**
	0.8	0.05	0.06
10	0.5	0.13	**0.38**
	0.6	0.11	0.24
	0.7	0.05	0.22
	0.8	0.06	0.19
15	0.5	0.03	**0.25**
	0.6	0.08	0.24
	0.7	0.10	0.17
	0.8	0.15	0.23

TABLE 4.7: Assignment: **M2V4**. Human annotator 2 is chosen to be gold standard here. Same settings as like table 4.5.

num. of theme sentences	distance thresold	n-gram	Phrase_kappa
10	0.6	3-gram	0.10
	0.6	4-gram	0.17
	0.7	3-gram	0.18
	0.7	4-gram	0.07
15	0.6	3-gram	0.14
	0.6	4-gram	0.14
	0.7	3-gram	0.22
	0.7	4-gram	0.21

TABLE 4.8: Assignment: **M2V2**. Human annotator 2 is chosen as gold standard who annotated 253 sentences of this assignment. kappa score of phrase model when sentence level classification are used and when either 3-gram or 4-gram phrase model are used

assignments.

Assignment	**num of sentences**	**kappa score**
M2V3	50	0.44 | **0.59**
	75	0.47 | 0.54
	100	0.38 | 0.44
	125	0.39 | 0.47
M2V2	50	0.81 | 0.79
	75	0.82 | **0.87**
	100	0.81 | 0.87
	125	0.82 | 0.82
M2V4	50	0.43 | 0.37
	75	0.42 | 0.46
	100	0.41 | **0.48**
	125	0.32 | 0.34

TABLE 4.9: Inter-annotator agreement on essay level. Second column represents the number of sentences taken randomly that were labelled by both coaches from different essays. Unweighted and weighted kappa in last column.

num. of theme sentences	**distance thresold**	***Phrase*_kappa**	***USE*_kappa**
5	0.5	0.08 ∣ 0.13	0.22 ∣ 0.19
	0.6	-0.06 ∣ -0.02	**0.25** ∣ **0.33**
	0.7	0.02 ∣ 0.15	0.23 ∣ 0.29
	0.8	0.14 ∣ 0.25	0.08 ∣ 0.28
10	0.5	0.14 ∣ 0.27	0.19 ∣ 0.16
	0.6	0.11 ∣ 0.10	**0.29** ∣ 0.26
	0.7	0.06 ∣ 0.28	0.27 ∣ **0.48**
	0.8	0.15 ∣ 0.23	0.24 ∣ 0.46
15	0.5	0.13 ∣ 0.23	0.18 ∣ 0.18
	0.6	0.08 ∣ 0.27	0.20 ∣ 0.34
	0.7	**0.28** ∣ 0.22	0.25 ∣ **0.49**
	0.8	0.16 ∣ 0.12	0.16 ∣ 0.34
20	0.5	0.03 ∣ 0.11	**0.24** ∣ 0.37
	0.6	0.12 ∣ 0.12	0.15 ∣ **0.38**
	0.7	0.08 ∣ 0.11	0.18 ∣ 0.35
	0.8	0.10 ∣ 0.23	0.23 ∣ 0.34

TABLE 4.10: This is done for Assignment: **M2V3**. **Essay level** kappa score for both phrase model and USE model. In columns representing kappa, the number after vertical bar denotes weighted quadratic kappa.

num. of theme sentences	**distance thresold**	***Phrase*_kappa**	***USE*_kappa**
5	0.5	0.29 | 0.28	0.27 | 0.31
	0.6	0.01 | 0.16	**0.43** | **0.57**
	0.7	0.19 | 0.14	0.26 | 0.32
	0.8	0.06 | 0.20	0.30 | 0.51
10	0.5	0.17 | 0.36	**0.49** | 0.55
	0.6	0.17 | 0.25	0.31 | 0.36
	0.7	0.26 | 0.34	0.41 | **0.62**
	0.8	0.17 | 0.18	0.25 | 0.49
15	0.5	0.10 | 0.25	0.43 | 0.49
	0.6	0.23 | 0.29	0.43 | 0.54
	0.7	0.31 | 0.32	**0.49** | **0.63**
	0.8	-0.07 | 0.03	0.30 | 0.49
20	0.5	0.13 | 0.15	**0.50** | 0.49
	0.6	0.27 | 0.48	0.39 | 0.54
	0.7	0.36 | 0.43	0.45 | **0.59**
	0.8	0.11 | 0.29	0.41 | 0.44

TABLE 4.11: Assignment: **M2V2**. Essay level kappa score for both phrase model and USE model.

num. of theme sentences	distance thresold	*Phrase*_kappa	*USE*_kappa
5	0.5	0.02 | 0.08	0.18 | 0.23
	0.6	0.14 | 0.14	0.34 | 0.35
	0.7	0.19 | 0.37	**0.43** | **0.60**
	0.8	0.31 | 0.46	0.24 | 0.29
10	0.5	0.17 | 0.33	0.52 | 0.55
	0.6	0.21 | 0.45	0.43 | **0.66**
	0.7	0.30 | 0.44	**0.52** | 0.60
	0.8	0.21 | 0.43	0.27 | 0.42
15	0.5	0.28 | 0.38	**0.46** | **0.63**
	0.6	0.25 | 0.42	0.39 | 0.54
	0.7	0.26 | 0.40	0.23 | 0.32
	0.8	0.32 | 0.51	0.42 | 0.47

TABLE 4.12: This is done for Assignment: **M2V4**. Essay level kappa score for both phrase model and USE model.

Chapter 5
CONCLUSION

Deep learning models built for the task for semantic similarity calculation is better suited for *EssayIQ* system compared to word2vec based models. The *USE* model in EssayIQ aligns better with human annotator than *Phrase2vec* model.

In this thesis, we have made an preliminary attempt to develop and test a text analytics system to analyze open-ended reflection texts for online learning environment with large student enrollment. One of the purposes of reflection text analysis is to provide customized feedback to learners using feedback template through *EssayIQ* system. Our experiments perform quantitative evaluation of two models to show the effectiveness of finding themes in essay texts and suggesting feedback in a time-efficient manner. Despite the superiority of USE model, we need to improve it so that it can handle theme prediction and feedback suggestion with very close to human level performance.

Chapter 6
FUTURE WORK

Our developed *EssayIQ* system is a novel text analytics system capable of analyzing open-ended texts with customizable domain knowledge input. However, it is a quite primitive work in this research direction. Our longer term goal is to develop an efficient and scalable tool for online learning environment with a strong knowledge base for reflection debriefing text assignments. So far, our system can aid in providing feedback by plotting assignment analysis summary and by linking essay sentences to the closest theme knowledge base. Our longer-term plan includes automated feedback generation to learners in a wide range of online learning environments. Our current work is not capable of detecting multiple themes in a sentence. We want to further improve *EssayIQ* with the capability of highlighting in multiple clauses of single sentence or even in same clause.

Another way we want to improve *EssayIQ* is by employing phrase and word weight based sentence semantic similarity metric. Some words and phrases might be exclusive to certain themes. Higher weights will be given to those words and phrases present in essay sentences.

References

Altoe, F. and D. Joyner (2019). “Annotation-free Automatic Examination Essay Feedback Generation”. In: *2019 IEEE Learning With MOOCS (LWMOOCS)*, pp. 110–115. DOI: 10.1109/LWMOOCS47620.2019.8939630.

Attali, Yigal and Jill Burstein (2006). “Automated Essay Scoring With e-rater® V.2”. In: *The Journal of Technology, Learning and Assessment* 4.3. URL: https://ejournals.bc.edu/index.php/jtla/article/view/1650.

Cer, Daniel et al. (2018). *Universal Sentence Encoder*. arXiv: 1803.11175 [cs.CL].

demo. https://youtu.be/pDj9McNiaY8.

Foltz, Peter, Darrell Laham, and T. Landauer (Apr. 1999). “The intelligent essay assessor: Applications to educational technology”. In: *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*.

Hastings, Peter, Simon Hughes, and M. Anne Britt (June 2018). “Active Learning for Improving Machine Learning of Student Explanatory Essays”. In: pp. 140–153. ISBN: 978-3-319-93842-4. DOI: 10.1007/978-3-319-93843-1_11.

kaggle reference. https://www.kaggle.com/mikeortman/wikipedia-sentences/data.

Li, Li and Vijayan Sugumaran (Mar. 2018). “A cognitive-based AES model towards learning written English”. In: *Journal of Ambient Intelligence and Humanized Computing* 10. DOI: 10.1007/s12652-018-0743-1.

McNamara, Danielle et al. (Nov. 2014). “A hierarchical classification approach to automated essay scoring”. In: *Assessing Writing* 23. DOI: 10.1016/j.asw.2014.09.002.

Mikolov, Tomas et al. (2013). “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., pp. 3111–3119. URL: http:

//papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mittal, Himani and Mandalika Devi (Apr. 2016). "Machine Learning Techniques with Ontology for Subjective Answer Evaluation". In: *International Journal of natural language computing* 5. DOI: 10.5121/ijnlc.2016.5201.

Park, D. et al. (2018). "ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding". In: *IEEE Transactions on Visualization and Computer Graphics* 24.1, pp. 361–370.

Park, Deokgun et al. (2016). "Supporting Comment Moderators in Identifying High Quality Online News Comments". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, 1114–1125. ISBN: 9781450333627. DOI: 10.1145/2858036.2858389. URL: https://doi.org/10.1145/2858036.2858389.

Phrase model. https://radimrehurek.com/gensim/models/phrases.html.

Phrase2vec. https://towardsdatascience.com/word2vec-for-phrases-learning-embeddings-for-more-than-one-word-727b6cf723cf.

Reflective writing. https://en.wikipedia.org/wiki/Reflective_writing.

Shermis, Mark D. and Jill Burstein (2003). "Automated Essay Scoring : A Cross-disciplinary Perspective". In:

STSbenchmark. http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark.

Taghipour, Kaveh and Hwee Tou Ng (Nov. 2016). "A Neural Approach to Automated Essay Scoring". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1882–1891. DOI: 10.18653/v1/D16-1193. URL: https://www.aclweb.org/anthology/D16-1193.

Valenti, Salvatore, Francesca Neri, and Alessandro Cucchiarelli (2003). "An Overview of Current Research on Automated Essay Grading". In: *J. Inf. Technol. Educ.* 2, pp. 319–330.

Viera, Anthony and Joanne Garrett (June 2005). "Understanding Interobserver Agreement: The Kappa Statistic". In: *Family medicine* 37, pp. 360–3.