Rationale classification for educational trading platforms

Annie T.T. Ying[1] , **Pablo A. Duboue**[2]
[1]Cisco Vancouver AI Lab*
[2]Textualization Softwared Ltd.
Vancouver, Canada

Abstract

Stock market trading simulation platforms have become popular finance education tools in recent years. To encourage students to think through a trade order, many of such platforms provide a field called "rationale" in the trade order user interface. In this paper, we first present a novel problem called "thoughtful rationale classification" based on two studies: (1) an observational study on the factors affecting a finance professional assessment of a student's trading sophistication and (2) a qualitative study on 2,622 rationales. The two studies together reveal that when a student provides thoughtful rationales, defined as rationales that document external research, specific strategies, or any technical analysis performed, the student is likely to be assessed higher in terms of the trading sophistication. We show that labelling rationales as thoughtful or not is a well defined task and automate it using CNNs. We also compare baseline implementations using simple features and support vector machines over selected keywords.

1 Introduction

Stock market trading simulation platforms have become popular finance education tools in recent years. These platforms provide trading capabilities such as buying/selling (as well as shorting and covering short) on a variety of securities (equities, bond, options, futures), providing a realistic hands-on experience for a student learning to trade.

To encourage students to think through a trade order, many of such platforms provide a field called *rationale* in the trade order user interface. To go one step further, some of such platforms allow a professor to specify the *minimum of rationales* a student has to provide in order to get a participation score as part of their grade. However, simply requiring the minimum number of rationales does not reward a student who provides a thoughtful rationale such as[1]

> XX is expected to show strong performance following Trump's increased defence spending $YY billion

from a student who provides rationales such as "good company," "good stock," or "the stock is rising." A *thoughtful rationale* documents external research, specific strategies, or any technical analysis performed, as opposed to a rationale that lacks any type of thought or analysis.

In this paper, we first present a novel problem called *thoughtful rationale classification* based on two studies: (1) an observational study on the factors that affect a finance professional's assessment on a student's trading sophistication and (2) a qualitative study on 2,622 rationales from a trading simulation platform called EquitySim.[2] The two studies together reveal that when a student provides *thoughtful rationales*, a student is likely to be assessed higher in terms of the trading sophistication. On the other hand, a student who provides rationales that lack any type of thought or analysis are likely to be assessed as lower in trading sophistication. More importantly, there is evidence that introspection and retrospection lead to better learning [Koh *et al.*, 2018]. From this perspective, encouraging a student to write less trivial rationales will help a student learn irrespective of whether there is a correlation between better rationales and better trading performance.

The proposed thoughtful rationale classification is useful in a number of use cases:

- Trading sophistication assessment — Thoughtful rationales were consistently used as a trading sophistication marker from the observational study.
- Teaching good behaviour — Using this classifier can provide immediate feedback to a student to encourage more thought into a trade order.
- Engagement assessment — Counting the number of thoughtful rationales is a much better engagement metric than simply the number of rationales provided.

The contribution of the paper are as follows:

- We show that labelling rationales as thoughtful or not is a well-defined task.

*Work done at EquitySim, Inc. and prior to joining Cisco.

[1]The examples are paraphrases of real rationales.

[2]https://equitysim.com/

Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2019), pages 14-20, Macao, China, August 12, 2019.

- We automate this task using CNNs and compare it against several baseline systems.

This paper is organized as follows: in the next section, we discuss the observational study that motivated the creation of the task. In Section 3, we discuss the annotation and precise definition of the task. Section 4 discuss the different approaches we employed to solve the task, together with the evaluation results. We briefly touch on related work to our task in Section 5. Discussion of future work concludes the paper.

2 Motivating Thoughtful Rationales: Observational Study

In this section, we present an observational study which provided the motivation on the importance of classifying thoughtful rationales. The observational study involved an industry expert reviewing the overall trading sophistication of 11 randomly chosen subjects on a trading simulation platform, based on the past trading actions. Some of these trading actions include the type of security types (equities, bond, option, or future), the amount of the trade, and the rationale text provided along with a trade order.

The evaluator reviewed the past trading actions using the think-aloud protocol [Lewis and Rieman, 1993], i.e., verbalizing the thought process so that the transcript and systematic notes could be taken. This study is exploratory and qualitative in nature; thus, the evaluator was not given a specific set of instruction when they walked through the trading activities, other than "who are good" and "why."

To analyze the factors affecting the evaluation of past trading actions, we used a grounded theory approach [Creswell, 2008], assigning codes to the part of the transcript wherever the evaluator mentioned factors indicative of trading sophistication. The following are the factors.

Portfolio's risk and return (8 subjects)
Portfolio management, from the perspective of modern portfolio theory [Grinold and Kahn, 2000], is about the optimal management of risk and returns. A large positive return is obviously good but luck can be at play. Especially for assessing students' portfolios, return should not be the only factor being considered in evaluating trading sophistication. The evaluator in 8 out of 11 subjects explicitly mentioned risk and return, as well as their trade-offs. Some examples the evaluator explicitly commented on the subjects' portfolios are "good looking chart, steady, good return" (Subject 2); "risky, 2.42 beta which is quite high, but he delivers the return" (Subject 3), and "one big dip in one shot, big return" (Subject 4).

Portfolio diversification (7 subjects)
Diversification is one of the most fundamental strategies in portfolio management for mitigating risk. The evaluator in 7 out of 11 portfolios verbalized aspects of the portfolio related to diversification. For example, the evaluator commented that Subject 3's portfolio is "not good, 81% in one company which is not good, and lots of companies but tiny proportions," whereas for Subject 2, "industry pretty wide but still may not diversify properly but it's OK."

Rationale text (7 subjects)
The evaluator mentioned trade rationales in 7 out of 11 subjects explicitly. For example, on Subject 3, the evaluator lamented "Bad rationales like bullish, good day, rising" whereas on Subject 9, the evaluator commented "better rationales mentioning earning reports."

This is the context that motivates us to look into rationales and see whether we can construct a well-defined task for natural language processing, as described in Section 3.

Complex instruments (6 subjects)
Using complex investment instruments such as options and futures, in the student assessment context, is a sign of sophistication. For example, the evaluator commented on Subject 1 of having "futures: the riskiest thing, but a sign of sophistication" and Subject 2 on having "a bit of options, good."

Trading strategy (2 subjects)
The evaluator in two cases inferred the trading strategy. e.g., "sell when they are down to manage risk, have ranges in mind" (Subject 2) and "large bounds, a waiter, a tolerant trader" (Subject 9).

3 Thoughtful Rationales Definition and Labeling

To elucidate what would be useful to extract from trade rationale text, we conducted a manual analysis on a set of 2,622 rationales within the past trading meta-data from a random sample of 35 users from the EquitySim trading simulation platform. The manual analysis is based on an inductive, qualitative approach.

We found that the thoughtfulness of a rationale is fruitful to focus on for a natural language processing task: four levels of rationale thoughtfulness emerged from the manual analysis of the 2,622 rationales, ranging from a rationale containing little or no thought, to one containing an extensive amount of research or analysis. We chose not to focus on whether a rationale was factually correct—for example, if the rationale mentioned an earnings number, whether the number was actually correct—as this task is a significantly more challenging task for human to annotate, and also significantly more challenging to automate.

The rest of this section provides a precise definition of the task (Section 3.1) and the annotation guide (Section 3.2 and Table 1) constructed for the initial annotation of the 2,622 rationales (866 after de-duplication).[3] These 866 deduplicated rationales are our core evaluation data. Note we do not split the evaluation by users, mixing rationales from different users in the test and training sets. A more rigorous evaluation splitting at the user level is possible, but it was not investigated in the present work.

3.1 Definition

The four levels of rationale thoughtfulness emerged from the manual analysis of 2,622 rationales are as follows:

[3] As the annotation was done in batches over the different students, de-duplication was not considered an issue until later in the process.

Table 1: Types of content among the four levels of thoughtful rationale and examples

Type of content	L1	L2	L3	L4	Examples
Too general	244				"Good trade", "Investment", "To make money"
Non-sensical text	103				"oops", "No laughter... pen drop!!!"
Non-sensical chars	68				"asdfasdfasdf"
Bug	3				"The platform sold my position accidentally. Rebuying"
General price movements		1446			"stocks are rising", "bullish", "short term correction"
General company fundamentals		63			"The company is taking an exciting new direction!" "Expected news"
General economical and & political fundamentals		9			"Economy looking strong. Market is on a rally." "trump stock"
General comment on portfolio strategy (e.g., allocation)		13			"Diversify", "Pharmaceuticals", "Government Bond", "Oil prices rising hopefully"
Uncertain, mistake		29			"this was a gamble", "accidentally hit the trade button"
Technicals			81	18	Level 3: "Day range being tested", "XX Finance has been performing well." Level 4: mentioning mathematical calculations e.g., moving averages, Doji, Bollinger bands, inflection point
Fundamentals - company (e.g., earnings, news, structural events such as M&A, IPO)			127	188	Level 3: "I believe in the stability of this company due to its long history" "Growth with attractive valuation", "Good earnings upside", "Reveal New Product" Level 4: "Monopoly as in-flight internet service porvider, solid fundamentals. Betting to go up after earnings", "XX failed merger depressing price", "XX earning reports: bottom line YY% increase; investor expectations were too high", "XX Accused of Gouging Customers on Prescription Drugs"
Fundamentals - economical & political (e.g., interest rates, seasonal events, elections)			39	29	Level 3: "possible rate increase next week" "Projecting market to fall after election" "cyber monday sell off" Level 4: "XX vote for rate hikes will hurt euro value" "higher XX prices in future months especially with the possible uk EU exit."
Portfolio strategy (e.g., asset allocation, diversification, portfolio actions, limit orders, hedging, target price)			30	37	Level 3: "XX assets because they provide very stable cash flows over the long term", "Diversification across markets", "covering short. taking profits" Level 4: "XX has a virtual monopoly. Stock is undervalued. Target price=XX mid April."
External advice / personal experience			13	4	Level 3: "Buffett growth outlook" Level 4: "XX said: 'hold' to 'buy' + long term uptrend"
Total	418	1560	290	276	

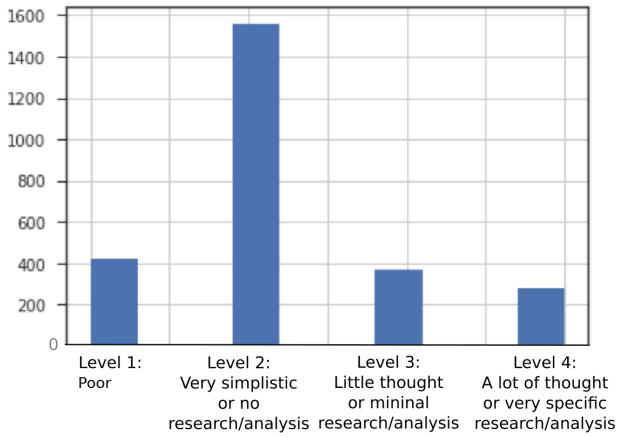

Figure 1: Distribution of rationales by levels of thoughtfulness

1. A rationale contains little or no thought.
2. A rationale contains research or analysis that is simplistic or too general.
3. A rationale contains specific research or analysis.
4. A rationale contains an extensive amount of research or analysis.

Figure 1 shows the distribution of the four levels of thoughtfulness.

3.2 Annotation Guide

Table 1 presents the results emerged from the manual analysis for each of the four thoughtfulness levels. The table and the rest of Section 3.2 serve as the annotation guide for this task. For each thoughtfulness level, Table 1 presents the type of content would be considered at what level (column "Type of Content") and the corresponding number of rationales for each of the four levels (columns "L1", "L2", "L3", and "L4"), as well as example rationales (column "Examples").

A level 1 rationale contains no or little thought.
Most of the rationales in level 1 are either too general or even non-sensical. There were a small portion of rationales that spoke about bugs users experience in the system.

A level 2 rationale contains simplistic or general research or analysis.
Though these rationales are better than the level 1's, the content still does not demonstrate any specific research or analysis. The large majority of these rationales (1,449) are simplistic reasons on the price movements and predictions, such as "stocks are rising." Level 2 rationales can also describe either company or economical/political fundamentals, though without much specific content, such as "The company is taking an exciting new direction!"

A level 3 or 4 rationale contains specific research or analysis, with level 4 having a significant amount.
For level 3, there were 81 rationales with content focusing on the technicals; for level 4, there were 188 rationales including specific mathematical indicators mentioned in the rationale text such as moving average [Bauer and Dahlquist, 1998], Doji [Bauer and Dahlquist, 1998] and Bollinger bands [Bauer and Dahlquist, 1998]. Additionally, there were 127 level 3 rationales and 29 level 4 rationales on the fundamentals of a company, such as structural aspects (e.g., merge and acquisitions), earnings or results, or news on a company. There were 39 level 3 rationales and 29 level 4 on economical fundamentals such as interest rates and other economical events.

We found that 85 rationales contain more than one type of contents: e.g., "Bottom, possibility for some M&A."

3.3 Inter rater agreement

To gauge the complexity of the task and the quality of the instructions, a second person, not familiar with the work, nor with trading concepts annotated 25 rationales in the two class category task (thoughtful vs. non-thoughtful, conflating levels 1-2 and 3-4 above).

Out of a chance agreement of 12.6, both annotators agreed in 19 times, for a Kappa statistic [Carletta, 1996] of 0.516 (a "moderate" agreement). The differences hinged in the understanding of technical terms, for example "covering" expressed an important trading concept that was lost to the second annotator.

This moderate agreement is encouraging and highlights either the need of expert annotators or, if we were to move to the use of crowd workers, the need of machine learning that can profit from noisy labels.

4 System

We experimented with five systems: two sanity baselines (Section 4.1), one SVM baseline (Section 4.2), and two deep learning models (Sections 4.3 and 4.4). The training data for the deep learning models consists of a set of 866 rationales (after de-duplication from 2,622) annotated as described in Section 3. The four thoughtfulness levels are conflated into two labels, *thoughtful* being level 3 or 4, and *not thoughtful* being level 1 or 2. The final dataset we are using has 403 rationales labeled as not thoughtful and 486 as thoughtful.

4.1 Sanity baselines

We approached the problem using existing tools. We started with two sanity baselines. The first sanity baseline uses two features of the rationale: its length and whether it contains a digit. If the length in characters was greater than a fixed threshold (30 characters) or if it contains a digit, then we consider the rationale to be thoughtful. This baseline achieves a precision of 0.812, a recall of 0.725, and a F-measure of 0.766 (Table 2). The second sanity baseline requires both signals to be present (length and digit). This baseline achieves a significantly higher precision of 0.988 but only a recall of 0.182, resulting in a relatively low F-measure of 0.308 overall. As neither of these baselines is trained, the results are over all the 866 rationales.

4.2 SVM

Next, we wonder whether the length threshold could be learned and whether the number of tokens besides the number of character might make for an informative feature. Moreover, using a broad lexicon is traditionally associated with

Table 2: Evaluation results

System	Prec	Rec	F1
Baseline	81.2	72.4	76.5
Strict Baseline	98.9	18.2	30.7
SVM	66.3	81.5	72.7
CNN	82.1	87.1	84.4

textual sophistication in language learning, so we added a third feature indicating the average IDF score for the top 3 highest IDF words in the rationale. As an IDF source, we used 21,000 articles from Thompson Reuters, for a total vocabulary of 133,000 word tokens. Training a Support Vector classifier using a Gaussian kernel on these three features (length in characters, length in tokens, average of the top 3 IDF scores) produced better recall than the baseline system at the expense of lower precision, for a diminished F-measure (third row in the table, evaluated using 10-fold cross-validation).

4.3 CNNs

We then trained a deep learning model for this text classification task using 10-fold cross-validation. The text classification model is trained using a convolutional neural network (CNN) [Goodfellow *et al.*, 2016] with residual connections [Honnibal, 2016] from the spaCy library [Honnibal and Montani, 2017]. The model assigns pre-trained position-sensitive vectors provided by the spaCy library to each word in a rationale. The document tensor (on a rationale) is produced by concatenating max and mean pooling, and a multilayer perceptron is used to predict an output vector, before a logistic activation is applied to each element. The value of each output neuron is the probability of the class (thoughtful or not). The neural network architecture is similar to the hierarchical attention network [Yang *et al.*, 2016] which has two levels of attention mechanisms applied at the word- and sentence-level. The difference with Yang et al's model is in the word embedding strategy (which uses sub-word features and Bloom embeddings [Serrà and Karatzoglou, 2017]) and that CNNs are used instead of BiLSTMs. Because the rationales are all single sentence, the sentence-level attention does not play a role in our case. But the word-level attention is definitely important.

4.4 Transfer learning

We decided to profit from 13,000 unannotated rationales and the strong baseline to experiment with multi-task learning [Caruana, 1997] and transfer learning using universal language models [Howard and Ruder, 2018]. For multi-task learning, we trained a LSTM and two dense layers for three tasks (Figure 2): the thoughtful rationale prediction baseline, predicting the type of operation (buy vs. sell and short) and the type of instrument (stock vs. options and bonds). The hope was the extra tasks will help make the network less tuned to just counting characters and detecting digits, as the network would be if it were trained only the baseline task.

For input, we used spaCy tokenization and word2vec embeddings [Mikolov *et al.*, 2013] over the 21,000 articles form

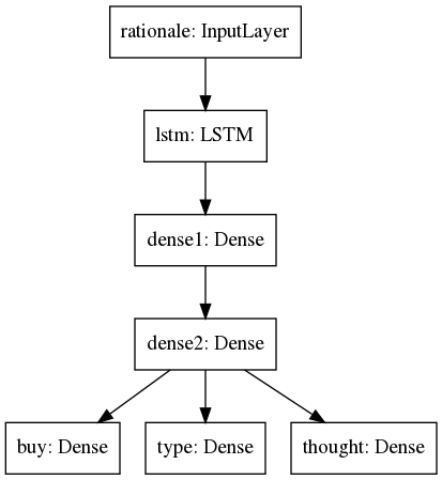

Figure 2: Multi-task network architecture

Thompson Reuters described before. The we used an embedding size of 50 dimensions with continuous-bag-of-words model. The total vocabulary size was 80,263 embeddings. When transforming the rationales into embeddings, due to the more colloquial, informal aspect of the input, there are many mispellings. Of the 321,538 different word types present in the input rationales, 18,058 are missing.

The parameters used are as follows:

- Input sequences where truncated to 40 tokens.
- The LSTM used a 50-unit memory.
- First dense layer used a 50 neurons, with ReLU activation.
- The second dense layer used 25 neurons, with ReLU activation.
- The output layers used a single neuron, with sigmoid activation.
- The network was trained for 12 epochs using a batch size of 256.

We experimented with different layer sizes and drop-out but these were the optimal parameters we obtained for this task.

This network was then transferred using only the thoughtful rationale task over the 866 annotated rationales. The evaluation was done as the average over ten bootstraps using 20% held-out for evaluation. We used a slated triangular learning rate with a base learning rate of 0.004 and a maximum learning rate of 0.01, on top of the Adam optimizer [Howard and Ruder, 2018], plus step-wise thawing of the layers. We used 5 epochs with a batch size of 64. The best accuracy we achieved was 83.3% (average over ten bootstraps), which is below what we obtained using spaCy's CNNs. And that is using an optimal architecture that is risking over-fitting. We continue to investigate other approaches.

4.5 Some qualitative validation

Finally, interviews with professors provided some initial positive qualitative evidence: For example, a professor described

a student with the most number of thoughtful rationales in his class as someone "you can always find him in our building reading the Wall Street Journal."

5 Related Work

The most common application of natural language processing over financial texts focus on sentiment analysis of online discussion around particular stocks [Bollen *et al.*, 2011; Pagolu *et al.*, 2016]. Text classification has also been applied to the finance domain to solve a variety of problems like predicting a firm's credit risk [Byanjankar *et al.*, 2015] and compliance [Fisher *et al.*, 2016], and evaluating loan application [Netzer *et al.*, 2018]. Here, we have taken an approach more closely aligned with natural language processing applications for language education [McNamara *et al.*, 2013; Kyle and Crossley, 2015; Gao *et al.*, 2018]. In that setting, trading rationale sophistication can be considered similarly to existing techniques to automatically evaluate the quality of an essay or a dialogue turn.

The type and quality of the language exhibited by a student has been gaining attention in the field of Educational Data Mining (EDM). Recent work by Crossley et al˙ [Crossley *et al.*, 2018] has tied linguistic features in the language used to talk with a pedagogical agent with their Math Identity. The concept of Math Identity [Nosek *et al.*, 2002] expresses how much of a "math person" the student identifies themselves. We have not taken such a deep analysis route ourselves with the thoughtful rationale identification but we find it an exciting direction. It is possible the type of features used by Crossley and colleagues can be useful for our task, too. It is also enriching to consider a Trading Identity similar to the Math Identity.

Our setting involves a human (the student) making complex decisions (trading) and explaining their actions through rationales. A mirror setting involves a computer system (an agent) interacting with a complex world and explaining its actions [Johnson, 1994; Lacave and Díez, 2002; Stumpf *et al.*, 2009; Lei *et al.*, 2016]. While the connection between the two topics is tenuous and we have not explored in our work, it might be possible to use our available data to automatically produce rationales. They can be used to show the student what a quality rationale for their trade might be and to make them double check their assumptions. Using training data for a generation system poses challenges, though, as poor quality training text will make for generator that produces poor quality output text [Reiter and Sripada, 2002]. In this setting, our work can be seen as pre-filtering quality text that can be then used to build an automatic explanatory system.

6 Conclusion

We have presented here the concept of *thoughtful rationales*, defined as a rationale that documents external research, specific strategies, or any technical analysis performed and showed that our analysis and discussion with professors indicate it is an indicator of trading sophistication. We then proceeded to automate the identification of thoughtful rationales through classification systems and experiments.

Some future improvements on our rationale classification system include features from the trade order, such as the security, trade amount, security type, and time of the order. We also want to explore new transfer learning models such as BERT [Devlin *et al.*, 2018].

A more challenging task would be to determine whether a rationale text displays an additional level of inference. For example, buying a certain stock based on an earnings report is simply reacting to the market where the news is already priced in. The highest level of a rationale would be one that involves an additional level of inference beyond simply reacting to the market, or a rationale that has a unique point of view differs from what the market expects.

Another useful task is to classify whether a rationale contains content on company fundamental, economic fundamental, or a technical piece of analysis. This is ongoing work on evaluating trading sophistication, where rationale thoughtfulness is a strong indicator.

Finally, we might want to expand the system to explore issues of Trading Identity or even a full-fledged trading rationale generation.

Acknowledgements

We would like to thank Justin Ling for support and sharing his finance insights in the observational study. We would also like to thank the reviewers and Denys Gajdamaschko for their useful questions and feedback.

References

[Bauer and Dahlquist, 1998] Richard J Bauer and Julie R Dahlquist. *Technical Markets Indicators: Analysis & Performance*, volume 64. John Wiley & Sons, 1998.

[Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[Byanjankar *et al.*, 2015] Ajay Byanjankar, Markku Heikkilä, and Jozsef Mezei. Predicting credit risk in peer-to-peer lending: A neural network approach. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 719–725. IEEE, 2015.

[Carletta, 1996] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, **22**(2):249–254, 1996.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[Creswell, 2008] John W. Creswell. *Research design: Qualitative, Quantitative, and Mixed Methods Approaches*, chapter Chapter 9: Qualitative Procedures. Thousand Oaks: Sage Publications, 2008.

[Crossley *et al.*, 2018] Scott Crossley, Jaclyn Ocumpaugh, Matthew Labrum, Franklin Bradfield, Mihai Dascalu, and Ryan S Baker. Modeling math identity and math success through sentiment analysis and linguistic features. *International Educational Data Mining Society*, 2018.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Fisher *et al.*, 2016] Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214, 2016.

[Gao *et al.*, 2018] Yanjun Gao, Patricia M Davies, and Rebecca J Passonneau. Automated content analysis: A case study of computer science student summaries. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 264–272, 2018.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, chapter Chapter 9: Convolutional Networks. MIT Press, 2016.

[Grinold and Kahn, 2000] Richard C Grinold and Ronald N Kahn. Active portfolio management. 2000.

[Honnibal and Montani, 2017] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, 2017.

[Honnibal, 2016] Matthew Honnibal. Embed, encode, attend, predict: The new deep learning formula for state-of-the-art nlp models, 2016.

[Howard and Ruder, 2018] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

[Johnson, 1994] W Lewis Johnson. Agents that learn to explain themselves. In *AAAI*, pages 1257–1263, 1994.

[Koh *et al.*, 2018] Aloysius Wei Lun Koh, Sze Chi Lee, and Stephen Wee Hun Lim. The learning benefits of teaching: A retrieval practice hypothesis. *Applied Cognitive Psychology*, 32(3):401–410, 2018.

[Kyle and Crossley, 2015] Kristopher Kyle and Scott A Crossley. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786, 2015.

[Lacave and Díez, 2002] Carmen Lacave and Francisco J Díez. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127, 2002.

[Lei *et al.*, 2016] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, November 2016. Association for Computational Linguistics.

[Lewis and Rieman, 1993] C. Lewis and J. Rieman. *Task-Centered User Interface Design: A Practical Introduction*, chapter Chapter 5: Testing The Design With Users. Self-published, 1993.

[McNamara *et al.*, 2013] Danielle S McNamara, Scott A Crossley, and Rod Roscoe. Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods*, 45(2):499–515, 2013.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Netzer *et al.*, 2018] Oded Netzer, Alain Lemaire, and Michal Herzenstein. When words sweat: Identifying signals for loan default in the text of loan applications. *Columbia Business School Research Paper*, (16-83), 2018.

[Nosek *et al.*, 2002] Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. Math= male, me= female, therefore math$\neq$ me. *Journal of personality and social psychology*, 83(1):44, 2002.

[Pagolu *et al.*, 2016] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE, 2016.

[Reiter and Sripada, 2002] Ehud Reiter and S. Sripada. Should corpora texts be gold standards for NLG? In *Proceedings of Second International Conference on Natural Language Generation INLG-2002*, pages 97–104, Arden House, NY, 2002.

[Serrà and Karatzoglou, 2017] Joan Serrà and Alexandros Karatzoglou. Getting deep recommenders fit: Bloom embeddings for sparse binary input/output networks. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 279–287. ACM, 2017.

[Stumpf *et al.*, 2009] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662, 2009.

[Yang *et al.*, 2016] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.