Using Word Embeddings to Analyze Teacher Evaluations: An Application to a Filipino Education Non-Profit Organization

Francesca Vera
Stanford University
Stanford, CA 94305
fvera@alumni.stanford.edu

Abstract

Analysis of teacher evaluations is crucial to the development of robust educational programs, particularly through the validation of desirable qualities being reflected on in the text. This research applies Natural Language Processing techniques on a real-world dataset from a Filipino education non-profit to explore insights from analyzing evaluations written by Teacher Fellows who assess their own progress. Prior to this research, only qualitative assessment had been conducted on the text. Inspired by the use of word embedding similarities to capture semantic alignment, we utilize GloVe embeddings to determine to what extent these evaluations reflect concepts critical to measuring the competency of Teacher Fellows and upholding the organization's Vision and Mission. As Fellows' quantitative ratings improved, so too did their demonstration of competency in the text. Further, Teacher Fellow language was consistent with the organization's Vision and Mission. This research therefore showcases the possibilities of NLP in education, improving our understanding of Teacher Fellow evaluations, which can lead to advances in program operations and education efforts.

1 Introduction

Applying Natural Language Processing (NLP) techniques to improve the quality of education programs is a crucial step in ensuring the NLP community's contributions to Social Good. Utilizing NLP unlocks the potential of computationally examining texts that were once only qualitatively analyzed or overlooked because of the difficulty in assessing the text. Textual data is plentiful in an educational setting, ranging from comments about student experiences to teacher evaluations and reflections. We look at applying NLP processes to teacher evaluations, which are a basis for documenting teacher growth and performance – components that will impact the quality of the education students receive.

The main contribution of this paper is that it introduces a computational framework through which teacher evaluations can be analyzed, so that insights gained can enhance educational programs. Applied on a real-world dataset from a Filipino education non-profit organization, word embedding similarities reveal which desirable traits within teachers are contained in these Teacher Fellow (a term used by the organization to describe its teachers in training) evaluations and their alignment with the organization's Vision and Mission. The motivation behind this work was to determine what qualities Teacher Fellows embody at different stages, as well as determining if their self-reflection was calibrated with manager (those training Teacher Fellows) evaluations. The organization was also curious if certain competencies emerged more than others.

With this application, we hope to improve the organization's teacher development efforts. Program changes that this piece of work inspired included refinement of the organization's evaluation tools and prompts, reflection on their competency definitions, greater discussions between Teacher Fellows and managers, alignment of interventions provided to Teacher Fellows, and increased support in the overall journey. We envision that this research can be scaled and flexibly applied to other bodies of textual data to serve future educational and Social Good purposes.

2 Related Work

2.1 NLP in Education

The uses of NLP in the education space include: to better understand natural language learning; to improve teaching materials; to develop learning applications; and to enhance student output, as out-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 382–389
August 1–6, 2021. ©2021 Association for Computational Linguistics

lined by Dr. Alhawiti (2014) in a survey of the state of NLP applications in education. Relevant to teacher evaluations, research by Rajput et al. (2016) implements a lexicon-based sentiment analysis tool on these evaluations to gain more insight on student feedback. Not only did the tool prove highly correlated with quantitative ratings, but also it provided clearer understanding of teacher performance through its sentiment score. Further, Tzacheva et al. (2019) present a model that detects multiple emotions within student comments on teacher performance in order to increase positivity in student emotions and improve student experience in the classroom. Our research differs from past NLP explorations in that it is mostly concerned with teacher self-evaluation and employs the use of word embeddings to derive a set of pre-defined concepts reflected in the text.

2.2 Word Embedding Similarities

Word embeddings are vector representations of words that capture their semantic features. These vectors can be assessed in relationship to one another through Euclidean distance or cosine similarity, used interchangeably to measure the semantic similarity (Pennington et al., 2014) between the words whose corresponding vectors are being compared. Embedding similarity to affirm the semantic alignment in language is a common technique in NLP. In exploring the semantic similarity between two texts, Kenter and de Rijke (2015) propose the proximity between embeddings as related to semantic proximity. To quantify gender and ethnic stereotypes using embeddings, Garg et al. (2017) computed the Euclidean distance between group words and neutral words to measure the strength of association between the two sets. Similarity can also be calculated on a sentence level, even when the structure of the sentence is ignored, but the associated embeddings of the words within the sentence are considered by averaging. (Faruqui et al., 2014; Yu et al., 2014) We adopt a combination of these past uses of word embeddings by calculating the similarity of sets of words to determine semantic proximity between them, where one of the sets contains sentences for which an associated embedding must be assigned.

3 Dataset

This paper applies NLP techniques on a real-world dataset from a Filipino education non-profit organization since the data consisted of many textual entries. The organization continues to conduct qualitative assessment on this particular dataset, but this paper explores the first application of quantitative analysis on it.

The dataset was made up of Summer Institute (SI) reflections, mid-year Competency Based Evaluations (CBE's), and end-year Competency Based Evaluations. CBE's allow a Teacher Fellow the opportunity to reflect on their progress, as well as provide a formal space for an instructional coach's assessment. Split into four domains: Personal Leadership, Servant Leadership, Change Management, and Critical Learning, CBE's prompt the Fellow to write on "Critical Incidents, Strengths, and Areas of Growth" and give themselves a rating out of 4.0 per domain. A coach repeats the process based on their assessment of the Fellow. Overall, each CBE consists of 8 text entries and 8 ratings. SI reflections differ from CBE's in that the prompts are about high-level, philosophical ideas in education and there are no quantitative ratings.

For the sake of consistency, only Fellows who had SI reflections, mid-, and end-year CBE's were chosen. There were 14 such Fellows. These evaluations were collected in 2019, making it still relevant to the present-day programs of the organization. In total, there were 126 unique text entries: 14 SI reflections, 56 CBE entries by Teacher Fellows, and 56 CBE entries by instructional coaches. Alongside these text entries were 56 ratings by Teacher Fellows and 56 ratings by instructional coaches.

Although the dataset only consisted of 14 Fellows, there were 126 text entries of average length 204 words, which was significant enough to move forward with exploring this application.

4 Method

4.1 Keyword List Development

The organization developed keyword lists to encapsulate the ideas that it was interested in analyzing. A Data and Impact Assessment Manager created these lists based on three important aspects: Competencies, Core Values, and Mission. These Competencies are the listed "indicators" that Fellows should develop throughout the duration of the program and after completion. They were defined by the organization's Fellowship Program team, who have years of experience and pivotal knowledge of the Fellows' impact on students and communities. The "indicators" were condensed into lists of Competency keywords by careful selection of words that carried the most meaning. A similar process was done to create Core Values and Mission lists, which represent the organization's Vision and Mission. Core Values were also embedded in the "indicators," so keywords were extracted from there. The Mission keyword list came directly from the organization's Mission statement. Because the Data and Impact Assessment Manager was not on the Program team that wrote the "indicators," keyword selection was relatively objective, ensuring that certain concepts were not favored nor dismissed.

Although there may be subjectivity within keyword choices, common NLP tasks exist where keywords are grouped by theme into sets (e.g. determining which occupation words are neutral in gender bias embedding explorations (Garg et al., 2017)). The method of having an internal expert choose keywords that represented competencies based on pre-defined qualities that Teacher Fellows should achieve most closely mirrors methods for past NLP datasets where experts from other fields have hand-labeled entries (e.g. cross-validating entries in hate speech datasets (Jha and Mamidi, 2017)) – despite the subjectivity in this exercise. Most importantly, the organization fealt confident that the keywords chosen accurately reflected its desired competencies, core values, and mission.

4.2 Embedding Similarity Calculation

Inspired by the concept that cosine similarity between embeddings measures semantic similarity between words, we calculate the similarity between the keyword lists and teacher evaluations.

For each keyword, we find the keyword embedding. If the keyword was present in the predefined embeddings, we used that embedding. In some cases, however, there were multiple words that made up a keyphrase; for these items, we took the average of embeddings for every word in the keyphrase and assigned that as the embedding. Next, we find the embedding for the teacher evaluations, which we will refer to as the evaluation embedding. Each evaluation was treated as a single document of text with only one assigned embedding. In the pre-processing of this text, punctuation was stripped, spelling was checked through Microsoft Word, and documents were tokenized. Similar to what was done with keyphrases, we averaged the embeddings of every token, and this average was taken as the evaluation embedding.

Competency Domain	Mid-Year Similarity	End-Year Similarity
Personal Leadership	0.498	0.523
Servant Leadership	0.603	0.608
Change Management	0.531	0.532
Critical Learning	0.599	0.600

Table 1: Fellow Competency Similarity Scores

Competency Domain	Mid-Year Similarity	End-Year Similarity
Personal Leadership	0.519	0.523
Servant Leadership	0.632	0.634
Change Management	0.557	0.560
Critical Learning	0.631	0.629

Table 2: Coach Competency Similarity Scores

After finding the keyword embedding and the evaluation embedding, we calculate their cosine similarity, which becomes their similarity score. This process was repeated for every evaluation and keyword. To find the overall score for Competencies, Core Values, or Mission, we average across all Fellow-keyword similarity scores per list.

4.3 Experiments

The following experiments were conducted on the dataset with GloVe 50-dimensional embeddings. A similarity score was calculated for each Fellow across the organization's Competency domains: Personal Leadership, Servant Leadership, Change Management, and Critical Learning, which had unique keyword lists each. This process was repeated for Fellows' and coaches' mid-year and end-year CBE entries, resulting in 16 similarity scores (tables 1 and 2). We also include the Fellows' and coaches' quantitative ratings out of 4.0 from the CBE's (tables 3 and 4) for later comparison. The Competency keywords with the highest similarities are listed out (tables 5 and 6). We calculated the similarity between Fellow evaluations and the Core Values and Mission lists to measure alignment with Vision and Mission. This was done on the Fellows' SI reflections, mid-year CBE's, and end-year CBE's (table 7). We also list the top keywords for Core Values and Mission (table 8).

Competency Domain	Mid-Year Score	End-Year Score
Personal Leadership	2.619	2.833
Servant Leadership	2.381	2.524
Change Management	2.214	2.393
Critical Learning	2.250	2.536

Table 3: Fellow CBE Ratings

Competency Domain	Mid-Year Score	End-Year Score
Personal Leadership	2.500	2.786
Servant Leadership	2.191	2.548
Change Management	2.143	2.446
Critical Learning	2.321	2.750

Table 4: Coach CBE Ratings

5 Discussion of Results

Cosine similarity has maximum score 1.0, which occurs when perfect similarity is achieved; higher cosine similarity indicates higher semantic similarity. Thresholds differ according to vector space and can be calculated through a variety of methods. We employ a cutoff of > 0.37 (Orkphol and Yang, 2019) to indicate similarity based on a predictive model validated by human perception of relevance, as human intuition of similarity is fundamental to how our application would be used in practice. However, we acknowledge that to achieve a truly appropriate cutoff, we would have to replicate the empirical study with our own relevant prompts to generalize for this task.

Fellow and coach mid-year and end-year CBE's display Competency similarity scores above our chosen threshold, indicating that their essence was present in the evaluations. We also notice an improvement in all Competency domains from mid-year to end-year evaluations for Fellows and coaches. This upward trend mirrors the increase in quantitative ratings, suggesting that as Fellows improve in practice, their reflections become more aligned with the domain concepts. The domain with highest similarity was Servant Leadership, while the lowest was Personal Leadership, although the differences between similarities are minimal. It is important distinguish what is reflected on in the evaluations from how the Fellows are rated in practice: for example, Personal Leadership ratings were among the highest out of 4.0, but the similarity scores were the lowest, meaning there may be inconsistencies between what is written and how Fellows are rated. Coaches produced higher similarity scores than Fellows did, which is expected due to coaches who speak more explicitly in terms of the Fellow's competency.

Clear themes emerge from looking at the top keywords for Competency domains. Personal Leadership focuses on life skills that Fellows may improve upon. For Servant Leadership, there is emphasis on the relationship aspects of teaching. Because Change Management is concerned with executing plans for the larger community, the top keywords express big picture thinking. In Critical Learning, the topics cover knowledge and lesson formation.

Personal Leadership	Servant Leadership
Manages Time	Relationship Building
Personal Development	Strong Relationships
Work Habits	Common Goals
Improve	Shared Goals
Describes Drives	Positive Relationships

Table 5: Top 5 Competency Keywords pt. 1

Change Management	Critical Learning
Higher Order Thinking	Working Knowledge
Effective Plans	Lesson Plan
Systems Thinking	Subject Matter
Big Picture	Deliver Lessons
Considering	Subject Content

Table 6: Top 5 Competency Keywords pt. 2

Since the Core Values and Mission keyword lists captured the Vision and Mission of the organization, the significant similarity scores across SI reflections, mid-year, and end-year CBE's indicate that the Vision and Mission was expressed in the Fellows' writing. The top 3 keywords for Core Values are especially salient, as "get the job done" was consistent with the Critical Learning Competency domain, and "working with others" and "working as a team" directly related to Servant Leadership.

5.1 Debrief with the Organization

We set up a formal discussion with the organization's Fellowship Program team, whose expertise in the program, evaluations, and Fellow performance led us to further insights. They endorsed the outcome that Servant Leadership produced the highest similarity scores. They theorized that the interpersonal and community aspects of Servant Leadership were already strong within this cohort of Fellows. When looking at the top 5 keywords for Servant Leadership, the themes of fostering relationships and shared goals were unsurprising to them. The second highest scoring domain, Criti-

Keyword List	SI	Mid-Year CBE	End-Year CBE
Core Values	0.686	0.677	0.679
Mission	0.637	0.620	0.623
Competency	0.565	0.577	0.566

Table 7: Core Values and Mission Similarity Scores

Core Values	Mission
Get the Job Done	Programs
Working with Others	Life Skills
Working as a Team	Developing Students

Table 8: Top 3 Core Values and Mission Keywords

cal Learning, also lined up with their expectations; because this domain is specifically concerned with pedagogical knowledge, Critical Learning tends to be discussed at length, as would be typical in standard teacher evaluations. Further, they said it made sense that coaches produced higher similarity scores than Fellows did since coaches are not only more familiar with concepts, but are also given templates for being explicit in their writing.

The Program team appreciated the conceptual alignment between the evaluations and the Core Values and Mission, claiming that throughout the training process, coaches encourage Fellows to incorporate these components in their practice. After looking at the top 3 Mission keywords, the Program team emphasized that Fellows are coached on "life skills" and "developing students" – two of the most important components of the program.

They noted that one area of improvement that would be easily implemented in future applications of this process would be to update keyword lists as the "indicators" themselves get updated. Because this project inspired reflection within the organization regarding what qualities it is evaluating teachers on, any further analysis should consider appropriate changes in competencies.

It may also be important to note that the Fellowship Program team's enthusiasm for this project meant that we received dynamic feedback during the entire process.

6 Conclusion

We present an application of word embedding similarities to evaluate how Teacher Fellow evaluations align with requirements for Competency and a non-profit's Vision and Mission. This analysis adds a new element to the interpretation of textual data that would have otherwise been only qualitatively examined. Moving forward, the organization aims to apply this method to other batches of Fellows. Other members of the organization, such as Admissions and Alumni Program teams, also conduct evaluations on which the method could be applied. A comparison may be conducted to see what certain cohorts were more willing to reflect on in the text and how their focuses differ. More sophisticated models, including Tf-idf weighting and a higher level of pre-processing, are possible, and further validation through correlation with quantitative ratings can help clarify the value of this model. Having demonstrated that Fellows' evaluations captured ideas contained within Competency, Core Values, and Mission keywords, we are confident that the organization can incorporate this analysis to achieve its goal of providing children access to high quality education across the Philippines.

Acknowledgments

I would like to acknowledge the Filipino education non-profit organization that provided the project scope and dataset that shaped this paper. In particular, Gaby, Ed, and Jo – who offered valuable collaboration, insights, and guidance.

Ethics and Positive Impact Considerations

NLP for Positive Impact aims to promote innovative ways NLP research will positively impact society, and this paper presents a case study of an NLP application for the social good cause of education. We consider the impact of this paper as positive because the improvements to the non-profit's programs (due to this research) will advance the organization's mission to provide Filipino children with the highest standard of education possible. This paper may also inspire positive impact in other situations where similar applications of NLP could be used for social good.

There is also an opportunity to incorporate the UN theme for International Women's Day, "Women in Leadership," because of the relationship between education and empowering women. It is worth noting that in the Philippines, teachers are mostly women, so this research could inform their professional leadership journey, including how to better train and support them. We envision many possibilities for this application in a profession dominated by women in the Philippines.

For context, the Filipino education non-profit focuses on improving teacher quality and addressing system-level educational challenges. This organization that provided the data allowed the submission of this paper. We focus on its Fellowship Program that trains Teacher Fellows to significantly improve student learning outcomes. Because the Fellowship Program is concerned with the growth of Teacher Fellows in their ability to deliver lessons of high quality and create positive change in educational communities, analyzing their evaluations is a worthwhile task in continuing the efficacy of this program. The Fellowship Program reaches over 10,000 public school students in the Philippines annually. Since the organization would be the main beneficiary of this technology, this technology has the potential to improve the Fellowship Program, which in turn will impact the Teacher Fellows and their students. A secondary beneficiary would be other similar non-profit organizations or operations teams that run educational programs.

The dataset used in this application was directly provided to us by the organization's Data and Impact Assessment team. The handover of data was formally documented and a non-disclosure agreement was signed. Further, we met with the Instructional Coaching Team to confirm the consent of the original authors of the texts. With regards to dataset privacy considerations, per the request of the organization and to protect the privacy of the Teacher Fellows and instructional coaches, the evaluations were completely anonymized. For the same reasons, the data cannot be submitted for review or replication. We chose not to include direct quotations from the evaluations (which could have been used to strengthen discussion of results) to ensure the authors' privacy.

An overall evaluation of using word embeddings for similarity tasks was considered. (Faruqui et al., 2016) Further, word embeddings have been shown to carry biases (e.g. gender and ethnic stereotypes) (Garg et al., 2017), so these biases may manifest in the outcomes. Because the similarity threshold was defined by a model based on human perception, the biases of the human participants may also factor into the resulting value. We acknowledge that the research was conducted in English, which may exclude other languages where word embeddings in those languages cannot be applied in a similar way.

References

Khaled Alhawiti. 2014. Natural language processing and its use in education. *International Journal of Advanced Computer Science and Applications*, 5:72–76.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. 2014. Retrofitting word vectors to semantic lexicons. *CoRR*, abs/1411.4166.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word embeddings quantify 100 years of gender and ethnic stereotypes. *CoRR*, abs/1711.08412.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Tom Kenter and Maarten de Rijke. 2015. Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1411–1420, New York, NY, USA. Association for Computing Machinery.

Korawit Orkphol and Wu Yang. 2019. Word sense disambiguation using cosine similarity collaborates with word2vec and wordnet. *Future Internet*, 11(5).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Quratulain Rajput, Sajjad Haider, and Sayeed Ghani. 2016. Lexicon-based sentiment analysis of teachers' evaluation. *Appl. Comp. Intell. Soft Comput.*, 2016:1.

Angelina Tzacheva, Jaishree Ranganathan, and Rajendra Jadi. 2019. Multi-label emotion mining from student comments. In *Proceedings of the 2019 4th International Conference on Information and Education Innovations*, ICIEI 2019, page 120–124, New York, NY, USA. Association for Computing Machinery.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *CoRR*, abs/1412.1632.

A Appendix

A.1 Keyword Lists

A.1.1 Personal Leadership

Personal Leadership, Self-Management, Self-Awareness, Self-Regulation, Motivation, Responsibility, Personal Development, Excellence, Describes Emotions, Describes Drives, Describes Values, Personality, Manages Resources, Manages Energy, Manages Emotions, Manages Time, Deliverables, Work Habits, Quality Output, Personal Motivations, Goals, Persevering, Rallying, Challenges, Setbacks, Constructive Feedback, Improve

A.1.2 Servant Leadership

Servant Leadership, Relationship Building, Strong Relationships, Healthy Relationships, Rewarding Relationships, Common Goals, Positive Relationships, Respect, Humility, Empathy, Expresses Clearly, Influences, Shared Goal, Expresses Effectively

A.1.3 Change Management

Change Management, Planning, Strategic, Effective, Ambitious, Realistic, Short-Term Goals, Mid-Term Goals, Long-Term Goals, Community, Effective Plans, Resources, Vision, Education Reform, Collaboration, Stakeholders, Scale, Innovation, Creative, Valuable, Sustainable, Enhancement, Community, Capacity Building, Collaborating, Respecting, Considering, Existing Practices, Existing Procedures, Resources, Efforts, Higher Order Thinking, Creative, Divergent, Convergent, Critical, Analytical, Excellence, Systems Thinking, Big Picture, Social, Political, Cultural, Evidence, In-Depth, Effective Decisions, Time-Critical, Communities, Factors, Proposes, Contextualized, Logical, Perspectives, Execution

A.1.4 Critical Learning

Critical Learning, Content Knowledge, Subject Matter, Contextualized, Applied, Learning, Teaching, Measurable Changes, Broadened Opportunities, Working Knowledge, Subject Content, Lesson Plan, Mastery, Delivery Lessons, Community Engagement Activities, Captivate, Care, Classroom Management, Confer, Teaching Framework, Pedagogical Knowledge

A.1.5 Core Values

Excellent Education, Inclusive Education, Relevant Education, Working as a Team, Excellent Results, Listens, Learns, Get the Job Done, Acts with Respect, Acts with Kindness, In the Face of Ambiguity, Builds Strong Partnerships, Values Strong Partnerships, Collaboration, Humility, Success of Whole, Integrity, Values Consensus, Shared Outcomes, Accepting, Sets Aside Personal Ego, Respect, Positive Relationships, Responsibility to Learn, Extend Helping Hand, Kindness, Inside and Outside Organization, Individual Role, Maximize Learning, Effectively Performs, Takes Action, Different Stakeholders, Collective Success, Working with Others, Takes on Tasks Outside Role, Analyzes Problem, Community Engagement, Relate Positively, Connect, Short-Term Impact, Mid-Term Impact, Long-Term Impact, Desired Outcomes, Seeks Perspectives, Sustainable Solution, Invites, Receptive to Feedback, Acts Decisively, Contribute, Translate Feedback, Commits to High Standards, Assistance, Improve Performance, Respectfully Asserts, Shared Goals, Makes Decisions

A.1.6 Mission

Positively Impact, Academics, Life Skills, Functional Literacy, Education Reform, Partners, Network, Shared Goal, Programs, Developing Students, Long-Term