Structured Paragraph Embeddings of Financial Earnings Calls

Porter Jenkins
Penn State University

ABSTRACT

Financial earnings calls contain rich information about the quarterly performance and future projections of public companies. Such information is highly relevant to developing trading strategies and understanding economic trends. However, due to the unstructured nature of call transcripts important signals can be difficult to extract. In this preliminary work, we propose a novel paragraph embedding method that leverages the structure inherent in the Q&A format of earnings calls. We show that the proposed method improves classification performance over more general methods and provides a useful measure of similarity between paragraphs.

ACM Reference Format:
Porter Jenkins. 2020. Structured Paragraph Embeddings of Financial Earnings Calls. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3366424.3382085

1 INTRODUCTION

Financial earnings calls are a rich source of information detailing the quarterly financial performance of publicly traded companies. In such calls, company leadership (i.e., managers) give an account of their company to the public. Additionally, select Wall Street investors (i.e., analysts) participate in the call to ask follow-up questions to managers. The transcripts of the earnings calls are made available to the public online in raw text form. Information in these earnings call transcripts relates to both historical performance and future projections of a company's financial targets. As such, transcripts represent a rich source of information that is of key interest to developing asset trading strategies and better understanding economic trends [8].

However, adequately representing raw text data remains a core challenge in Natural Language Processing (NLP). Many general NLP techniques exists for representing documents and words. For example, Latent Dirichlet Allocation (LDA) [1] is an unsupervised generative model for discovering topics from raw documents. Additionally, doc2vec is an extension of the word2vec embedding framework for documents. However, both of these approaches fail to exploit the structure of earnings call transcripts. Figure 1 depicts the typical question and answer (Q&A) format of earnings calls. In nearly all cases, a call begins with managers giving prepared remarks. The call then ends with a Q&A between analysts and managers. This structure likely reveals some semantic relationships between documents. The main idea of the current work is to leverage the structure of the quarterly earnings calls to learn better document embeddings.

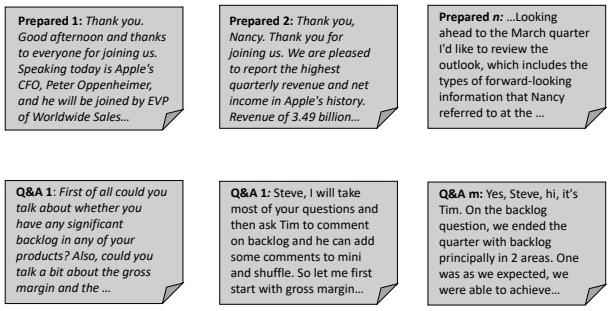

Figure 1: An example earnings call. Each call begins with a set of n prepared remarks, typically each from a different manager. After the prepared remarks, the call progresses into a question & answer (Q&A) format, where analysts ask questions and managers give answers.

However, exploiting this structure for better document embeddings is challenging. Primarily, existing document embedding techniques do not incorporate the Q&A structure for context. LDA makes assumptions about topic distributions through the use of Bayesian priors, while doc2vec assumes the proper context comes from the words within a document. We show that simply applying these methods results in inferior classification performance.

To solve these challenges, we extend the StarSpace framework [9] to incorporate the structured information in quarterly earnings transcripts. Specifically, we make assumptions regarding the semantic relationships between documents by constructing a document-document graph, wherein prepared remarks are adjacent to all other documents in a call, and specific questions and answers between managers and analysts also share an edge. We use the graph to generate positive samples of related documents. Our cost function then tries to maximize similarity between related documents and minimize similarity between negative samples.

In summary, our contributions are as follows:

- We build upon the StarSpace embedding framework by proposing a structured approach for generating related documents and embedding financial earnings calls.
- We show that the proposed embeddings outperform other document baselines in a supervised, speaker classification task. Through a series of case studies, we also show that our method is very effective at identifying document-to-document and word-to-document similarities.

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3382085

2 PRELIMINARIES

In this section we formally define the problem in terms of inputs and outputs.

In our dataset, we observe a set of earnings call transcripts, $C = \{c_i : i <= m\}$. Within each call, c_i, we see two types of documents, prepared remarks, $p_{ij} \in \mathcal{P}_i$, and question-answer dialogue between managers and analysts, $q_{ik} \in Q_i$. Therefore our document set is comprised of prepared remarks and question-answer remarks: $\mathcal{D} = (\bigcup_{i=1}^{m} \mathcal{P}_i) \cup (\bigcup_{i=1}^{m} Q_i)$.

Additionally, because we observe the speaker of each document, $d_i \in \mathcal{D}$, we can construct a document-document graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each document $r_i \in \mathcal{V}$ is a vertex in the graph. The edges are heuristically constructed and denote that two documents, from the same call c_i, share a semantic relation. In our experiments, we model documents has sharing an edge for all prepared and question-answer document pairs, ($\langle p_{ij}, q_{ik} \rangle$) as well as any two subsequent question-answer pairs with different speakers ($\langle q_{ij}, q_{ik} \rangle$). The latter is intended to model the shared semantic relationship between questions and answers. Moreover, note that all of the edges are constructed from documents with the same call, c_i.

3 MODEL

In this section we propose a novel language embedding framework that exploits the structure inherent in the Q&A format of quarterly earnings calls. Our approach is grounded in the generalized StarSpace framework proposed in [9]. We extend this approach by constructing a semantic document-document graph, as depicted in Figure 2.

We assume that each document, d_i, is comprised of a bag of words, $d_i = \{w_1, w_2, ...w_l\}$. Our dictionary, $\mathcal{W}$ is the set of all words observed across all documents. We maintain an embedding matrix $\mathbf{Z} \in \mathbb{R}^{|\mathcal{W}| \times h}$, where each row, $\mathbf{Z}_{(:,i)}$ corresponds to the h-dimensional embedding vector for word, w_i. To obtain the embedding for each document, we can sum over its bag of words, $\sum_{j \in d_i} \mathbf{Z}_{(:,j)}$.

We train our model to compare documents by optimizing the following cost function:

$$\sum_{(a,b) \in E^+} \sum_{b \in E^-} L^{batch}(sim(a,b), sim(a,b_1^-), ..., sim(a,b_k^-)) \quad (1)$$

where E^+ is a set of related documents, and E^- is a set of randomly generated unrelated documents. We use cosine similarity for sim, and margin ranking loss, $L^{batch} = \max(0, \mu - sim(a,b))$. Our key contribution lies in the construction of the sets E^+ and E^-.

For each batch, we randomly sample a set of edges from our edge list, $e_i \in \mathcal{E}$. Each edge provides two documents, d_i, and d_j, that are semantically related. Additionally, for each d_i we construct a set of negative samples, E^-, by sampling documents that are not adjacent to d_i. This sampling strategy allows us to better learn from context and leverage the structure of the Q&A format of earnings calls.

4 EXPERIMENTS

In the following section we perform experiments to validate and explore the information that the proposed embedding method is able to extract from raw financial documents. We first describe the dataset in more detail. Next, we use the features learned from several representation learning techniques to classify speakers as managers or analysts. Finally, we present case studies qualitatively demonstrating how the proposed framework captures document-to-document and word-to-document similarity.

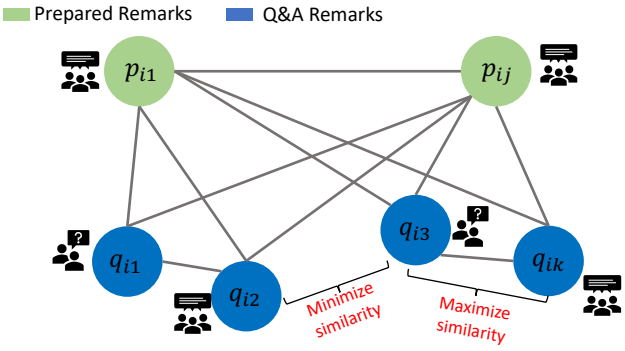

Figure 2: A depiction of the structure used to train document embeddings. For each earnings call, c_i, we assume that the prepared remarks are fully connected to every other document. Additionally, we assume that sequential questions from analysts, and responses from managers are also adjacent. We then train the model to maximize similarity between adjacent documents and minimize similarity from all others. It is important to note that such adjacencies only occur within a call, c_i.

4.1 Dataset Description

Quarterly earnings announcement conference call data for this project are obtained from Refinitiv (formerly known as StreetEvents Data Feed) through Thomson Reuters. We analyze 10,000 call transcripts from 2001-2013, which amounts to 615,603 unique text documents (paragraphs).

4.2 Speaker classification

In the following section we perform a speaker classification experiment. Our desire is to classify each document to either the `manager` or `analyst` class. Intuitively, these two classes of speakers are expected to speak in different ways. Managers present information related to their company's performance, and analysts ask follow-up questions to what managers have revealed. Our hypothesis is that this signal related to speaker category can be detected from raw text alone.

We compare the proposed approach to two NLP techniques presented in the literature:

- **Latent Dirichlet Allocation** is an unsupervised, generative topic model for natural language [1]. Using Bayesian inference algorithms such as variational inference or Markov chain Monte Carlo, LDA seeks to compute the posterior distribution of topics within a document.
- **doc2vec** is an unsupervised text embedding algorithm. Similar to word2vec, doc2vec represents each document as a fixed-length vector, trained to predict its context. In this

Embedding Model	Accuracy	AUC (ROC)	AP
LDA	0.5678	0.4997	0.5670
doc2vec	0.5694	0.5969	0.6885
Proposed	**0.5809**	**0.6314**	**0.7390**

Table 1: Speaker classification results for three embedding methods used as features with an L-1 logistic regression classifier.

case, each document's context is comprised of the words contained in the document [4].

We embed each document as a h–dimensional vector. We let $h = 32$ in all cases. For LDA, we treat the document topic distribution as an embedding vector.

We then train a logistic regression model with L-1 regularization to classify documents to either the `manager` or `analyst` class. We compare all embedding methods in terms of classification accuracy, area under the curve (AUC) - receiver operating characteristics (ROC), and average precision (AP). Results are reported in Table 1.

Overall, we see the proposed embedding approach outperform both LDA and doc2vec across all of the reported evaluation metrics. LDA provides a quite naive representation of documents and results in poor performance. In many cases, the topic distribution is very sparse, with only one or two components receiving any probability mass. Such sparsity likely makes it difficult to for the logistic regression classifier to learn any significant patterns.

Doc2vec provides better performance, especially seeing lift in AUC (ROC) and AP. This is likely due to the ability of the algorithm to incorporate document context. However, both LDA and doc2vec are outperformed by the proposed method. We see increases in accuracy, AUC (ROC) and AP over the other two baselines. The proposed framework is likely able to achieve such convincing results because it builds on the context prediction ideas from doc2vec, but does so in a more stuctured, and semantically meaningful way. Additionally, we present the ROC curves in Figure 3. At nearly every classification threshold, the proposed method gives better results than doc2vec and LDA.

4.3 Embedding Case Studies

We also present two case studies to illuminate the knowledge that the proposed embedding framework has discovered from raw text.

First, we query a set of documents and find the three nearest neighbors in the learned embedding space. Intuitively, documents close in embedding space should share semantic similarities. In Table 2 we show the results for two query documents and their neighbors.

The first is a short document that is a question about marketing and advertising. Each of the top three neighbors shares very similar themes related to marketing and advertising.

The second document is longer, and is composed of a manager's answer related to the seasonal changes of product demand. Similarly, we see that each of the three neighbor documents also discusses product seasonality. This example is also noteworthy because of the length of the documents. Longer documents are typically harder for an algorithm to understand because they likely contain multiple topics or subjects. However, our approach appears to be somewhat robust to document length since all three neighbors share a high degree of semantic similarity.

Finally, in Table 3 we present a word-to-document query of the embedding space. In Table 3 we choose three query words, *profitability*, *demand*, and *seasonality*, and the three most similar documents. In each case, the query word seems to be the main subject of the neighbor document. For instance, the query word *seasonality* returns phrases such as "*...is there any seasonality....?*" and "*No, there's not that much seasonality.*" That being said, each query word returns a relatively short document, which is expected. Because we sum over word embeddings to obtain document embeddings, longer documents will likely be further away from any given word embedding.

Table 3: Word-to-document comparison. We select three query words and display their three nearest documents in embedding space.

Query Word	Neighbor 1	Neighbor 2	Neighbor 3
profitability	In terms of TranSwitch's operating profitability?	We don't talk about profitability by segment.	In terms of your profitability (multiple speakers)
demand	We're seeing very good demand for those this year.	No, the demand is actually been higher than it's ever been.	Okay, so it's a general increase in demand?
seasonality	Okay. Is there any seasonality to the AccuRoute business?	No, there's not much seasonality. A little bit. But –	But you have seasonality every year though.

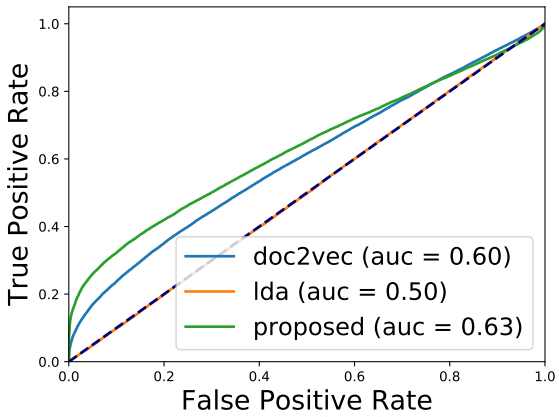

Figure 3: Area under the curve (AUC) of the ROC for a document classifer using different embeddings strategies.

5 RELATED WORK

The current work is related to a number of existing studies. Many works seek to embed words in vector space for improved NLP

Table 2: Document-to-document comparison. We select three query documents and display their three nearest neighbors in embedding space. Only document neighbors are considered.

Query Document	Neighbor 1	Neighbor 2	Neighbor 3
You're referring to the kind of marketing and advertising Google will be doing?	Yes. Just the type of advertising and marketing that you're doing this year versus in the past.	No way. It's a little bit like advertising and marketing.	In terms of increase you've talked before about offline advertising for the floral business. What are you doing in terms of offline advertising and how do you see that growing, going forward?
Yes, Steve, hi, it's Tim. On the backlog question, we ended the quarter with backlog principally in 2 areas. One was as we expected, we were able to achieve a supply demand balance on all models of G5, that's Power Mac and iMac, with the exception of the Power Mac 2.5GHz SKU, and so we did end it with backlog there. Secondly, we ended with backlog on iPod exiting the holiday season. Relative to your question on shuffle and Mac mini, we just announced these yesterday, as you know. We're very pleased with the reception that we received yesterday and that we're getting today, but frankly it's too early to gauge the demand on these. We obviously have an internal forecast on both and had 1 supply up to those forecasts.	Hi, good morning. First question kind of relates to your backlog. Your backlog has been running the past couple of quarters and this quarter at a pretty healthy low teens. Your revenue has been coming in in kind of the high single-digit range. And you specifically have commented this quarter about the strength of at-once, so I'm trying to figure out kind of that mapping or the disconnect between backlog growth and your top-line growth.	I think the key thing from our perspective is that for the second quarter net shipments were up about 4%. And backlog at the end of the quarter was not much changed from the beginning, which is kind of good news and bad news. But like $100,000 change in the backlog from the end of Q1 to the end of Q2. Now, during the quarter, the backlog peaked much higher. 4 million bucks or so higher. So during the quarter we began to eat into the backlog. Since the end of the quarter we've taken about, a little – maybe 2.5 million further reduction in the backlog. In fact, the math that I'm getting to is that shipments were up about 4%, and the backlog was steady during that upshipments period.	Okay. Just had few other quick ones. Last couple of quarters the backlog I think you guys reported at 190 this quarter, versus about 300 million last year at this time. Your backlog has been a pretty good proxy for next quarter or two sales. Just kind of wondering where the disconnect is now because a backlog of 190 given how it's been tracking, say six quarters, would indicate a run rate of sales below your guidance. Just wondering if could you walk me through the disconnect there. Is it just a timing issue?
Sure. The industry normally sees a mid double-digit seasonal decline from the December quarter to the March quarter. In addition, to our natural PC seasonality the iPod business has become a much larger component of our revenue, and with our 90% share of the hard drive-based MP3 player market here in the U.S. it seems reasonable that we could experience something closer to the typical seasonal demand for these consumer products, which I'm told is in the range of 50%.	And then last question I have. In terms of SMTEK's business, how seasonal is it? It seems like in the past it was relatively seasonal – if they just look at what they contributed now, would imply maybe on a flat run rate they might contribute an additional of $10 million to June. But is it a seasonally down quarter for them like it appears, it may be or may be towards the North of that. Are they typically very seasonal or do you see many seasonal patterns in their order book?	Sure, Ross. Well, we don't specifically give guidance, but I think it is appropriate to give you more of an indication. If you look at our seasonality, our historical seasonality, for the three business groups, historically we have seen our Transportation and Standard Products grow on a sequential basis from Q4 to Q1. Historically we have seen our wireless business decline from Q4 to Q1 and our networking, as discussed, is flat.	I just missed that then. Then just last of all on seasonality, last Q3 you had like a 21% sequential growth. Is that a seasonal issue? I guess that was the quarter that Verizon came in, I assume.

tasks. However, these methods are all generally optimized for words and not entire documents [5] [6] [2]. Other work seeks to extend these to the document-level, including LDA [1] and doc2vec [4]. As mentioned above, these methods can be improved by accounting for the Q&A structure in earnings calls. Finally, other studies explore the economic relationships in financial documents using sentiment analysis [7] or word embeddings [3] to predict uncertainty and asset volatility. To the best of our knowledge, no specific embedding method for financial documents exists.

6 CONCLUSIONS AND FUTURE WORK

In this preliminary work, we present a novel framework for learning representations of the quarterly earnings calls of publicly traded companies. These earnings calls are a rich source of information that can be leveraged for a variety of tasks. We demonstrate that the proposed document embeddings outperforms more general methods on a speaker classification task. Additionally, we show that our framework provides a very strong sense of document-to-document and word-to-document similarity.

Future work will explore two aspects of the problem. First, we will extend our semantic document-document graph framework by learning graph attention weights between documents, d_i and d_j. Second, we plan to explore other classification and regression tasks with the learned embeddings.

REFERENCES

[1] Blei, D., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research* (2002).

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Kilian Theil, C. e. a. Word embeddings-based uncertainty detection in financial disclosures. In *Proceedings of the First Workshop on Economics and Natural Language Processing Melbourne, Australia, July 20, 2018* (2018).

[4] Le, Q., and Mikolov, T. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014* (2014).

[5] Mikolov, T. e. a. Efficient estimation of word representations in vector space.

[6] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proc. of NAACL* (2018).

[7] Rekabsaz, N. e. a. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. *arXiv preprint arXiv:1702.01978* (2017).

[8] Wigglesworth, R. Ai decodes trading signals hidden in jargon. online: https://www.ft.com/content/23ae43d4-b3ec-11e7-a398-73d59db9e399, 2017.

[9] Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., and Weston, J. Starspace: Embed all the things! *arXiv preprint arXiv:1709.03856* (2017).