Sentiment Lexicon Induction and Interpretable Multiple-instance Learning in Financial Markets

by

Chengyao Fu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2020

Author's Declaration

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

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

Abstract

Sentiment analysis has been widely used in the domain of finance. There are two most common textual sentiment analysis methods in finance: *dictionary-based approach* and *machine learning approach*. The dictionary-based method is the most convenient and efficient method to extract sentiments from the text, but the words in the dictionary are limited and cannot capture the full scope of a particular domain. Additionally, it is expensive and unsustainable to manually create and maintain domain-specific dictionary using expert opinions. Deep learning models become mainstream methods in sentiment analysis because of their better performance by utilizing extra information on a larger corpus and more complex model structures. However, deep learning models often suffer from the interpretability problem.

This thesis is an attempt to address the issues of both methods. It proposes a machine learning method to do a corpus-based sentiment lexicon induction, which extends the sentiment dictionary that is customized to analyze corporate conference calls. The new extended dictionary is shown to have a better performance than the original dictionary in terms of the three-day returns of the companies in the MSCI universe. It also proposes a highly interpretable attention-based multiple-instance learning model to perform sentiment classification. It also shows that the newly proposed model has comparable accuracy performance to the state-of-the-art sequential models with better interpretability. A keyword ranking is also generated by the model as a by-product. A new sentiment dictionary is also generated by the deep learning method and shows even better performance than both the extended dictionary and the original dictionary.

Acknowledgements

Foremost, I would like to express my deep gratitude to my supervisors, Professor Alan Huang and Professor Yuying Li. Without their guidance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. Their insightful feedback, and immense knowledge helped me in all the time of research and writing of this thesis. I would like to thank you very much for your support and understanding over these past two years.

I would also like to show gratitude to my readers, Professor Peter Forsyth and Professor George Labahn. Thank you for taking your valuable time to read my thesis. Your feedback and comments are very valuable.

In addition, I would like to thank Doctor Grant Wang for his patient help during my internship at AGF Investment Inc.. I could not have completed this thesis without his support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter describes the research motivation and presents the contribution of this thesis. It also details the structure of the paper.

1.1 Research Motivation

Evidence suggests that corporate conference calls[1] contain information that can trigger a significant movement in stock prices. The extraction of sentiment information from conference calls is consequently of interest to investors. The sentiment analysis of different text resources (e.g., news, tweets, reviews, disclosures, and conference calls of companies) has gained significant attention as it can extract signals for examining the effects on the market in different ways, namely correlation with the price movement [37], volume of trades [19], and volatilities. Sentiment analysis is a crucial problem in the category of text classification. It is also a fundamental task in natural language processing (NLP), which involves the investigation of people's opinions or sentiments towards entities such as events, products, institutes, and news. A large number of studies [82, 58] have focused on this subject using a range of techniques—from rule-based methods, including the exploitation of sentiment lexicons, semantic patterns, and grammatical analysis, to the early machine learning methods such as support vector machine (SVM), naïve Bayes, and random forest, which combine the bag-of-words representation of the texts. Recent progress on deep learning

[1] One example of Apple (AAPL) Q3 2020 Earnings Call Transcript can be found in https://www.fool.com/earnings/call-transcripts/2020/07/31/apple-aapl-q3-2020-earnings-call-transcript.aspx

has further increased the performance by utilizing the large dataset and introducing more complicated models such as convolutional neural networks [39], long short-term memory (LSTM) [30], and transformer networks [81].

However, most studies frame the problem as a supervised task that involves a large amount of data with a ground truth associated with each record in the dataset. In finance, data are typically unlabeled, which renders the unfeasibility of most supervised methods. Furthermore, a machine learning model trained on a general dataset usually performs poorly on a domain-specific (e.g., finance) dataset due to the radically different distribution of the samples in the domain-specific dataset. As a result, investors still use a rule-based method by calculating the polarity scores based on the raw counts of the sentiment words from a pre-defined dictionary. The Loughran-McDonald (LM) dictionary, which is an extensively used sentiment dictionary in the financial industry, analyzes more than 50,000 earnings reports during the 1994–2008 period [50]. However, it presents some limitations; for instance, the word list is small, thus hindering the coverage of the full scope of language. Moreover, the LM dictionary is constructed on reports (written English), but the conference text is the transcribed form of speeches (spoken English). Written English and spoken English have different word distributions. Numerous studies attempted to tackle this problem by applying more sophisticated deep learning models such as LSTM [30]. However, the lack of interpretability of deep learning models has slowed down the progress of the application of the model because investors prefer to know the rationale behind the prediction before they make a decision based on the prediction.

This thesis is another attempt to extract information from financial text data, and try to predict the short-term returns by detecting sentiments of financial conference calls. It explores a method that expands the LM dictionary in a way that fully customizes conference call data. It also proposes a highly interpretable deep learning method for sentiment analysis with a by-product of an extensive word list.

1.2 Contributions of this Work

The thesis is an attempt to exploit different sentiment lexicon extraction methods on financial conference call data. It initially reviews the popular methods that are generally adopted in the sentiment analysis, with a focus on the recently proposed deep learning models. The limitations and advantages of each method are subsequently discussed from the perspective of financial conference call data.

Machine learning models such as word embeddings [56] provide rich representations in a latent space by learning from a large corpus. This thesis expands the LM dictionary using a sentiment-aware word embedding, and the new dictionary outperforms the LM dictionary on the correlation test between the sentiment polarity scores and three-day returns. The better performance is consistently observed in all sections of the conference call (presentations, analyst questions, and executive answers).

Multiple-instance learning [38] is widely used in imagine classification for learning the properties of the sub-images that characterize the target scene. This thesis proposes an attention-based multiple-instance learning model [32] to conduct a sentiment analysis at the sentence level, and the model reaches a performance that is comparable to the state-of-art-model on a standard sentiment dataset without sacrificing interpretability. This thesis consequently provides strong evidence for attention as an "explanation" for predictions in contrast with previous research [33]. It is also the first work to apply attention-based multiple-instance learning to text data and use attention scores for extracting sentiment words. A new sentiment dictionary generated by this method displays significant outperformances over the state-of-art financial sentiment dictionary on financial conference calls.

1.3 Structure of the Thesis

This thesis is organized into several chapters. Chapter 2 introduces the background knowledge and presents the previous related studies about sentiment analysis in general and within the financial domain in particular.

Chapter 3 investigates corpus-based methods (Section 2.9.3) for sentiment lexicon induction. The sentiment-aware word2vec (senti-word2vec) proposed in [92] is implemented to extract the sentiment lexicons from conference call data by investigating word relations.

Chapter 4 proposes a highly interpretable attention-based multiple-instance learning (att-MIL) model to perform a sentiment analysis on text data, with the capacity to extract words that are associated with a sentiment label, which is either positive or negative.

Chapter 5 presents a new dictionary created by the senti-word2vec model. It also provides a comparison of the performance between the new dictionary and the LM dictionary [50] on the conference calls of all companies within the MSCI universe from 2008 to 2018. The improved interpretability of the att-MIL model is also discussed in both IMDb [51] and the conference call dataset. A new att-MIL sentiment dictionary is also generated and compared with both the LM dictionary and the extended LM dictionary.

Chapter 6 concludes the thesis and discusses the potential future studies.

Chapter 2

Background and Related Research

The work presented in this thesis entails a sentiment analysis within the natural language process, with a focus on the application to the financial domain. The goal of sentiment analysis in finance is to assign sentimental polarity scores (positive or negative) to financial documents with possible explanations. The succeeding section introduces the dataset used in this work and the requisite background knowledge, followed by a literature review in this area.

2.1 Dataset

Two datasets are investigated in this thesis: financial conference call dataset and the IMDb movie review dataset. The datasets are described below.

2.1.1 Financial Conference Call Dataset

For companies, a financial conference call is a means of relaying information to all the interested parties. A financial conference call is largely conducted immediately after the release of a company's financial results for each quarter.

The financial conference call dataset[1] contains the transcript versions of all the public conference calls from 2008 to 2018. Each conference typically comprises two parts. The

[1] The data are available in this website: https://wrds-www.wharton.upenn.edu/login/?next=/pages/support/manuals-and-overviews/compustat/capital-iq/transcripts/.

first part includes the presentations about the overview of all the significant issues that affected the company's performance in the last quarter. The executives of the company, including the chairman, CFO, and CEO, normally make such presentations. The second part of a conference generally ends with a question-and-answer session in which the analysts from investment banks can raise some questions regarding the company. The work in this thesis aims to extract domain-specific sentiment information from the dataset.

2.1.2 IMDb Movie Review Dataset

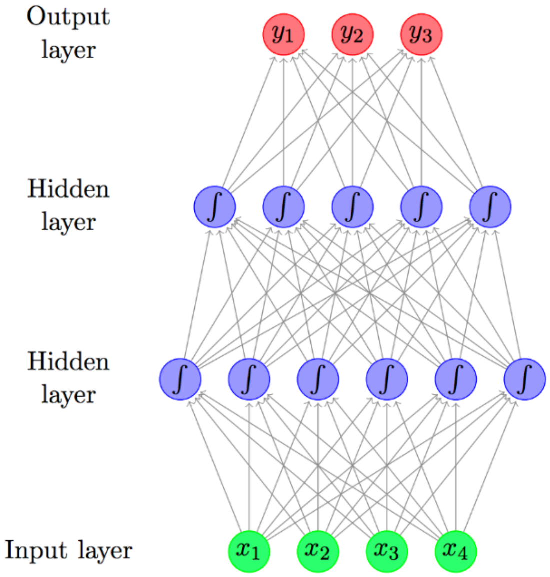
IMDb movie review dataset [51] is designed for the binary sentiment classification, and it substantially contains more data than previous benchmark datasets. It provides a set of 25,000 highly polarized movie reviews for training and 25,000 for testing. The IMDb movie review dataset is a standard dataset for the binary sentiment classification that has been used in numerous studies. The work in this thesis involves the evaluation of the methodology against the standard dataset and the comparison of the results with previous state-of-the-art methods.

2.2 Multilayer Perceptrons

Multilayer perceptrons (MLPs) are often referred to as simple vanilla feed-forward neural networks. They are extensively used models created through regression analysis. Multilayer perceptions are universal function approximators, as indicated by Cybenko's theorem [15]. The network comprises one or more hidden layers, learning a complex hidden representation in the latent space before outputting the results. Figures 2.1 1 illustrates an example of an MLP.

An MLP can approximate any function $g(x)$ by learning the best parameter θ for $f(x;\theta)$. The network consists of many layers chained together, which is represented by a nested function. For example, a network with three fully connected layers can be represented by $f(x) = f_3(f_2(f_1(x)))$, and each layer i can be represented by $f_i(W_i * x_i + b_i)$, where f_i is the nonlinear activation function. Here W_i and b_i are the parameters and bias, and x_i is the input of the layer.

Activation functions are applied to the output of each layer. This aspect provides the model with additional power to describe the arbitrary or non-linear relations between inputs and outputs. Popular activation functions includes sigmoid, tanh, and rectified linear unit (ReLU), with a softmax function usually applied to the output of the last layer



Source: https://machinelearningmastery.com/primer-neural-network-models-natural-language-processing//

Figure 2.1: Example of an MLP structure with two hidden layers

for the multi-label classification problem. The softmax layer is a generalized version of the logistic regression classifier, in which the output of the softmax can be used for representing generalized bernoulli distribution. The softmax layer is expressed in (2.1).

$$P(y = j|x) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}} \tag{2.1}$$

where x is the input to the softmax layer, w_j is the weights associated with the class j, and K is the total number of classes.

An MLP classifier is trained by minimizing a loss function between the estimated distribution $q(x)$ and the ground truth distribution $p(x)$. One popular loss function for classification problem is the cross entropy loss shown in (2.2)

$$H(p, q) = -\sum_{\forall x} p(x) log(q(x)) \tag{2.2}$$

where x is the discrete variable.

2.3 Convolutional Neural Networks

A convolutional neural network (CNN) [80] is a regularized version of the aforementioned MLP: instead of having fully connected layers, a CNN extracts the local features in the data, such as an object in an image or a phrase in a sentence, by applying different shared-weight "filters" as an attempt to capture the spatial (images) and temporal (texts) dependencies. A fully connected layer is impractical to train and prone to overfitting when inputs have large dimensions (e.g., images where each pixel is a dimension). Figure 2.2 illustrates an example of a CNN structure for text data.

A filter (or kernel) is the implementation of a convolution. It maps a sub-region of the image into a single value, and the same filter is applied to all the sub-regions in the image. Multiple filters with different sizes or weights are usually applied to extract different local relations. A pooling layer is constantly applied to the output of the filters to reduce the spatial dimensions, whereby max-pooling or average-pooling is used for outputting the maximum or average number in every sub-region around which the filter convolves.

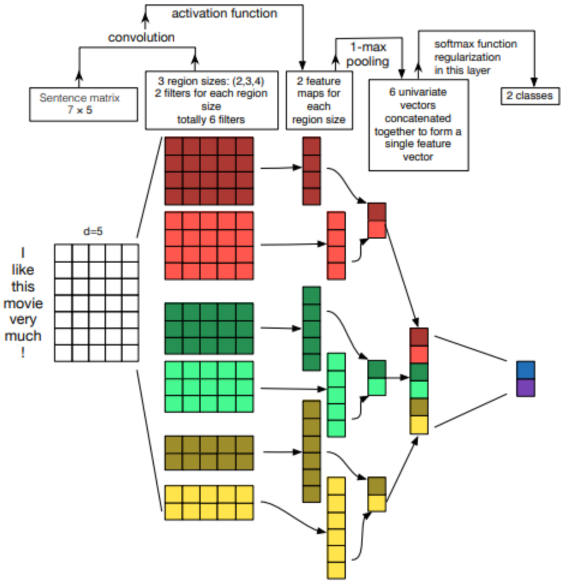
The combination of the convolutional layer and pooling layer allows the network to be translation invariant [94]. In other words, the model can identify the local features regardless of the location of the features in the input data.

A CNN has been shown to achieve the state-of-the-art performance on numerous tasks for images [13, 14, 46, 54, 44]. It has also reached the advanced performance on several NLP tasks such as sentiment analysis [39] and language modeling [34].

2.4 Recurrent Neural Networks

A recurrent neural network (RNN) is a specialized version of a neural network, with its unit forming a directed graph along a temporal sequence. As a consequence, it usually models sequential data such as texts and audios, in which the length of the inputs can vary, and the order of the input features matters. For example, language modeling is a suitable task for the RNN structure, whereby it models the distribution over sentences $p(w_1, ..., w_T)$. The application of the chain rule of conditional probability enables us to decompose the distribution into a sequence of conditional probabilities, as shown in (2.3)

$$p(w_1, ..., w_T) = \prod_{t=1}^{T} p(w_t | w_1, ..., w_{t-1}) \tag{2.3}$$


activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

Source: https://arxiv.org/pdf/1510.03820.pdf

Figure 2.2: Example of a CNN for text classification

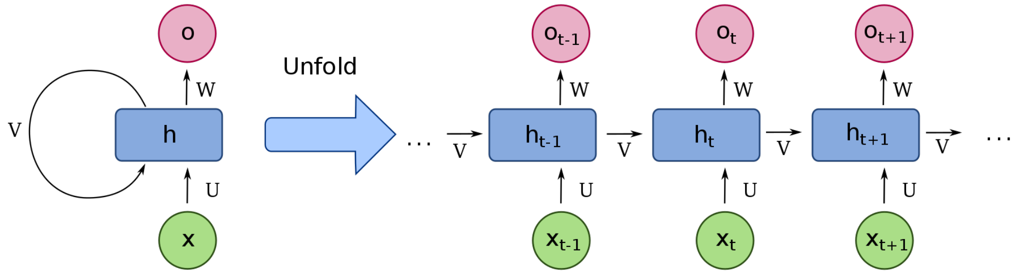
where w_t is the word w at position t in the sentence.

An example of an unrolled RNN structure is presented in Figure 2.3

The RNN model functions as a long-term memory cell by adding information flows between hidden states, thereby allowing the information of the first input to be stored in the hidden state to contribute to the final prediction. This model is in contrast to the memoryless Markov chain model, which assumes that the conditional probability distribution of future states only depends on the present state. The RNN and its variants (LSTM [30] and GRU [11]) have gained considerable attention because of their demonstrated effectiveness in broad practical applications such as machine translation [74], speech recognition [25], and hand-writing recognition [26].

2.4.1 Long Short-term Memory

The aforementioned vanilla RNN is incapable of capturing long-term dependency in practice due to the vanishing gradients problem during the training stage. This drawback makes the vanilla RNN impossible to train on long sequence data. Therefore, some researchers attempted to solve the problem using the gradient norm clipping strategy [59]. Hochreiter



Source: https://commons.wikimedia.org/wiki/File:Recurrent_neural_network_unfold.svg

Figure 2.3: Example of a fold RNN structure (left) to an unfold RNN structure (right)

and Schmidhuber first proposed an LSTM [30] unit to address the issue by introducing the gating mechanisms to adaptively decide the specific information to forget or remain at each time steps.

The structure of an LSTM unit is illustrated in Figure 2.4. Based on the RNN structure, LSTM unitizes three types of gates—input, forget, and output gates—with cell states and hidden states to remember the important information while forgetting the noisy information along the temporal sequence. In the subsequent formulas, we denote $\cdot$ as the inner product, $\sigma(\cdot)$ as the sigmoid function, and $\odot$ as the element-wise multiplication.

A forget gate (2.4) initially performs a linear combination between the learned weights of the forget state W_f and h_{t-1}, x_t, and the sigmoid function σ then maps the results to a number between 0 and 1, where 0 represents "totally forget the information" and 1 denotes "keep all the information." The forget gate is later used for determining the particular information to keep or forget in the previous cell state C_{t-1}

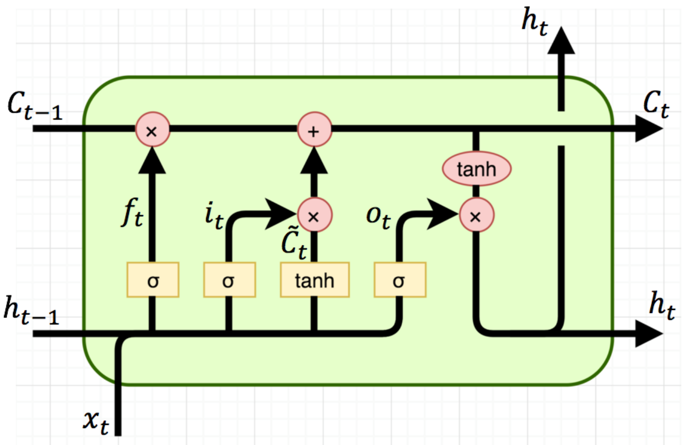
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.4}$$

The input gate (2.5) and output gate (2.6) follow the same pattern as the forget gate. It decides on the specific information to keep in the candidate cell state $\tilde{C}_t$ (2.7) and the final cell state C_t (2.8).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.6}$$

The candidate cell state $\tilde{C}_t$ (2.7) is obtained by $[h_{t-1}, x_t]$ going through a tanh layer. The final cell state C_t (2.8) is composed of the previous cell state C_{t-1} scaled by the forget



Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Figure 2.4: Structure of an LSTM unit

gatef_t plus the current candidate state $\tilde{C}_t$ element-wise multiplied by the input gate i_t

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.7}$$

$$C_t = f \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{2.8}$$

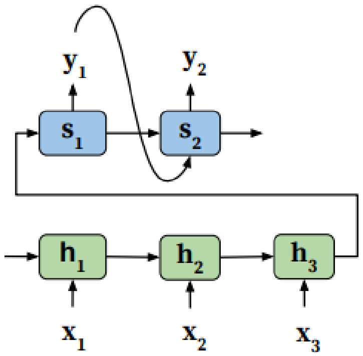
Finally, the output (or current hidden state) h_t (2.9) is computed based on the current cell state C_t, projected by tanh and gated by output gate o_t mentioned above.

$$h_t = o_t \odot tanh(C_t) \tag{2.9}$$

The LSTM is shown to successfully capture long-term dependency. Its auto-regressive nature makes it the dominant model for sequential data such as language modeling [73, 31].

2.5 Attention

The attention mechanism was first introduced in neural machine translation to learn the word alignment between different languages [2]. The intuition behind attention is the notion of relevance. For instance, when one performs a translation task, the only subset of the words in the sentence is relevant to the prediction of the next word. Similarly, in sentiment analysis, only certain words are relevant to the final prediction. Attention also



Source: https://arxiv.org/pdf/1904.02874.pdf

Figure 2.5: Traditional sequence-to-sequence model

relieves the burden for LSTM to learn long dependency, as it utilizes every hidden state of the LSTM instead of only the last one. Furthermore, each hidden state simply needs to capture the short-term dependency around the word [48].

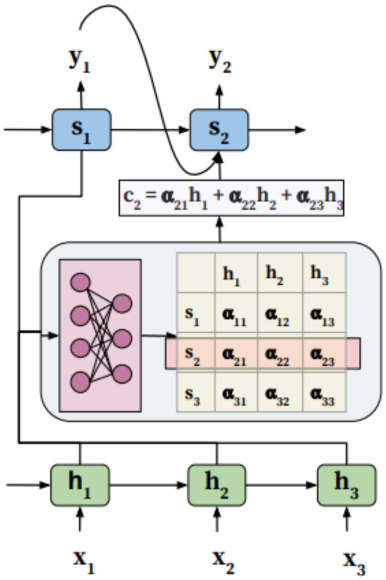
2.5.1 Attention for the Sequence-to-sequence Model

As depicted in Figure 2.5, the sequence-to-sequence model (seq2seq) [74] for machine translation consists of an encoder and a decoder. The encoder (usually an RNN) learns the representation of the sentences (known as the context vector for sentences) from the source language as the last hidden state of the RNN, and the decoder (usually an RNN) takes the last hidden and cell state of the encoder as the initial state based on which it starts to generate the translated version of texts in the target language.

As indicated in previous research [8], the fixed-length context vector is prone to forget the first part of the sentence if the sentence is extremely long. Another study [2] subsequently uses the attention mechanism to define the context vector C as the weighted average of all the hidden states h of the encoder (2.10). The weights, which are referred to as attention scores α_{ij}, are learned using a MLP (2.12) with a softmax layer (2.11) during training (see Figure 2.6).

$$C_i = \sum_j \alpha_{ij} h_j \tag{2.10}$$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{j'} exp(e_{ij'})} \tag{2.11}$$


y_1
y_2
s_1
s_2
$c_2 = \alpha_{21}h_1 + \alpha_{22}h_2 + \alpha_{23}h_3$
h_1 h_2 h_3
s_1 α_{11} α_{12} α_{13}
s_2 α_{21} α_{22} α_{23}
s_3 α_{31} α_{32} α_{33}
h_1
h_2
h_3
x_1
x_2
x_3

Source: https://arxiv.org/pdf/1904.02874.pdf

Figure 2.6: Sequence-to-sequence model with attention

$$e_{ij} = v_a^T tanh(W_a[s_{i-1}; h_j]) \tag{2.12}$$

Here v_a and W_a are weights in the MLP to be learned.

2.5.2 Self-attention

For tasks such as sentiment analysis or any other classification problem where the output of the model is a single value instead of a sequence, attention can be used for modeling the relevance of the hidden states with respect to the final prediction. Yang [89] has proposed a hierarchical attention network for sentiment analysis using self-attention to identify important words and sentences that contribute the most to the prediction.

As attention is capable of learning long-term dependency within the sentence, researchers proposed the transformer model [81], which abandons the RNN structures and only uses attention to model languages. Further works based on transformer and self-attention such as BERT [17] and XLNet [88] are shown to be the state-of-the-art model on most NLP tasks.

2.5.3 Attention as Explanation

Attention is introduced under the hypothesis that the attention scores are highly correlated to the degree of importance or relevance of the input tokens to the output prediction. Nevertheless, in practice, whether attention scores can be used for explaining the prediction of the model is still debatable. Some previous works claim that attention provides interpretability to some extent [2, 86, 48, 89, 8] by showing certain use cases of their models. However, Jain and Wallace [33] evaluated this assumption by conducting extensive experiments on a large number of NLP tasks and concluded that attention could not provide meaningful explanations.

2.6 Text Representation

2.6.1 Bag-of-words Models

The bag-of-words model is a discrete representation for documents in NLP. In the model, each document in the corpus is represented as a row in a sparse matrix, and each word is represented as a column in the matrix; meanwhile, the entry is the term frequency. For example, the entry M(i,j) = 5 signifies that word j appears five times in the document i. This basic method provides a pathway of transformation from raw texts to numerical matrices that the machine can understand. The bag-of-words model ignores the order and grammar of the texts and represents the document as word frequencies.

Numerous studies have indicated the success of the bag-of-words model. Although this model remains the mainstream text representation of most machine learning models, it presents some drawbacks. For example, the bag-of-words model requires careful feature engineering for optimal results, and the sparsity of the matrix brings challenges to model training. Some variants of the bag-of-words model, including word-ngrams [7] and TF-IDF [4], further improve the model by addressing some of the shortcomings.

2.6.2 Distributed Word Representation

Discrete representations such as bag-of-words suffer from limited word vocabulary and ineffectiveness in capturing the semantic relations between words. Distributed word embedding was introduced to map words into a much lower dimensional space in comparison to the size of the vocabulary. Information about the semantic relations between words is gained by exploiting a large corpus. The **distributional hypothesis** [28] provides the theoretical foundation for distributed word embedding. It assumes that the semantic of a word is defined by its context. Words with a similar context should have similar meanings. Different methods of creating distributional word embeddings have been proposed, but all aim to project the words into vector representations in which similar words have similar vectors. The semantic relationship is usually captured by the distance measurement between word vectors such as cosine similarity. Mapping the words into vectors is primarily conducted through two models, namely **count models** and **predict models** (more popularly referred to as neural language model). In an empirical comparison and evaluation of the two models, one study [3] concludes that predict models outperform count models with a large margin.

2.6.2.1 Count Model

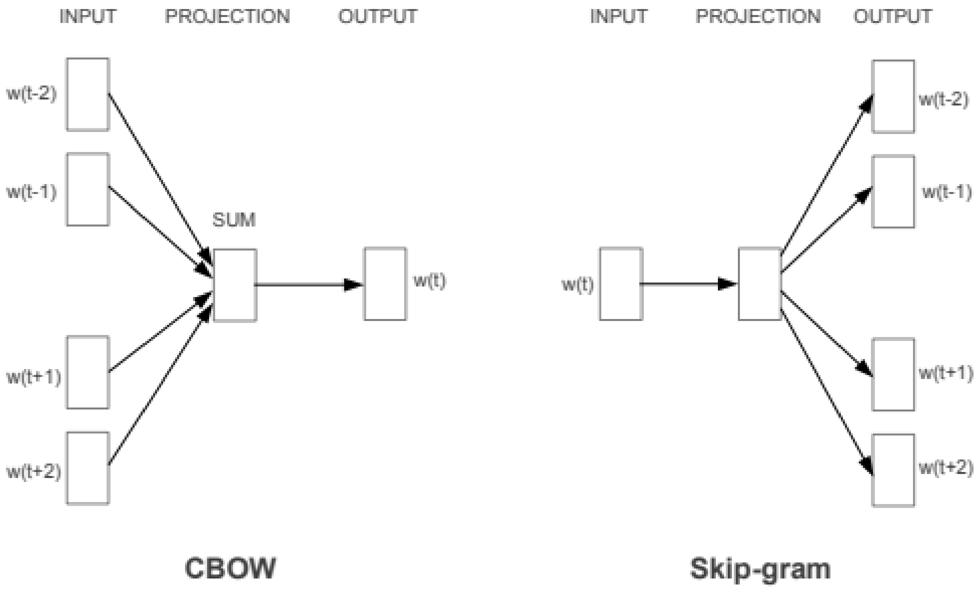
Count models aim to represent words by the raw co-occurrence counts of other words in the context of a large corpus with some weighting schemes such as pointwise mutual information (PMI) [12] (2.13) and log-likelihood ratio. Dimension reduction techniques such as singular value decomposition (SVD) [24] and latent semantic analysis (LSA) [45] are usually applied to compress the vector space. Pointwise mutual information is defined as follows:

$$PMI(W_1 = w_1, W_2 = w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{2.13}$$

where $P(w_1, w_2)$ is the frequency of word w_1 and word w_2 appearing together in the corpus within a window size. $P(w_1)$ and $P(w_2)$ are the frequency of word w_1 and w_2 separately. Moreover, PMI can be interpreted as "observation over expectation," in which a positive value signifies that w_1 and w_2 occur together more than expected under independence assumption, zero represents independence, and a negative value indicates that w_1 is likely to appear only when the w_2 does not appear in the window of a given size.

2.6.2.2 Predict Models (Word2vec)

Predict models frame the vector estimation problem as a self-supervised task, which maximizes the likelihood of the co-occurrence of the center word and context words. Word2vec is the first proposed and most well-known predict model that maps the word into a high dimensional continuous space, in which the representation can also retain the semantic information that words contain [56, 55]. The two forms of the word2vec model are continuous bag-of-words (CBOW) and continuous skip-gram. The CBOW model aims to predict

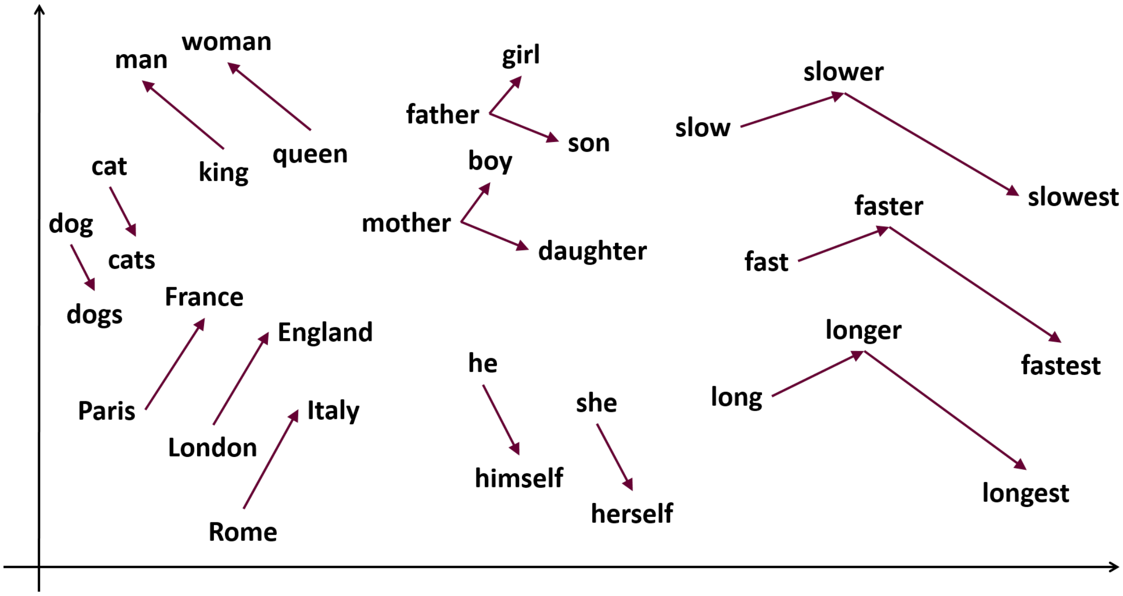


Source: [55]

Figure 2.7: Word2vec embedding model

the current word based on a window of surrounding words, whereas the continuous skip-gram model seeks to predict the surrounding words given the current words. Both models attempt to maximize the likelihood of the co-occurrence of the center word and context words. The structure of the models is illustrated in Figure 2.7. Word2vec model contains only a simple MLP with one hidden layer. The dimension of the hidden layer is a choice of the amount of information that one intends to keep in the compressed latent space. Smaller dimensions indicate a greater loss of information but faster training time, whereas larger dimensions are characterized by more information but are computationally expensive. Empirically, 300 is the most common choice of dimensions. One study [90] suggests that the optimal number for dimensions is around 300 both empirically and theoretically. One variant of word2vec is GloVe [60]. In contrast to word2vec, which only captures the local relation within the window size, GloVe takes advantage of the global context to learn the word relations.

Previous work and empirical evidence have revealed that word2vec is capable of learning rich semantic relationships between words if the model is trained on a large corpus (see Figure 2.8). One famous example is that the analogy of "man to woman as king to queen" can be solved by word2vec embeddings.



Source: https://samyzaf.com/ML/nlp/nlp.html

Figure 2.8: Word2vec latent space

Pre-trained word embeddings such as word2vec and GloVe are currently the standard inputs for most deep learning models for NLP tasks, as they are trained on a large corpus that enables the model to learn the rich and universal semantic relationship between words.

2.7 Model Interpretability

Neural network models, especially deep learning models, are black-box models [27]. The lack of interpretability hinders the progress of deployment of a powerful model for solving real-world problems. Better interpretability not only increases the user's confidence about the model's decision [67] but also illustrates the strength and weakness of the model [22]. In finance, a strong interest has been raised in the investigation of the methods for explaining the black-box models, given the crucial importance of understanding the justification of the prediction when making a key financial decision. However, most of the deep learning models provide great predictions but with little explanation.

Gradient-based saliency maps [71] are the most flexible and convenient means of interpreting the neural network model, as they can be applied to all differentiable models. Saliency maps define the explanation of the model's prediction through the contribution of the model's input, and the gradients of the loss with respect to the inputs are the natural tool for measuring the contributions of inputs. A more sophisticated method called integrated gradients [72] was introduced to explain the model. This method initializes baseline inputs with no information (usually represented by inputs with 0 values). Input contributions are measured by integrating the gradients from baseline inputs into the original inputs.

The attention mechanism is another popular method for interpreting deep learning models (mentioned in Section 2.5.3). However, it is limited to only explaining the model with attention structures.

2.8 Sentiment Analysis in Finance

The most common textual sentiment analysis methods in finance are the *dictionary-based approach* and the *machine learning approach*. These methods are detailed below.

2.8.1 Dictionary-based Approach

The dictionary-based approach calculates the polarity score (often referred to as "tone") for each text document by counting the positive words and negative words in a pre-defined sentiment dictionary [47]. Documents are represented by the "bag-of-words" model (Section 2.6.1), which ignores the linear ordering of the words in a text. The most extensively used sentiment dictionary is Harvard IV-4.[2] Numerous financial studies have utilized the Harvard IV-4 dictionary to derive sentiments from financial texts [76, 42, 16]. However, nearly three-fourths of the words identified as negative by the Harvard IV-4 dictionary are typically not considered as negative in the financial context. For example, words such as "tax" and "liability" are in the negative word list in the Harvard IV-4 dictionary, but they are not negative in the financial context. To tackle the problem, researchers Tim Loughran and Bill McDonald manually created the LM dictionary that is specific to the finance domain from the 10-K filings [50]. The LM dictionary has become the most widely adopted dictionary in both the financial industry and academia. The dictionary-based method is the most convenient and efficient method of extracting sentiments from the text. However, some drawbacks are apparent. First, the words in the dictionary are limited and incapable of capturing the entire scope of a particular domain. Despite the availability of domain-specific dictionaries such as the LM dictionary, the word list is relatively small. Second,

[2] http://www.wjh.harvard.edu/~inquirer/homecat.htm

the manual creation and maintenance of domain-specific dictionary using expert opinions is both expensive and unsustainable.

2.8.2 Machine Learning Approach

Machine learning models usually learn rich information about the distribution of the data from a large training set and make statistical inferences on the test set. In sentiment analysis, researchers typically formulate the problem as a classification problem with the output being either positive or negative; the inputs are the text data represented by bag-of-words (Section 2.6.1). With the introduction of the pre-trained word2vec embeddings (Section 2.6.2.2), deep learning models such as RNN (Section 2.4), and CNN (Section 2.3) become the mainstream methods for sentiment analysis because of their better performance via the utilization of extra information on a larger corpus and more complex model structures. However, as mentioned in Section 2.7, deep learning models often suffer from the interpretability problem.

2.9 Sentiment Lexicon Generation

There are mainly three methods being investigated in the previous work to generate sentiment lexicons. These methods are manually annotated lexicons, thesaurus-based method, and corpus-based method.

2.9.1 Manually Annotated Lexicons

Most of the early works generated sentiment tokens by manually annotating terms with respect to emotions or sentiments. Dictionaries such as Harvard IV-4 and the LM dictionary (Section 2.8.1) are both manually developed from different forms of expert knowledge. In 1999, Bradley and Lang [6] developed a set of Affective Norms for English Words (ANEW), which includes the ratings of 1,034 words. The ratings cover three perspectives according to the theory of emotions [52]. The first and the most relevant aspect of sentiment is the valence of the emotion, from **unhappy** to **happy**. The second aspect describes the level of arousal evoked by the word. The third aspect represents the dominance of the word. Further research [84] extended ANEW to 13,915 English words (E-ANEW) using the same methodology.

2.9.2 Thesaurus-based Method

The thesaurus-based method is also known as a graph-based approach. The general idea is to use the underlying encoded relations between words and a small set of pre-defined seed words through which new words can be induced. The prevalent graph is WordNet defined by Millers *et al.* [57]. Kamps *et al.* [35] proposed measures for determining the semantic orientation of adjectives for three factors by exploiting WordNet. The study [64] generated the sentiment (positive or negative) of a word by re-framing it as a semi-supervised label propagation problem in WordNet. In the graph, each word is a node with a sentiment label, and the edges between them describe the relations.

2.9.3 Corpus-based Method

The corpus-based method relies on the processing of a large corpus and the generation of new sentiment words through co-occurrence statistics. One study [78] combined the co-occurrence measurement and information retrieval to induce the sentiment of new phrases, calculating the PMI (2.13) between the new phrase and word "*good*" or "*bad*" on the web search. Other researchers [20] used the deep learning word embeddings (Section 2.6.2.2) that are enhanced by the crowd-powered filter to classify words into 200 pre-built categories. Kiritchenko *et al.* [41] created the sentiment lexicon for social media by exploiting words that co-occur with sentimental hashtags (e.g., #good or #bad) or emoticons (e.g., :) or :().

2.10 Multiple-instance Learning

Multiple-instance Learning (MIL) [38] is a variant of supervised learning for weakly annotated data [18]. Instead of every instance having an associated label, a single label is assigned to a bag of instances [53]. The main goal of the MIL model is to predict the label distribution at the bag level without knowing the label for each instance in the bag. For example, Keeler [38] used MIL to recognize the handwritten postcode without knowing the position and value of each individual digits. Research interest in simultaneously inferring the individual instance labels and bag labels has recently intensified [49, 95, 87, 43, 32], as detecting the instance-level labels can help with understanding the model prediction. For example, in finance, stock selection has been examined under the framework of MIL [53]. The process involves the selection of the top 100 stocks with the highest returns in the positive bag and bottom stocks with the lowest returns in the negative bag. It attempts to distinguish among stocks that outperform due to fundamental reasons (positive instances in the positive bag), stocks that outperform due to flukes (negative instances in the positive bag), and stocks that underperform (all instance in the negative bag). In the medical field, MIL has been used for detecting the pixels in cancer cells in the histology image [32].

The MIL primary problem (bag-label inference) may be solved in two ways. The first approach projects the instances into a low dimensional space, followed by a bag-level classifier taking the input from the latent space. The second method simply aggregates the results of the instance-level classifier as the prediction of the bag-level label [63, 65]. Although the second method is capable of inferring the instance-level labels, it suffers from poor performance on instance classification [36]. The stability of the instance label is also empirically evaluated [10].

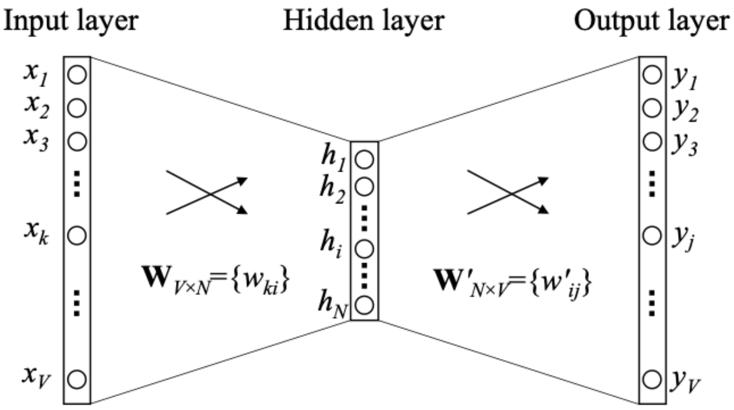
Chapter 3

Corpus-based Methods for Sentiment Lexicon Induction on Financial Conference Calls

In this chapter, we discuss the corpus-based word similarity for domain-specific sentiment lexicon induction. Inspired by [92], we use the sentiment-aware word2vec (senti-word2vec) (Section 3.2) model to extend the current state-of-art sentiment dictionary (Loughran-McDonald dictionary) for financial conference texts. Section 3.1 introduces the general continuous bag-of-words [55] algorithm for learning the semantic relations between words and its limitations on the sentiment relations between words. Section 3.2 describes the process of extending the LM dictionary using senti-word2vec in an unsupervised manner.

3.1 Continuous Bag-of-words for Word Similarity

As mentioned in Section 2.9.3, the corpus-based method for sentiment word induction is based on the relations between words that have been learned from a large corpus. Continuous bag-of-words (Section 2.6.2.2) is a state-of-art model of learning the semantic relations between words.



Source: https://arxiv.org/abs/1411.2738

Figure 3.1: Structure of the continuous bag-of-words model with one word considered in context

3.1.1 Continuous Bag-of-words

Continuous bag-of-words (CBOW) is the simplest model for learning the probability of a word based on its context words. For example,

"The cat jumped over the puddle."

Assume the word "jumped" is the target word. CBOW tries to predict the target word "jumped" based on its context of five words (or neighbouring words) namely, { "The", "cat", "over", "the", "puddle" }. The size of the context (i.e., the number of the neighbouring words) is a hyper-parameter we can choose based on the data. To illustrate the model, we start with a simple CBOW model with only one word considered per context.

3.1.1.1 CBOW with One-word Context

Figure 3.1 shows the network model with only one word considered per context. This means the model will predict one target word given one neighbouring word. In the aforementioned example, the CBOW model tries to predict the target word "jumped" based on one neighbouring word { "cat" }. In our setting, assume a vocabulary of size V, and the hidden layer size is N. The adjacent layers are fully connected. Let the input of the model be the one-hot encoding vector of a word, which means only one unit out of V units,

$(x_1, x_2, ..., x_V)$, is 1 for a given neighbouring word, and all others are 0. In this case, we assume x_k is 1 to represent the neighbouring word, w_I with index k.

The weights between the input layer and the hidden layer can be represented by a matrix $W \in \mathrm{I\!R}^{\mathrm{VxN}}$. Each row of W is the N-dimension vector representation v_w of the associated word of the input layer. Formally, row i of W is v_w^T. Given a context with one word, w_I with an index of k in the vocabulary, and $x = (x_1, x_2, ..., x_V)$ is the one-hot encoding of the word w_I with only unit x_k is 1, we have:

$$h = W^T x := v_{w_I}, \tag{3.1}$$

which practically copy the k-th row of W to h. v_{w_I} is a vector representation of the input word w_I to be learnt.

The weights between hidden layer and the output layer is a matrix $W' \in \mathrm{I\!R}^{\mathrm{NxV}}$. After a linear combination of this matrix W' and hidden layer h, we can generate a score u_j for every word in the vocabulary,

$$u_j = {v'}_{w_j}^T \cdot h \tag{3.2}$$

where v'_{w_j} is the j-th column of the matrix W'. A softmax function then is used to obtain the posterior distribution over all words in the vocabulary.

$$p(w_j|w_I) = y_j = \frac{exp(u_j)}{\sum_{j'=1}^{V} exp(u_{j'})} \tag{3.3}$$

where y_j is the output of the j-th unit in the output layer. Combining (3.1), (3.2), and (3.3), we obtain:

$$p(w_j|w_I) = y_j = \frac{exp({v'}_{w_j}^T \cdot v_{w_I})}{\sum_{j'=1}^{V} exp({v'}_{w_{j'}}^T \cdot v_{w_I})} \tag{3.4}$$

Note that v_w and v'_w are two different representations coming from rows of W and columns of W', respectively. In the subsequent analysis, we denote v_w as the "**input vector**" of the word w, and v'_w as the "**output vector**" of the word w.

The training objective as noted in [68] is to maximize (3.4), the conditional probability of observing the actual output word w_O given the input context word w_I. We assume that the index of the output word w_O in the output layer is j^*.

$$\max p(w_O|w_I) = \max y_{j^*} \tag{3.5}$$

For optimization purpose, we minimize the negative log probability:

$$\min -log(p(w_O|w_I)) = \min -log(y_{j^*}) \tag{3.6}$$

Combining (3.4), and (3.6), we have:

$$\begin{aligned}\min -log(p(w_O|w_I)) &= \min -log(y_{j^*}) &(3.7)\\ &= -v'^T_{w_O} \cdot v_{w_I} + log \sum_{j'=1}^{V} exp(v'^T_{w_{j'}} \cdot v_{w_I}) := E &(3.8)\end{aligned}$$

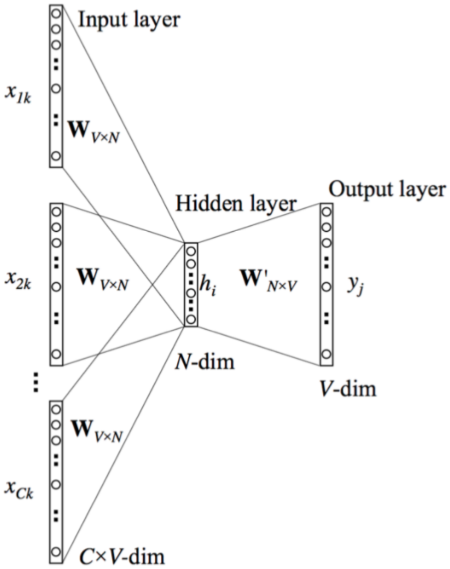
where we try to maximize the log probability $logp(w_O|w_I)$ over the "input vector", v_{w_I}, of input word w_I, the "output vector", v'_{w_O}, of the target word w_O, and the "output vectors", $v'_{w_{j'}}$, of all the words $w_{j'}$ in the vocabulary. $E = -logp(w_O|w_I)$ is the loss function to minimize.

3.1.1.2 Generalized CBOW with Multi-word Context

Figure 3.2 presents the structure of the CBOW model with a multi-word context input, which means the model will predict one target word given multiple neighbouring words. In the aforementioned example, the CBOW model tries to predict the target word "jumped" based on the neighbouring words, { "The","cat", ”over", "the", "puddle" }. In this case, the CBOW model takes the average of the one-hot encodings of the neighbouring words as the input, and use the inner product of the weight matrix W and the average vector as the hidden vector h (3.9).

$$\begin{aligned}h &= \frac{1}{C} W^T (x_{1k} + x_{2k} + ... + x_{Ck}) &(3.9)\\ &= \frac{1}{C}(v_{w_1} + v_{w_2} + ... + v_{w_C}) &(3.10)\end{aligned}$$

where C is the number of words in the context, $x_{1k,...,}x_{Ck}$ are the one-hot encodings of words $w_{I,1,...,}w_{I,C}$ in the context ({ "The","cat", ”over", "the", "puddle" } in our example), and v_w is the "input vector" of a word w. The loss function is



Source: http://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes01-wordvecs1.pdf

Figure 3.2: Structure of the continuous bag-of-words model with C words considered in context

$$E = -logp(w_O|w_{I,1}, ..., w_{I,C}) \tag{3.11}$$

$$= -u_{j^*} + log \sum_{j'=1}^{V} exp(u_{j'}) \tag{3.12}$$

$$= -v'^{T}_{w_O} \cdot h + log \sum_{j'=1}^{V} exp(v'^{T}_{w_j} \cdot h) \tag{3.13}$$

which is the same as the E in (3.8) except that h is defined in (3.9) instead of (3.1).

3.1.2 Negative Sampling

In the CBOW model, for each training instance, we have to iterate through every word w_j in the vocabulary: computing the net score u_j; probability y_j; prediction error e_j; and use the error to update all "output vectors" v'_j in the weight matrix $W' \in \mathbb{R}^{\text{NxV}}$ (i.e., "output vectors" of all words in the vocabulary). This aspect renders the infeasibility of training the model in a large corpus with billions of words in the vocabulary. The intuitive solution is to limit the number of the "output vectors" needed v'_j to be updated per training instance. One paper [56] proposed the use of negative sampling to reduce the computation cost of the optimization.

The idea of the negative sampling is fairly straight-forward. In order to tackle the complexity of updating too many "output vectors" per training instance, we only update a subset of them. The "output vectors" of output word (i.e., the ground truth, the positive sample) should be kept in our sample and get updated, and we need to sample the "output vectors" of few words (other than the ground truth) as negative samples. The sample process needs a probability distribution, and the distribution can be defined arbitrarily. We call the distribution the noise distribution and denote it as $P_n(w)$.

In the case of CBOW, [55] proposes a simplified training objective of negative sampling (3.14) and argues that this simplified objective is capable of producing high-quality word embeddings while reducing optimization cost.

$$E = -log\sigma(v'^{T}_{w_O} \cdot h) + \sum_{w \in \mathbb{W}_{neg}} log\sigma(-v'^{T}_{w} \cdot h) \tag{3.14}$$

where $\sigma(\cdot)$ is the sigmoid function, w_O is the output word (i.e., the positive sample), and v'_{w_O} is the "output vector" of the output word, h is the output of the hidden layer where

$h = \frac{1}{C} \sum_{c=1}^{C} v_{w_c}$, $\mathbb{W}_{neg} = \{w_j | j = 1, ..., K\}$ is the set of negative samples that are sampled based on $Pn(w)$ where K is the size of the negative samples. The simplified objective (3.14), compared to the original one (3.13), reduces the computation per training instance from $O(V)$ (size of the vocabulary) to $O(K)$ (size of the negative samples). [56] suggests that $K = 5$ is empirically good enough for training this objective. [23] provides a theoretical analysis as to why we use this simplified objective function.

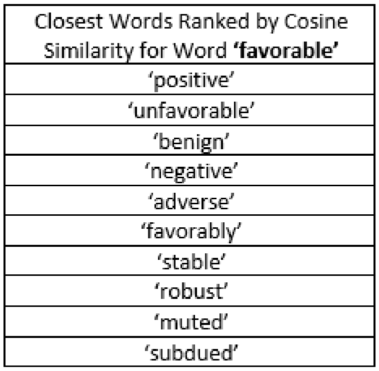
3.1.3 CBOW as Word2vec Embeddings

The word2vec embeddings are an umbrella term for models that transfer a word index into a word vector (i.e., word embedding). Formally, given a word index i, the word2vec embeddings serve as a look-up table to produce the corresponding word vector v_i. The learnt weight matrix $W' \in \mathbb{R}^{\mathrm{NxV}}$ of the CBOW model in Figure 3.2 is normally used in the word2vec embeddings. Each column in W' is the "output vector" v'_{w_i} for the word w_i with the index i. N is the pre-defined size of word vectors and it is also the size of the hidden layer in CBOW model. V is the vocabulary size. Thus, $W' \in \mathbb{R}^{\mathrm{NxV}}$ contains all the word vectors in the vocabulary.

The original paper of CBOW [55] shows that the CBOW model can successfully learn high-quality word vectors from large corpus with billions of words. The resulting vector representations can preserve similarity between words, meaning words with similar meaning tend to be close to each other based on the cosine distance between word vectors.

3.1.4 Word2vec on the Conference Call Dataset

Despite the existence of some pre-trained word embeddings such CBOW [55] and GloVe [60], most are trained on a general corpus such as Wikipedia. To learn rich semantic relations between words in the financial context, we train the CBOW model on financial conference call data from 2008 to 2018 and we use the weight matrix W' as the word vectors in word2vec embeddings. We do not conduct any text prepossessing, except for the tokenization of sentences into words because we intend to retain the semantic information as much as possible.



Closest Words Ranked by Cosine Similarity for Word **'favorable'**
'positive'
'unfavorable'
'benign'
'negative'
'adverse'
'favorably'
'stable'
'robust'
'muted'
'subdued'

Figure 3.3: Words that are similar to the word "favorite" ranked by cosine similarity between word2vec embeddings.

3.2 Senti-word2vec

3.2.1 Limitation of the General Word2vec Embeddings

The general word2vec embeddings (i.e., the output weight matrix W' of the CBOW model) are trained under the distributional hypothesis [28], in which the semantic meaning of a word is defined by its neighbor words (i.e., context words). Under this assumption, the learned word vectors ignore the sentiment perspective of the word. Prior research [92] has reported that words with similar vector representations may have opposite sentiment polarity scores. For example, the words "good" and "bad" are indeed likely to have similar neighbor words; thus, they are considered similar by CBOW even though they have opposite sentiments. We examine the word2vec embeddings trained on a conference call dataset (Section 2.1.1) using cosine similarity: We calculate the cosine similarity scores (i.e., inner product) between the word vector of word "favorable" $v'_{w_{favorable}}$, and all others word vectors in the matrix W' corresponding all other words in the vocabulary. We rank all other words in the vocabulary descendingly based on their associated cosine similarity scores. A larger score means the word vector generated by the model share more similarity with the vector of word "favorable". Figure 3.3 displays top ten words in the rank based on cosine similarity scores to the word "favorable". The resulting rank reveals that the word "unfavorite" ranked second as the most similar word to the word "favorite", which provides evidence of the limitation of the word2vec embeddings.

To overcome this problem, some researchers [75] attempted to change the training objective of the CBOW model to learning the semantic and sentimental meaning of words by supervised learning based on the sentiment polarity labels. This thesis adopts another strategy proposed by [92]. Instead of changing the training objective and re-training the model on the labeled dataset, we use the same objective of CBOW to learn the semantic meaning of the words in a self-supervised fashion and subsequently add sentiment information to each word as a post-processing step based on a pre-defined sentiment dictionary.

3.2.2 E-ANEW Sentiment Dictionary

The E-ANEW dictionary is a manually annotated sentiment dictionary by 1,827 participants in the study [84]. It contains 13,915 sentiment words with three different scores associated with each word. The three scores are in line with the theory of emotion [52] to measure the emotion of a word. The first and the most important score is the valence of the emotion triggered by the word, ranging from *unhappy* to *happy*. The second score describes the level of arousal evoked by the word (from *calm* to *excited*), and the third aspect represents the dominance (power) of the word of which the word denotes something weak or strong. In this thesis, we use the valence score to measure the word sentiment, which ranges from 1(extremely unhappy) to 9(extremely happy), and 5 is neutral.

3.2.3 Sentiment Lexicon Induction

The LM dictionary is manually generated based on the annually financial reports so most of the words in the dictionary are written English. Thus, the words in the dictionary are limited and cannot fully capture the sentiment of spoken English in transcripted financial conference calls. Therefore, our goal is to extend the LM dictionary to capture written and spoken English words.

We use the word2vec embeddings trained on the financial conference call data to identify the top k most similar words to the target words. In our example, target words are the words in the LM dictionary because our objective is to extend the LM sentiment word lists. We use the E-ANEW dictionary to re-rank the top k most similar words based on the sentiment distance defined as follows:

$$sentidist(w_1, w_2) = |v_1 - v_2| \tag{3.15}$$

where w_1, w_2 are the words and v_1, v_2 are the individual valence scores for w_1 and w_2 separately. We list the detailed steps and describe word induction:

1. Tokenize the conference call text into word-level tokens without any text prepossessing steps. For example, the sentence “The cat jumped over the puddle.” is considered as a list of tokens: [“The”, “cat”, “jumped”, “over”, “the”, “puddle”, “.”].

2. Train a CBOW neural model (Figure 3.2) on the financial conference call data with 300 units in the hidden layer ($N = 300$), 10 context words for input ($C = 10$), and 19,426 distinct words in the vocabulary ($V = 19426$). Use a negative sampling (Section 3.1.2) to reduce the training complexity.

3. Construct the word2vec embeddings using the weight matrix $W' \in \mathbb{R}^{\mathrm{NxV}}$ in the CBOW model as mentioned in Section 3.1.3. In our case $N = 300$ and $V = 19426$.

4. Identify the top 15 most similar words to every target word (i.e., words in the original LM dictionary) based on cosine similarity between word embeddings of the target words and the rest of the words.

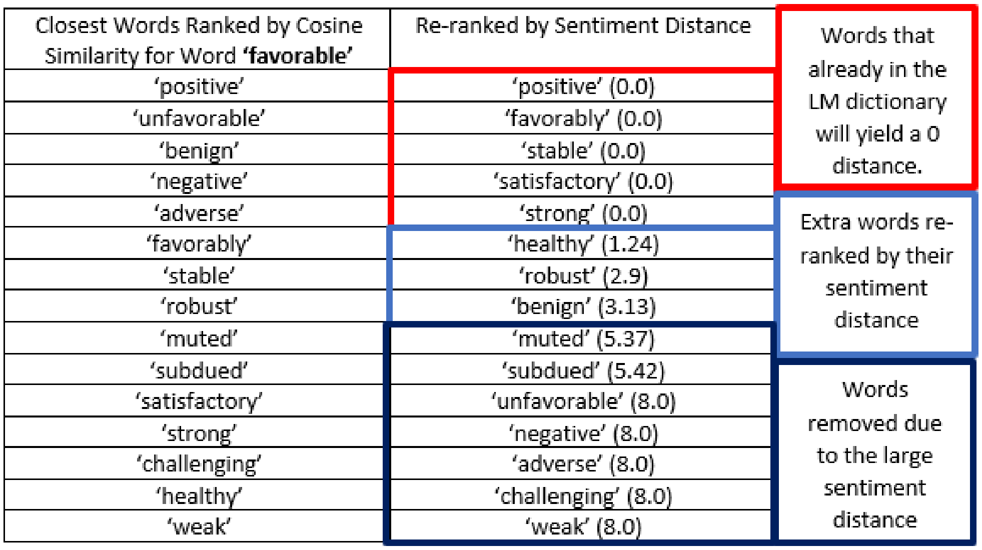
5. Reset the valence scores of the target words to an extreme number (1 for negative words, 9 for the positive words). By undertaking this step, we assume that the targets are either extremely positive or extremely negative because we believe that the LM dictionary is the ground truth in terms of financial texts.

6. Calculate the sentiment distance between the top 15 most similar words and the target words based on (3.15) and re-rank the word list according to the sentiment distances (3.15).

7. Remove all the words that have a sentiment distance larger than the threshold of 4 (the midpoint of the distance), as indicated in Figure 3.4

3.2.4 Threshold Tuning

We use 4 as the baseline threshold—given the scores ranging from 1 to 9, the maximum distance is 8 and the midpoint of the distance is 4. Words with a distance larger than 4 have opposite sentiments. However, based on the baseline, we can further trim the word list by decreasing the threshold values. The extended positive (or negative) word list consists of the original LM positive (or negative) words plus the additional words identified. The performance of the extended word list is measured by the information coefficient (IC), computed as Kendall’s tau [61], between the three-day excess returns (benchmarked to the



Closest Words Ranked by Cosine Similarity for Word **'favorable'**	Re-ranked by Sentiment Distance	
'positive'	'positive' (0.0)	Words that already in the LM dictionary will yield a 0 distance.
'unfavorable'	'favorably' (0.0)	
'benign'	'stable' (0.0)	
'negative'	'satisfactory' (0.0)	
'adverse'	'strong' (0.0)	
'favorably'	'healthy' (1.24)	Extra words re-ranked by their sentiment distance
'stable'	'robust' (2.9)	
'robust'	'benign' (3.13)	
'muted'	'muted' (5.37)	Words removed due to the large sentiment distance
'subdued'	'subdued' (5.42)	
'satisfactory'	'unfavorable' (8.0)	
'strong'	'negative' (8.0)	
'challenging'	'adverse' (8.0)	
'healthy'	'challenging' (8.0)	
'weak'	'weak' (8.0)	

Figure 3.4: Words that have a large *stnidist* (3.15) will be removed. Numbers in the parenthesis are the *sentidist*

MSCI US Index[1]) and the quarterly sentiment scores of all the companies across every sector within the MSCI universe from 2008 to 2018. The three-day returns are calculated based on the split-adjusted closing price covering the period from one business day before the earning call date and another business day afterwards. The optimal threshold with the best performance is identified by exhausting all the possible values from 0 to 4 with an increment of 0.1 and finalizing the positive and negative extended word lists with each threshold. The full evaluation metrics and results are presented in Chapter 5.

[1] The MSCI USA Index is designed to measure the performance of the large and mid-cap segments of the US market. With 636 constituents, the index covers approximately 85% of the free float-adjusted market capitalization in the US. The description can be found in: https://www.msci.com/documents/10199/471d55eb-ca0b-43c8-882c-ee161de1c422

Chapter 4

Interpretable Sentiment Analysis with Attention-based Multiple-instance Learning

4.1 Multiple-instance Learning

4.1.1 Motivation

Multiple-instance learning has been widely applied to various fields (e.g., stock selection [53], and computer vision [32]), but rarely it has been applied to the field of natural language processing. The possible reason is that most of the mainstream models are sequential (i.e., they take the order of the words into considerations), but the nature of the MIL model ignores the permutation of words. This assumption of independent words is usually believed to have worse performances. However, with the advancement of deep learning, the implementations of MIL models also allow the models to capture more complex relations between words even we assume they are independent. It is also easier to interpret MIL models than sequential models. Thus, we try to apply the MIL models to solve sentiment analysis problem.

4.1.2 Problem Formulation

In the case of classical binary supervised learning, the model seeks to predict a label y based on a given instance x. However, in the setting of a multiple-instance learning, a bag of instances $X = \{x_1, x_2, ..., x_K\}$ is given, in which we assume that every instance in the bag is independent and that the instances have no ordering between them (Condition 1). K can be different for diverse bags. Every bag X is also associated with a label Y. We also assume that every instance x_k in the bag has its own label y_k where $y_k \in \{0, 1\}$, for $k = 1, ..., K$ (Condition 2). The individual label y_k remains unknown during training, and we re-define the bag level label Y as follows:

$$Y = \begin{cases} 0, & \text{iff } \sum_k y_k = 0 \\ 1, & \text{otherwise} \end{cases} \tag{4.1}$$

Conditions 1 and 2 guarantee that the MIL model is **permutation-invariant**. Furthermore, the preceding equation can be written as

$$Y = \max_k \{y_k\} \tag{4.2}$$

However, directly learning the aforementioned objective (4.2) results in two major problems. First, gradient-based optimization methods encounter vanishing gradients with the max operator. Second, it is only feasible when an instance-level classifier is present.

To overcome both learning problems, we adopt the MIL problem by optimizing the log-likelihood function, $\theta(X)$, in which the bag label is distributed in the Bernoulli distribution with parameters. The bag probability $\theta(X)$ is the probability of the bag label $Y = 1$ given the bag X. Thus, bag-level labels are directly inferred by the model instead of aggregating the instance-level labels by the *max* operator.

4.1.3 Multiple-instance Learning Approaches

For MIL, the probability $\theta(X)$ must be **permutation-invariant** because every instance in the bag is independent and lacking in order. The MIL problem (Section 4.1.2) can be categorized into a special form of the fundamental theorem of symmetric functions with monomials given by the following theorems [93]:

Theorem 1: A scoring function for a set of instances X, $S(X) \in \mathbb{R}$, is a symmetric function (i.e., permutation-invariant to the elements in X), if and only if it can be decomposed in the following form:

$$S(X) = g(\sum_{x \in X} f(x)) \tag{4.3}$$

where $f(\cdot)$ and $g(\cdot)$ are suitable transformation functions.

(4.3) provides a general mechanism to model the bag-level probability $S(X)$.

Theorem 2: For any $\epsilon > 0$, a Hausdorff continuous function[1] $S(X) \in \mathbb{R}$ can be arbitrarily approximated by a function in the form of $g(\max_{x \in X} f(x))$, where f and g are continuous functions, that is:

$$|S(X) - g(\max_{x \in X} f(x))| < \epsilon \tag{4.4}$$

Both of the theorems frame the MIL into three steps:

1. A transformation of instances using function $f(\cdot)$;

2. A symmetric (permutation-invariant) function $\sigma(\cdot)$ is applied to aggregate the transformed instances. (e.g. $\sum$ in **Theorem 1** and max in **Theorem 2**); and

3. A transformation of aggregated instances using $g(\cdot)$.

In the case of MIL, $f(\cdot)$ and $g(\cdot)$ are called transformation functions. The permutation-invariant function $\sigma(\cdot)$ is referred to as MIL pooling. Different choices of the functions and pooling define the various strategies for implementing MIL. The two main strategies are as follows:

1. *Instance-based approach*: The transformation $f(\cdot)$ is an instance-based classifier that obtains a score (1 dimension) for each instance in the bag X, and the $\theta(X)$ is obtained by the MIL pooling over the instance scores. In this case, $g(\cdot)$ is simply an identity function.

2. *Embedding-based approach*: $f(\cdot)$ maps every instance into an embedding in a low dimensional space (usually more than 1 dimension). A bag-level representation is obtained by the MIL-pooling function aggregating on every unit of instance-level embeddings, which is independent of the order of the instances in the bags. $g(\cdot)$ is a bag-level classifier for predicting the label based on the bag-level embedding (i.e.,aggregated instances-level embeddings).

Previous research indicated that the latter approach performs better in terms of the bag-level classification because the performance of the former approach is highly dependent on the quality of the instance-based classifier, whereas the latter integrates the instance embeddings to reduce the bias of insufficiently trained instance-based classifier. However,

[1] Defined by [62], a Hausdorff continuous function $S(X) \in \mathbb{R}$ is a continuous set function w.r.t. Hausdorff distance $d_H(\cdot, \cdot)$. A proof of Theorem 2 can be found in the supplementary material of [62].

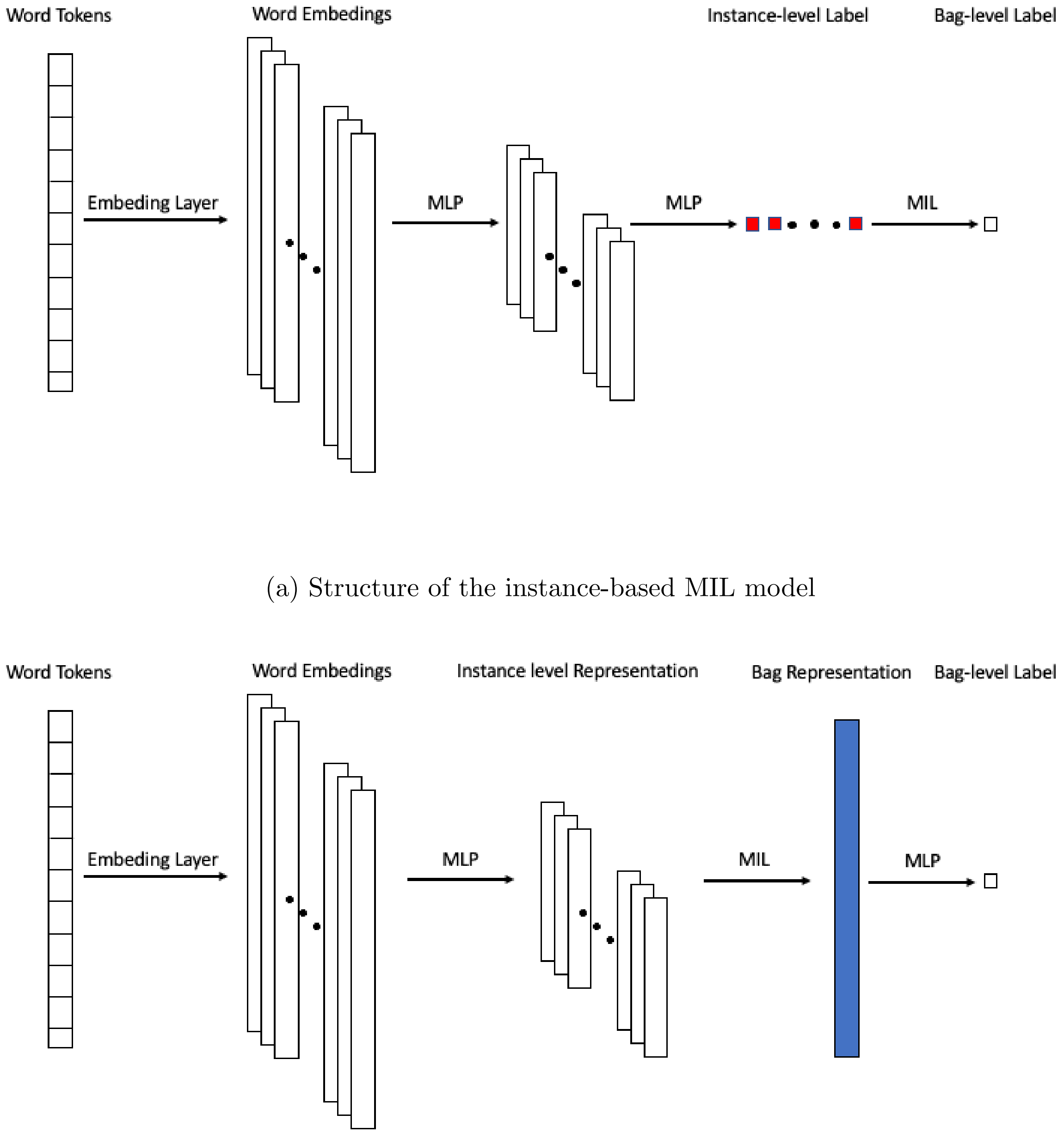
in contrast to the former approach, which induces the label for each instance in the bag by using the instance-based classifier, the latter approach fails to give each instance a label because it maps each instance into an uninterpretable low-dimensional embedding. This thesis demonstrates the process of modifying the second method to be more interpretable by using the attention mechanism with neural networks.

4.2 Multiple-instance Learning with Neural Networks

In most MIL classification problems, each feature is treated as an instance; $f(\cdot)$ is simply the identity function. For text data, however, the representation of a document can have a large number of dimensions. Thus, additional feature extraction is required. A feed-forward neural network $f_\psi(\cdot)$ with parameters ψ is used for parameterizing the function $f(\cdot)$ in Theorems 1 and 2. It transforms every instance x_k in the bag X into a lower-dimensional embedding $h_k = f_\psi(x_k)$ where $h_k \in H$. In the instance-based approach, $H = [0,1]$; Meanwhile, in the embedding-based approach, $H = \mathbb{R}^M$, where M is the embedding dimension. After aggregating all instance-level embeddings into one bag-level embedding $z \in H$, the $g(\cdot)$ function is also parameterized by a feed-forward neural network $g_\phi(\cdot)$ with parameters ϕ to predict the bag label Y. In the embedding-based approach, it maps the bag-level embedding z, into $[0,1]$. In the instance-based approach, the $g_\phi(\cdot)$ is an identity function, as $z \in H$ and $H = [0,1]$. An example of a model structure for text sentiment classification is illustrated in Figures 4.1a and 4.1b; Figure 4.1a is the structure for the instance-based MIL, whereas Figure 4.1b is the embedding-based approach.

4.3 Multiple-instance Learning Pooling

The definition of MIL requires that the pooling function $\sigma(\cdot)$ must be permutation-invariant. Theorems 1 and 2 provide examples of such pooling function, namely the max operator and the mean operator. Other operators such as convex max operator [63], noisy-or, and noisy-and [53], can replace the max operator in the Theorem 2; and a detailed proof of the replacement is presented in [62]. All of these alternative operators to the max operator are differential, which is suitable for any deep learning architecture.



(a) Structure of the instance-based MIL model

(b) Structure of the embedding-based MIL model

Figure 4.1: Structures of the different types of MIL models

4.4 Attention-based Multiple-instance Learning Pooling

As mentioned in [32], the preceding operators are characterized by some performance disadvantages because they are pre-defined and not trainable. For example, in binary classification, the max operator may be suitable for the instance-based approach because the bag will be positive (label 1) as long as any instance in the bag is positive, which aligns with the MIL assumption in (4.1). However, the max operator might fail in the embedding-based approach because every dimension in the hidden space is interpretable; simply selecting the max value among all the instances for every dimension to construct the bag-level embedding might not be adequately sophisticated. We use the attention mechanism as the new adaptive pooling function, which can learn a better pooling according to the specific data and task and potentially provide more interpretability than the pre-defined operators.

We use **self-attention** mentioned in Section 2.5.2 to replace the pooling function. It serves as the weighted average of the transformed instances (i.e., instance-level embeddings). Weights are learned by optimization training, and a final softmax layer is used to ensure that all the weights are summed up to one, which is invariant to the number of instances. Let $H_K = \{h_1, h_1, ..., h_K\}$ be a bag of K embeddings:

$$z = \sum_{k=1}^{K} \alpha_k h_k \tag{4.5}$$

where

$$\alpha_k = \frac{exp(v^T tanh(W h_k^T))}{\sum_{j=1}^{K} exp(v^T tanh(W h_j^T))} \tag{4.6}$$

where v and W are weights in MLP to be learned. The weighted average of h_k (4.5) satisfies Theorem 1 where $a_k h_k$ can been seen as a part of the $f(\cdot)$ function and the value of z (i.e., the bag-level embedding) does not depend on the permutation of h_k in bag H_K (i.e., permutation-invariant). The neural network structure is depicted in Figure 4.2. This proposed self-attention mechanism has been largely used for LSTM and transformer for text data. All the previous models take the sequence into consideration, but we assume that words are independent of each other (i.e., one-gram assumption). This assumption allows us to perform MIL models on text data. Moreover, the later analysis (see Chapter 5) suggests that it provides better interpretability and comparable performances with sequence models (i.e.,LSTMs)

4.4.1 Gated Attention Mechanism

The $tanh(\cdot)$ used in the aforementioned attention mechanism can be insufficient for learning the complexity because $tanh(x)$ is almost linear for $x \in [-1, 1]$. A gated attention mechanism[32] is proposed to add a more learnable non-linearity to the original $tanh(\cdot)$. The new attention score is calculated as follows:

$$\alpha_k = \frac{exp(v^T tanh(Wh_k^T) \odot sigm(Uh_k^T))}{\sum_{j=1}^{K} exp(v^T tanh(Wh_j^T) \odot sigm(Uh_j^T))} \tag{4.7}$$

where $\odot$ is the element-wise multiplication, and $sigm(\cdot)$ is the sigmoid function. However, in practice, the gated attention mechanism does not necessarily outperform the original one. Their comparison is presented in the experiment in Chapter 5.

4.4.2 Interpretability

As mentioned in Section 2.5.3, attention should provide some explanation for the model decision. In the case of MIL for the positive classification of the bag ($Y = 1$), high attention weight should be given to the instance that is most likely to be positive ($y_k = 1$). The attention scores naturally provide the interpretability for the deep learning models by giving the positive instances in the bag more weights. For example, in the case of sentiment analysis, if a sentence is predicted to be positive (bag label is positive, $Y = 1$), then positive words in the sentence should be given higher attention weights. Contrary to the instance-based approach, the attention-based MIL does not give an explicit label to each instance in the bag, but it provides some level of interpretability that the embedding-based approach does not offer.

4.4.3 Sentiment Analysis as Multiple-instance Learning

Although an ample number of studies have investigated multiple-instance learning, only few were conducted in the context of text data, in which the assumption is that each word is an instance and each sentence is a bag. The potential reason is that MIL makes the assumption that every instance in the bag is independent of each other, whereby the permutation of the instances does not matter (similar to the one-gram assumption); meanwhile, mainstream deep learning language models take the sequence of words into account. For example, LSTM learns the probability of a sentence by learning the accumulated conditional probability of all the words in order within the sentence. Language models effectively work in tasks such as text generation (i.e., generation of the next words based on the previous words), in which the sequence of the words plays an essential part of the task.

However, in the context of explainable sentiment analysis, we argue that the MIL model is better than sequence-based models because first, the sequence of the word is not as essential in the sentiment classification task as in most other NLP tasks; traditional bag-of-words models (no sequence) perform as effectively as the sequence-based model with much less computational resource. For example, in the IMDb movie review challenge (Section 2.1.2), Doc2vec with the bag-of-words method [77](no word sequence) ranks first in the competition, beating all the sequential language models with complicated structures and a large number of parameters such as BERT and LSTM.[2] Second, assuming the independence of each word in the sentences can provide more interpretability. Most language models use a complicated model structure to learn the interrelation between words, which makes the models difficult to explain. For example, the hidden state of LSTM at each time step contains not only the information of the current word but also all the words that are before it, causing difficulty in explaining the contribution of each word. However, in MIL, every instance only represents one word, which consequently facilitates the confirmation of the contribution of each word.

In this case, we propose MIL, as illustrated in Figure 4.2. Words in a sentence are presented by pre-train word embeddings, and an MLPs model transfers the representations to instance-level embeddings, followed by a self-attention layer to form a contextual vector to represent the bag (i.e., the bag-level embedding) by aggregating all instance-level embeddings; a bag-level classifier then is used for classifying the bag.

[2] The ranking can be found in `https://paperswithcode.com/sota/sentiment-analysis-on-imdb`.

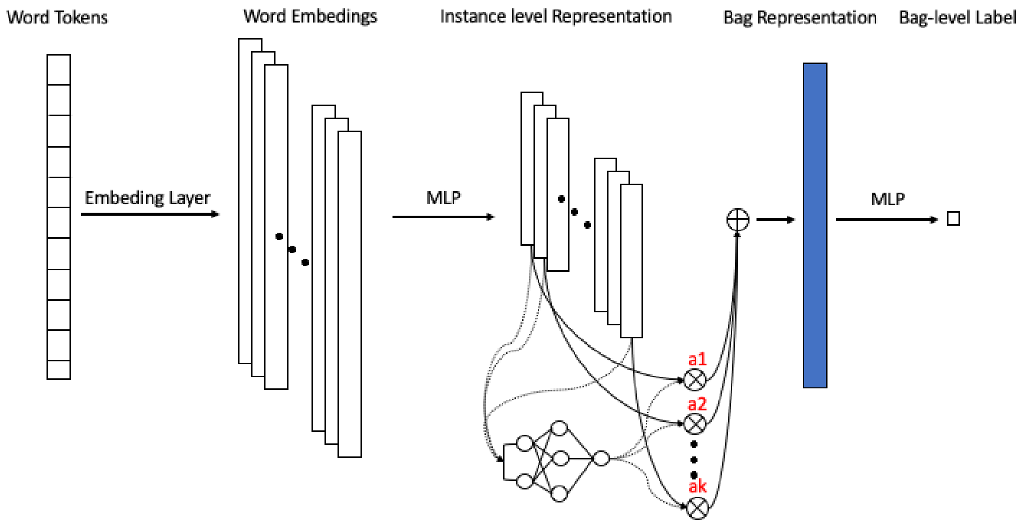


Figure 4.2: Structure of the attention-based MIL model; the MIL pooling is implemented by self-attention mechanisms

Chapter 5

Computational Results

5.1 Evaluation of the Extended Sentiment Lists

Generated by the sentiment lexicon induction (Section 3.2.3) using the senti-word2vec model (Section 3.2) with a baseline threshold of 4, the word list has 517 positive words and 465 negative words. We generate different sentiment lists for various thresholds. The word counts of each list are presented in Appendix A.1. This section describes the metrics used in this work and presents the results for every new dictionary.

5.1.1 MSCI Dataset

Companies of interest are in the MSCI universe. The target companies' quarterly conference calls from 2008 to 2018 are selected; companies with less than 10 conference calls during the period are omitted. The final list comprises 579 companies. The breakdown of the companies into each sector is presented in Table 5.1. For every conference call, we only keep **Answers** from **Executives** in the **Q & A Session**. Text pre-proposing includes tokenization, removal of special characters, and deletion of stop words.[1]

[1] Stop words are generally words that are not considered to add information content to the question at hand. The stop word list used here is the GeneraricLong list provided by the University of Notre Dame: https://sraf.nd.edu/textual-analysis/resources/.

Table 5.1: Company count for each sector in the MSCI universe

MSCI Sectors	**Company Count**
Communication services	30
Consumer discretionary	68
Consumer staples	32
Energy	31
Financials	80
Health care	73
Industrials	80
Information technology	86
Materials	28
Real estate	38
Utilities	30
Total	579

5.1.2 Polarity Scores

The quantification of the sentiment of a sentence as a polarity score is a common approach in sentiment analysis. The polarity score is calculated by some combination of the number of positive words, negative words, and all the words in a sentence. In this work, we delineate three different polarity scores to quantify the sentiment of a conference call of company i at event time t: $pos\_pcent^i_t$, $neg\_pcent^i_t$, and $posneg\_diff^i_t$. These scores are defined in (5.1).

$$\begin{aligned} pos\_pcent^i_t &= \frac{PW^i_t}{TW^i_t} \\ neg\_pcent^i_t &= \frac{NW^i_t}{TW^i_t} \\ posneg\_diff^i_t &= \frac{PW^i_t - NW^i_t}{TW^i_t} \end{aligned} \tag{5.1}$$

where PW^i_t is the number of positive words in the conference call of company i at event time t, NW^i_t is the number of negative words in the conference call of company i at event time t, and TW^i_t is the total number of words in the conference call of company i at event time t. We calculate all three scores for each conference call as the sentiment scores for the call.

5.1.3 Three-day Returns

Three-day excess return E_t^i at the event time t for the company i is defined in (5.2), where R_t^i is the three-day return at the event time t for the company i (5.3), and $\hat{I}_t$ is the three-day MSCI U.S. Index at time t (5.4).

$$E_t^i = R_t^i - \hat{I}_t \tag{5.2}$$

$$R_t^i = \frac{(P_{t+1}^i - P_{t-1}^i)}{P_{t-1}^i} \tag{5.3}$$

$$\hat{I}_t = \frac{(I_{t+1} - I_{t-1})}{I_{t-1}} \tag{5.4}$$

where P_{t+1}^i is the stock price of the company i on one business day after time t, P_{t-1}^i is the stock price of company i on one business day before time t, I_{t+1} is the MSCI US Index on one business day after time t, and I_{t-1} is the MSCI US Index on one business day before time t.

We believe that the three-day excess returns can capture the stock performance triggered by a certain event in excess of the benchmark (MSCI U.S. Index). The three-day excess returns E_t^i (5.2) of companies in MSCI universe on their quarterly conference call dates are calculated. We later use the three-day excess returns E_t^i to tune the threshold of the dictionary (Section 5.1.5) because it requires less computation. The three-day returns R_t^i (5.3) of all available conference calls in the dataset are calculated. We later use the three-day returns R_t^i as one variable of the correlation analysis with the polarity scores.

5.1.4 Kendall's τ Coefficient

Kendall's τ coefficient is a correlation measure for ordinal data. It measures the similarities of the ranks of the data when ranked based on their values [61]. The score will be high if the two observations being measured have a similar or the same ranking, and vice versa.

Let (x_1, y_1), (x_2, y_2),...,(x_n, y_n) be a set of observations of the joint random variables X and Y respectively. Any pair (x_i, y_i) and (x_j, y_j) is called *concordant* if and only if both $x_i < x_j$ and $y_i < y_j$ or if both $x_i > x_j$ and $y_i > y_j$; however, it is *discordant* if and only if both $x_i < x_j$ and $y_i > y_j$ or if both $x_i > x_j$ and $y_i < y_j$. The Kendall's τ is defined in (5.5). An explicit expression for Kendall's τ is defined in (5.6).

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}} \tag{5.5}$$

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j) \tag{5.6}$$

where $sgn(z)$ is the sign of z.

Note that the range of scores is between -1 and 1, where 1 suggests that x and y have a perfect positive correlation, -1 suggests that x and y have a perfect negative correlation, and 0 suggests that x and y are independent from each other.

5.1.5 Optimal Threshold

For every threshold, we calculated the Kendall's τs between three-day excess returns and two polarity scores: the *pos_pcent* and *neg_pcent* separately for each conference call of all the target companies. The result is shown in Table 5.2. From the results, we can observe that the threshold of 3.5 provides the best result for the **Negative** list and the threshold of 2.8 provides the best result for the **Positive** list. We combine the two optimal lists in addition to the original LM dictionary as the extended LM dictionary.

Table 5.2: Correlations between excess three-day returns and the sentiment scores calculated by the new dictionary with different thresholds

Threshold	**Neg_pcent**	**Pos_pcent**	**PosNeg_diff**
0	-5.86%	6.76%	8.05%
0.1	-5.86%	6.76%	8.05%
0.2	-5.86%	6.76%	8.05%
0.3	-5.86%	6.76%	8.05%
0.4	-5.86%	6.76%	8.05%
0.5	-5.86%	6.76%	8.05%
0.6	-5.86%	6.76%	8.05%
0.7	-5.79%	6.80%	8.04%
0.8	-5.79%	6.81%	8.04%
0.9	-5.78%	6.75%	8.05%
1.0	-5.79%	6.80%	8.13%
1.1	-5.80%	6.81%	8.15%
1.2	-5.79%	6.85%	8.18%

Continued on next page

Table 5.2 – continued from previous page

Threshold	Neg_pcent	Pos_pcent	PosNeg_diff
1.3	-5.81%	6.89%	8.45%
1.4	-5.82%	6.95%	8.45%
1.5	-5.82%	6.91%	8.46%
1.6	-5.86%	7.03%	8.53%
1.7	-5.88%	7.04%	8.55%
1.8	-5.81%	6.99%	8.51%
1.9	-5.79%	7.17%	8.58%
2.0	-5.86%	7.12%	8.56%
2.1	-5.87%	7.38%	8.70%
2.2	-5.79%	7.51%	8.49%
2.3	-5.83%	7.37%	8.48%
2.4	-5.86%	7.30%	8.50%
2.5	-5.79%	7.22%	8.33%
2.6	-5.79%	7.37%	8.55%
2.7	-5.82%	7.30%	8.41%
2.8	-5.79%	**7.55**%	8.41%
2.9	-5.87%	7.54%	8.36%
3.0	-6.18%	7.51%	8.38%
3.1	-6.21%	7.37%	8.61%
3.2	-6.23%	7.33%	8.56%
3.3	-6.31%	6.86%	8.14%
3.4	-6.21%	6.93%	8.19%
3.5	**-6.34**%	6.97%	8.32%
3.6	-6.28%	6.91%	8.29%
3.7	-6.13%	6.94%	8.24%
3.8	-6.22%	6.94%	8.22%
3.9	-5.95%	6.83%	8.12%
4.0	-6.16%	6.86%	8.21%

5.1.6 Word Comparison With the LM Dictionary

The extended LM dictionary consists of the original LM dictionary and 275 new positive words and 324 negative words. Furthermore, we stem[2] all the extra words and words in the LM dictionary into their word forms (e.g., “depressed”, “depression”, “depressing”, and “depresses” are all stemmed into the form “depress”). The LM dictionary contains 151 distinct positive word roots and 916 distinct negative word roots, while the new words generated contain 275 distinct positive word roots (See full list in Appendix A.2.1) and 269 distinct negative word roots (See full list in Appendix A.2.2). fourteen of the positive word roots and fifteen of the negative word roots are shared by the original LM dictionary as well. Table 5.3 displays the common word roots shared by both of the word lists.

Table 5.3: Word roots shared by both the new words and the words in the LM dictionary

Postive Word	Negative Word
excit	fatal
prosper	victim
pleas	depress
beauti	wast
inspir	neg
encourag	worri
desir	ban
profit	ridicul
effect	neglig
inspir	sever
insight	shock
confid	dispos
attract	turbul
solv	drop
advanc	inact
advanc	

[2] stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form.

5.1.7 Cross-sectional Analysis

We also conduct cross-sectional analysis to investigate the correlation between the three-day returns and the polarity scores on monthly bases.

5.1.7.1 Test Dataset

To evaluate the general performance of the dictionary, we decide to use data beyond just the companies within the MSCI universe. We use all conference calls available in the dataset (Section 2.1.1) from 2010 to 2018. For every conference call, we keep all sections namely, (1) presentation sections; (2) question sections; and (3) answer sections. There are a total 249,194 observations in presentation sections, 232,438 observations in question sections, and 243,785 observations in answer sections. Text proposing includes tokenization, removal of special characters, and deletion of stop words.[3]

5.1.7.2 Sample Groups

Each month's conference calls are categorized into different sample groups:

- All samples: all the conference calls;
- Positive samples: all the conference calls with positive $posneg\_diff$ (Section 5.1.2) calculated by LM dictionary, resulting in 29,611 observations in presentations sections, 138,564 observations in question sections, and 43,677 observations in answer sections;
- Negative samples: all the conference calls with negative $posneg\_diff$ (Section 5.1.2) calculated by LM dictionary, resulting in 215,742 observations in presentations sections, 78,007 observations in question sections, and 191,760 observations in answer sections;
- Top-bottom samples: all the conference calls with the top 10% and bottom 10% $posneg\_diff$ (Section 5.1.2) calculated by LM dictionary, resulting in 115,389 observations in presentations sections, 41,701 observations in question sections, and 65,790 observations in answer sections.

[3] Stop words are generally words that are not considered to add information content to the question at hand. The stop word list used here is the GenericLong list provided by the University of Notre Dame: https://sraf.nd.edu/textual-analysis/resources/.

In each sample group, conference calls are divided into three sections based on contents: presentations, questions, and answers. To evaluate the effectiveness of the extended LM dictionary relative to the LM dictionary, we calculate the Kendall's τ correlations between three-day returns and the polarity scores calculated by two different dictionaries.

5.1.7.3 Significant Out-performances

108 months (from 2010 to 2018) are analyzed using the aforementioned steps. We compare the **mean** and **median** of the correlation scores. T-test[70] and Mann–Whitney U-test[85] are also performed to determine if the out-performances of mean and median are statistically significant, respectively. Below, we only report statistically significant results.

5.1.7.3.1 All Samples

Question section: Table 5.4 illustrates that *neg_pcent* scores generated by the extended dictionary outperform the benchmark's mean by 1.15% and median by 0.95%.

Table 5.4: Comparison of the correlations generated by two dictionaries from the question section for all samples

	neg_pcent
LM mean (benchmark)	-4.85%
Extended LM mean	-6.00%
Outperform percentage	1.15%
P value (t-test)	0.63%
LM median (benchmark)	-4.87%
Extended LM median	-5.82%
Outperform percentage	0.95%
P value (u-test)	0.40%

5.1.7.3.2 Positive Samples

Question Section: Table 5.5 illustrates that *neg_pcent* scores generated by the extended dictionary outperform the benchmark's mean by 2.20% and median by 1.95%; *pos_pcent* scores generated by the extended dictionary outperform the benchmark's median by 1.12%; and $posneg\_diff$ scores generated by the extended dictionary outperform the benchmark's mean by 1.40% and median by 1.67%.

Table 5.5: Comparison of the correlations generated by two dictionaries from the question section for positive samples

	neg_pcent	*pos_pcent*	*posneg_diff*
LM mean (benchmark)	0.11%	3.39%	3.80%
Extended LM mean	-2.09%	4.52%	5.20%
Outperform percentage	2.20%	1.13%	1.40%
P value (t-test)	0.00%	12.75% (not significant)	4.04%
LM median (benchmark)	0.38%	3.71%	3.35%
Extended LM median	-1.57%	4.83%	5.97%
Outperform percentage	1.95%	1.12%	1.67%
P value (u-test)	0.00%	4.28%	1.90%

5.1.7.3.3 Negative Samples

Question Section: Table 5.6 illustrates that *neg_pcent* scores generated by the extended dictionary outperform the benchmark's mean by 1.69% and median by 1.34%; and *posneg_diff* scores generated by the extended dictionary outperform the benchmark's median by 0.50%.

Table 5.6: Comparison of the correlations generated by two dictionaries from the question section for negative samples

	neg_pcent	*posneg_diff*
LM mean (benchmark)	-0.86%	4.29%
Extended LM mean	-2.56%	4.93%
Outperform percentage	1.69%	0.64%
P value (t-test)	0.05%	19.43% (not significant)
LM median (benchmark)	-0.83%	4.89%
Extended LM median	-2.17%	5.38%
Outperform percentage	1.34%	0.50%
P value (u-test)	0.03%	8.28%

Presentation Section Table 5.7 illustrates that *pos_pcent* scores generated by the extended dictionary outperform the benchmark's median by 2.15%.

Table 5.7: Comparison of the correlations generated by two dictionaries from the presentation section for negative samples

	pos_pcent
LM mean (benchmark)	1.55%
Extended LM mean	2.52%
Outperform percentage	0.98%
P value (t-test)	47.01%
LM median (benchmark)	0.78%
Extended LM median	2.93%
Outperform percentage	2.15%
P value (u-test)	9.64%

5.1.7.3.4 Top-bottom Samples

Question Section: Table 5.8 illustrates that *neg_pcent* scores generated by the extended dictionary outperform the benchmark's median by 0.79%; and *posneg_diff* scores generated by the extended dictionary outperform the benchmark's median by 1.47%.

Table 5.8: Comparison of the correlations generated by two dictionaries from the question section for top-bottom samples

	neg_pcent	*posneg_diff*
LM mean (benchmark)	-0.19%	2.59%
Extended LM mean	-1.44%	3.73%
Outperform percentage	1.25%	1.14%
P value (t-test)	24.33% (not significant)	25.46% (not significant)
LM median (benchmark)	0.05%	2.24%
Extended LM median	-0.85%	3.67%
Outperform percentage	0.79%	1.42%
P value (u-test)	9.59%	1.47%

5.1.7.4 Discussion

The out-performance of *neg_pcent* scores is consistently significant in the question section for every sample group. The out-performance of *pos_pcent* scores is significant in the question section of the positive samples, and in the presentation section of the negative samples; and the out-performance of *posneg_diff* scores is significant in the question section of the positive samples, negative samples, and top-bottom samples.

Overall, the extended LM dictionary shows its edge over the benchmark dictionary on the question section of positive samples. Furthermore, *neg_pcent* scores of the extended dictionary also display advantages across every sample group.

5.1.7.5 Sensitivity Test

To evaluate the sensitivity of the dictionary in terms of different threshold, we construct a new dictionary based on the second highest correlation of *pos_pcent*, and *neg_pcent* scores in Table 5.2, namely threshold 3.3 for **Negative** list, and threshold 2.9 for **positive** list. We denote the new dictionary as the extended LM-sec dictionary. We perform the same cross-sectional evaluation on this new dictionary. Below, we only report statistically significant results.

5.1.7.5.1 All Samples

Question section: Table 5.9 illustrates that *neg_pcent* scores generated by the extended dictionary outperform the benchmark's mean by 1.03% and median by 0.81%.

Table 5.9: Comparison of the correlations generated by two dictionaries from the question section for all samples

	neg_pcent
LM mean (benchmark)	-4.85%
Extended LM-sec mean	-5.88%
Outperform percentage	1.03%
P value (t-test)	1.50%
LM median (benchmark)	-4.87%
Extended LM-sec median	-5.68%
Outperform percentage	0.81%
P value (u-test)	0.77%

5.1.7.5.2 Positive Samples

Question Section: Table 5.10 illustrates that *neg_pcent* scores generated by the extended dictionary outperform the benchmark's mean by 1.88% and median by 1.34%;

pos_pcent scores generated by the extended dictionary outperform the benchmark's median by 1.40%; and *posneg_diff* scores generated by the extended dictionary outperform the benchmark's mean by 1.18% and median by 1.55%.

Table 5.10: Comparison of the correlations generated by two dictionaries from the question section for positive samples

	neg_pcent	*pos_pcent*	*posneg_diff*
LM mean (benchmark)	0.11%	3.39%	3.80%
Extended LM-sec mean	-1.98%%	4.42%	4.98%
Outperform percentage	1.88%	1.02%	1.18%
P value (t-test)	0.02%	16.86% (not significant)	7.48%
LM median (benchmark)	0.38%	3.71%	3.35%
Extended LM-sec median	-1.72%	5.11%	4.91%
Outperform percentage	1.34%	1.40%	1.55%
P value (u-test)	0.00%	5.99%	2.21%

5.1.7.5.3 Negative Samples

Question Section: Table 5.11 illustrates that *neg_pcent* scores generated by the extended dictionary outperform the benchmark's mean by 1.65% and median by 1.59%.

Table 5.11: Comparison of the correlations generated by two dictionaries from the question section for negative samples

	neg_pcent
LM mean (benchmark)	-0.86%
Extended LM-sec mean	-2.51%
Outperform percentage	1.65%
P value (t-test)	0.09%
LM median (benchmark)	-0.83%
Extended LM-sec median	-2.42%
Outperform percentage	1.59%
P value (u-test)	0.03%

Presentation Section Table 5.12 illustrates that *pos_pcent* scores generated by the extended dictionary outperform the benchmark's median by 2.15%.

Table 5.12: Comparison of the correlations generated by two dictionaries from the presentation section for negative samples

	pos_pcent
LM mean (benchmark)	1.55%
Extended LM-sec mean	2.52%
Outperform percentage	0.98%
P value (t-test)	47.01%
LM median (benchmark)	0.78%
Extended LM-sec median	2.93%
Outperform percentage	2.15%
P value (u-test)	9.64%

5.1.7.5.4 Top-bottom Samples

Question Section: Table 5.13 illustrates that *neg_pcent* scores generated by the extended dictionary outperform the benchmark's median by 1.11%; and $posneg\_diff$ scores generated by the extended dictionary outperform the benchmark's median by 1.39%.

Table 5.13: Comparison of the correlations generated by two dictionaries from the question section for top-bottom samples

	neg_pcent	$posneg\_diff$
LM mean (benchmark)	-0.19%	2.59%
Extended LM-sec mean	-1.46%	3.78%
Outperform percentage	1.27%	1.19%
P value (t-test)	23.71% (not significant)	23.09% (not significant)
LM median (benchmark)	0.05%	2.24%
Extended LM-sec median	-1.16%	3.64%
Outperform percentage	1.11%	1.39%
P value (u-test)	9.45%	3.07%

The extended LM-sec dictionary generated by choosing the second highest thresholds, outlines similar significant out-performances with the extended LM dictionary. Both outperform the original LM dictionary in the same sections of the same sample groups with similar out-perform percentage. This illustrates that this method of extending the LM dictionary is not sensitive to a slight threshold change.

5.2 Evaluation of the Attention-based Multiple-instance Learning for Sentiment Analysis

In this section, we demonstrate the experiments to evaluate the attention-based multiple-instance learning model (att-MIL) (Section 4.4) on two different datasets, namely IMDB movie reviews (Section 2.1.2), and financial conference calls (Section 2.1.1). The evaluation tackles two major research questions: (1) Will the model have a analogous performance compared to the current state-of-the-art model on the binary bag-level classification (document sentiment classification where labels are 1s for positive documents and 0s for negative documents)? (2) Will the attention scores successfully capture the important words in the document? To answer the first question, we will use several metrics:

- $Classification\ Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
- ROC curve (receiver operating characteristic curve)[4]

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

All of the metrics are evaluated to show the comparison between att-based MIL, bi-directional LSTM, and bi-directional LSTM with self-attention. To answer the second question, we present the visualization of the document with the attention score for each word to ascertain whether the model is capable of capturing the most important word in a sentence. Finally, as a byproduct of the model, we can rank the words by their attention scores to construct an important word list in the corpus in terms of sentiment.

5.2.1 IMDb Dataset

Previous literature investigated the usefulness of the att-MIL model on the image dataset [32]. In this work, we examine how the model performs in the context of sentiment analysis

[4] The ROC curve is produced by plotting the true positive rate (TPR) against the false positive rate (FPR) at various classification threshold settings, where $TPR = \frac{TP}{TP+FN}$, and $FPR = \frac{FP}{FP+FN}$.

on the text data. We test our model on the standard sentiment dataset: the IMDb movie review dataset described in Section 2.1.2. The IMDb movie review dataset contains 25,000 polarized movie reviews in both training and testing datasets. The ground truths are labeled based on the scores associated to the reviews. A negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10. Each score is entered by the user who writes the review in the IMDb website.

We test the following four models in the experiment:

1. Bi-directional LSTM (Bi-LSTM) with 200 hidden neurons and 200 time steps implemented the same as in [25];

2. Bi-directional LSTM with self-attention Bi-LSTM-att) with 200 hidden neurons and 200 time steps implemented the same as in [21];

3. Attention-based multiple-instance learning(att-MIL) described in Section 4.4 (Figure 4.2); and

4. Gated attention-based multiple-instance learning (gated-att-MIL) described in Section 4.4.1 (Figure 4.2).

An exhaustive hyperparameter searching is impossible to conduct due to the limited computational resource. We use pre-train GloVe [60] word embeddings with 300 dimensions as the input layer to all models. The first two models (i.e., Bi-LSTM, and Bi-LSTM-att) are sequential models which consider the order of elements in inputs. In our case, the models take a movie review as an input with the correct permutation of words. For example, a review like “The movie is so good.“ has to be the exact permutation of (“The”, “movie”, “is”, “so”, “good”, “.”). Any change on the order of the words will result different prediction scores. Meanwhile, the last two models (att-MIL, and gated-att-MIL) treat the input review as a set (or bag) of words, which ignores the order of words. The models produce the same prediction scores regardless of the positions of words in a input like {“The”, “ movie”, “is”, “so”, “good”, “.”}.

Adam [40] is used for optimizing the models. Dropout layers with a rate of 0.5 are also used to prevent overfitting. The evaluation of the models is performed on the same test dataset with 25,000 examples of polarized movie reviews. We use the classification accuracy, precision, recall, F1 score, and ROC curve as the evaluation metrics.

The results are displayed in Table 5.14. The ROC curves are depicted in Figure 5.1. The results indicate that the gated-att-MIL and att-MIL can have a comparable performance with mainstream sequential models (i.e., LSTM) across all of the metrics. Such results represent evidence that the sequence of the words is not as important in the text classification task as the other NLP tasks, and the gated attention mechanism does not show significant advantages than the normal attention mechanism.

Table 5.14: Results on the IMDb test dataset

METHOD	ACCURACY	PRECISION	RECALL	F-SCORE
Bi-LSTM	0.8855	0.8878	0.8856	0.8854
Bi-LSTM-att	0.8874	0.8883	0.8874	0.8873
att-MIL	0.8856	0.8862	0.8856	0.8855
gated-att-MIL	**0.889**	**0.8891**	**0.889**	**0.889**

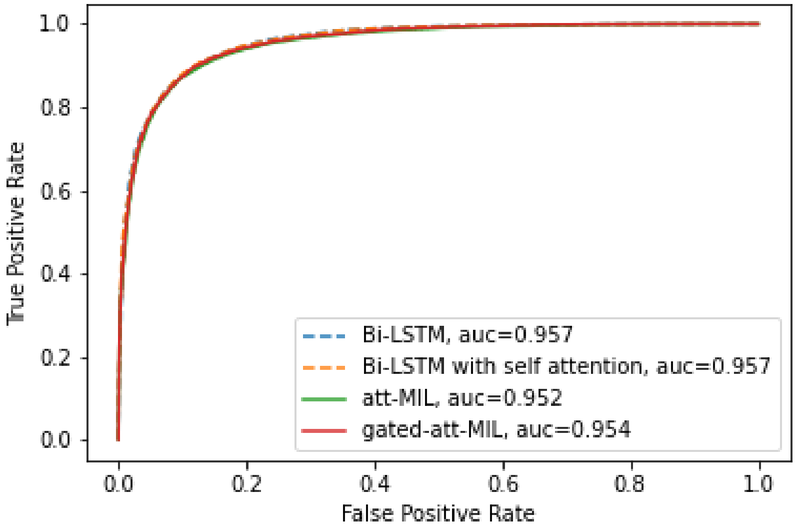


Figure 5.1: ROC curves for all the models

5.2.2 Highly Ranked Words

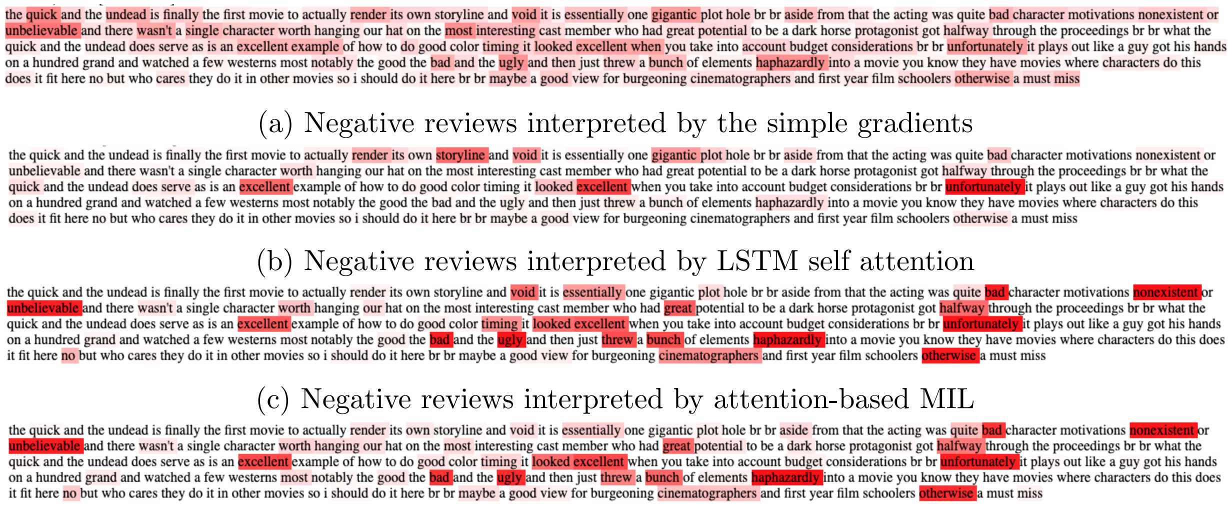
One unique feature of the att-MIL and gated-att-MIL is the capacity to generate an important word list based on the attention scores associated with each word. Most of the attention mechanisms in the previous literature are context-dependent; that is, a word will have different attention scores based on the context the word is in and the position the word is at in the sentence. For instance, the attention scores in the Bi-LSTM-att are non-deterministic. The same word will be assigned different attention scores based on its position in the inputs of the LSTM. However, our model att-MIL assumes that every instance (word) in the bag (sentences) is independent (one-gram assumption); thus, the model gives each word in the corpus a unique score regardless of the context. We assume that the attentions represent how important the model thinks the word is. We are able to rank the words based on the attention scores. Table 5.15 displays the top 10 most negative words ranked by the models. Appendix B.1 lists the top 500 words ranked by the attention scores. Both models can successfully extract the negative words. This feature is highly useful for binary classification problems for applications such as keyword extraction and interpretation of the model.

Table 5.15: Top 10 negative words ranked by att-MIL and gated-att-MIL based on the attention scores

Rank	att-MIL	gated-att-MIL
1	forgettable	unwatchable
2	unfunny	forgettable
3	unwatchable	unfunny
4	insipid	4/10
5	4/10	uninspired
6	uninspired	disappointing
7	disappointing	tedious
8	lackluster	lackluster
9	3/10	worst
10	tedious	underwhelming

5.2.3 Visualization

This section evaluates whether the attentions trained by the att-MIL and gated-attMIL are useful for explaining the decision of the model by highlighting the important words in the sentence. It also presents a comparison of the results of this work with the popular visualization methods proposed in the previous literature. Furthermore, this section provides the results of the four different methods for explaining the model, namely (1) gradient-based method (simple gradients) proposed by [83]; (2) LSTM self-attention [48]; (3) attention-based MIL (Section 4.4); and (4) gated attention-based MIL (Section 4.4.1).



the quick and the undead is finally the first movie to actually render its own storyline and void it is essentially one gigantic plot hole br br aside from that the acting was quite bad character motivations nonexistent or unbelievable and there wasn't a single character worth hanging our hat on the most interesting cast member who had great potential to be a dark horse protagonist got halfway through the proceedings br br what the quick and the undead does serve as is an excellent example of how to do good color timing it looked excellent when you take into account budget considerations br br unfortunately it plays out like a guy got his hands on a hundred grand and watched a few westerns most notably the good the bad and the ugly and then just threw a bunch of elements haphazardly into a movie you know they have movies where characters do this does it fit here no but who cares they do it in other movies so i should do it here br br maybe a good view for burgeoning cinematographers and first year film schoolers otherwise a must miss

(a) Negative reviews interpreted by the simple gradients

the quick and the undead is finally the first movie to actually render its own storyline and void it is essentially one gigantic plot hole br br aside from that the acting was quite bad character motivations nonexistent or unbelievable and there wasn't a single character worth hanging our hat on the most interesting cast member who had great potential to be a dark horse protagonist got halfway through the proceedings br br what the quick and the undead does serve as is an excellent example of how to do good color timing it looked excellent when you take into account budget considerations br br unfortunately it plays out like a guy got his hands on a hundred grand and watched a few westerns most notably the good the bad and the ugly and then just threw a bunch of elements haphazardly into a movie you know they have movies where characters do this does it fit here no but who cares they do it in other movies so i should do it here br br maybe a good view for burgeoning cinematographers and first year film schoolers otherwise a must miss

(b) Negative reviews interpreted by LSTM self attention

the quick and the undead is finally the first movie to actually render its own storyline and void it is essentially one gigantic plot hole br br aside from that the acting was quite bad character motivations nonexistent or unbelievable and there wasn't a single character worth hanging our hat on the most interesting cast member who had great potential to be a dark horse protagonist got halfway through the proceedings br br what the quick and the undead does serve as is an excellent example of how to do good color timing it looked excellent when you take into account budget considerations br br unfortunately it plays out like a guy got his hands on a hundred grand and watched a few westerns most notably the good the bad and the ugly and then just threw a bunch of elements haphazardly into a movie you know they have movies where characters do this does it fit here no but who cares they do it in other movies so i should do it here br br maybe a good view for burgeoning cinematographers and first year film schoolers otherwise a must miss

(c) Negative reviews interpreted by attention-based MIL

the quick and the undead is finally the first movie to actually render its own storyline and void it is essentially one gigantic plot hole br br aside from that the acting was quite bad character motivations nonexistent or unbelievable and there wasn't a single character worth hanging our hat on the most interesting cast member who had great potential to be a dark horse protagonist got halfway through the proceedings br br what the quick and the undead does serve as is an excellent example of how to do good color timing it looked excellent when you take into account budget considerations br br unfortunately it plays out like a guy got his hands on a hundred grand and watched a few westerns most notably the good the bad and the ugly and then just threw a bunch of elements haphazardly into a movie you know they have movies where characters do this does it fit here no but who cares they do it in other movies so i should do it here br br maybe a good view for burgeoning cinematographers and first year film schoolers otherwise a must miss

(d) Negative reviews interpreted by gated-attention-based MIL

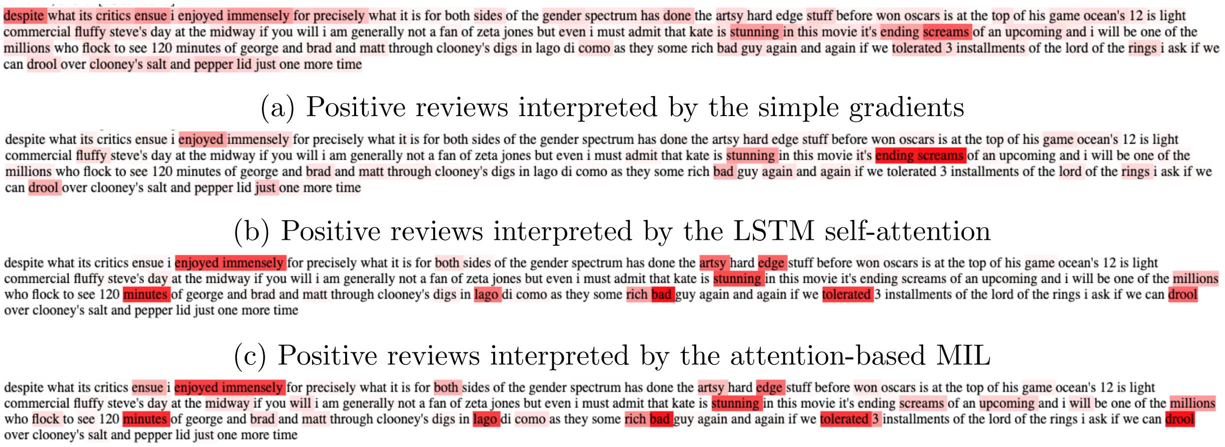
Figure 5.2: Comparison of the different interpretation methods; each score is rescaled as $s_k = (s_k - \min(S))/(\max(S) - \min(S))$

A negative example and a positive example are given in Figure 5.2 and Figure 5.3 respectively. Each method assigns a score to each word in the sentence, that is associated with the contribution of the word for making the prediction. We visualize sentences using red highlights based on the scores. Higher scores result in a darker highlight, and vice versa.

The simple gradient method (Figures 5.2a, 5.3a) scatteredly assigns the scores to most words in the sentence, hence failing to focus on the most important words for the decision making. The LSTM self-attention (Figures 5.2b, 5.3b) can successfully highlight some sentiment words in the sentence, but the interpretability is still unsatisfactory. Both att-MIL and gated-att-MIL (Figures 5.2c, 5.3c, 5.2d, 5.3d) can successfully assign higher scores to sentiment words that contribute to the decision of the models.

5.2.4 Multiple-instance Learning on Financial Conference Calls

This section investigates the performance, and use cases of att-MIL models on the financial conference calls. An att-MIL sentiment dictionary is generated for financial conference calls, and a correlation analysis is performed to compare the performance of the att-MIL dictionary with the LM dictionary and the extended LM dictionary.



despite what its critics ensue i enjoyed immensely for precisely what it is for both sides of the gender spectrum has done the artsy hard edge stuff before won oscars is at the top of his game ocean's 12 is light commercial fluffy steve's day at the midway if you will i am generally not a fan of zeta jones but even i must admit that kate is stunning in this movie it's ending screams of an upcoming and i will be one of the millions who flock to see 120 minutes of george and brad and matt through clooney's digs in lago di como as they some rich bad guy again and again if we tolerated 3 installments of the lord of the rings i ask if we can drool over clooney's salt and pepper lid just one more time

(a) Positive reviews interpreted by the simple gradients

despite what its critics ensue i enjoyed immensely for precisely what it is for both sides of the gender spectrum has done the artsy hard edge stuff before won oscars is at the top of his game ocean's 12 is light commercial fluffy steve's day at the midway if you will i am generally not a fan of zeta jones but even i must admit that kate is stunning in this movie it's ending screams of an upcoming and i will be one of the millions who flock to see 120 minutes of george and brad and matt through clooney's digs in lago di como as they some rich bad guy again and again if we tolerated 3 installments of the lord of the rings i ask if we can drool over clooney's salt and pepper lid just one more time

(b) Positive reviews interpreted by the LSTM self-attention

despite what its critics ensue i enjoyed immensely for precisely what it is for both sides of the gender spectrum has done the artsy hard edge stuff before won oscars is at the top of his game ocean's 12 is light commercial fluffy steve's day at the midway if you will i am generally not a fan of zeta jones but even i must admit that kate is stunning in this movie it's ending screams of an upcoming and i will be one of the millions who flock to see 120 minutes of george and brad and matt through clooney's digs in lago di como as they some rich bad guy again and again if we tolerated 3 installments of the lord of the rings i ask if we can drool over clooney's salt and pepper lid just one more time

(c) Positive reviews interpreted by the attention-based MIL

despite what its critics ensue i enjoyed immensely for precisely what it is for both sides of the gender spectrum has done the artsy hard edge stuff before won oscars is at the top of his game ocean's 12 is light commercial fluffy steve's day at the midway if you will i am generally not a fan of zeta jones but even i must admit that kate is stunning in this movie it's ending screams of an upcoming and i will be one of the millions who flock to see 120 minutes of george and brad and matt through clooney's digs in lago di como as they some rich bad guy again and again if we tolerated 3 installments of the lord of the rings i ask if we can drool over clooney's salt and pepper lid just one more time

(d) Positive reviews interpreted by the gated attention-based MIL

Figure 5.3: Comparison of the different interpretation methods; each score is rescaled as $s_k = (s_k - \min(S))/(\max(S) - \min(S))$
.

5.2.4.1 Dataset Construction

Quarterly conference calls (Section 2.1.1) from 2010 to 2018 are ranked based on their three-day returns (5.3) immediately after the conference call dates. The top 10% and bottom 10% of the sorted conference calls are selected to construct the new dataset. The top 10% conference calls are assigned to be positive samples, and the bottom 10% conference calls are assigned to be negative samples. To investigate the different predicting powers of each section of the conference calls to the three-day returns, we divide each conference call into three sections: presentations, questions, and answers, resulting in three different datasets. The presentation section has 25,035 observations with 20,028 observations for training, and 5,007 observations for testing. The question section has 24,930 observations with 19,944 observations for training, and 4,986 observations for testing. The answer section has 25,409 observations with 20,327 observations for training, and 5,082 observations for testing.

5.2.4.2 Models

Three attention-based MIL models (Section 4.4) are trained on three datasets, respectively, resulting in three different models: att-MIL on presentations, att-MIL on questions, and att-MIL on answers. The inputs of models are the pre-train GloVe embeddings[60] of the raw texts.

5.2.4.3 Out-of-sample Accuracy

Table 5.16 illustrates the out-of-sample accuracy of the models trained on different datasets. From the results, the question section offers the most information regarding the three-day returns. Surprisingly, the presentation section is relatively more correlated than the answer section regarding three-day returns. The presentation section is usually believed to have less predicting power because it is prepared by executives in advance.

Models	Test accuracy
att-MIL on presentations	65.0%
att-MIL on questions	67.4%
att-MIL on answers	62.5%

Table 5.16: The out-of-sample accuracy of att-MIL models on different datasets

5.2.4.4 Highly Ranked Words

We use the same method in Section 5.2.2 to rank the words for different sections of conference calls. The results of the word ranks after stemming are displayed in Table 5.17. It is noticeable that the presentation and answer sections share similar word distributions and ranks, as they both come from executives of companies while the question section is from financial analysts. Another interesting observation is that the model highly ranks positive words for the question section, but highly ranks negative words for the presentation and answer section. This observation may suggest that presentation and answer sections from executives generally have positive tones. This makes negative words more critical attributes to determine the actual labels of the text, while the general negative tones of the question section make positive words more critical for predicting labels.

Table 5.17: Top 50 words ranked by att-MIL models on different sections

Rank	**Presentation section**	**Question section**	**Answer section**
1	disappoint	delay	sustain
2	shortfal	shortfal	strength
3	delay	disappoint	impress
4	impact	caus	nice
5	challeng	soft	remark

Continued on next page

Table 5.17 – continued from previous page

Rank	Presentation section	Question section	Answer section
6	issu	paus	strong
7	slower	issu	gain
8	neg	impact	help
9	unfortun	cancel	outperform
10	slow	miss	improv
11	caus	blame	upsid
12	adasuv	disrupt	benefit
13	underestim	weak	congrat
14	frustrat	exacerb	phenomen
15	disrupt	sm	congratul
16	slowdown	resolv	excel
17	face	lose	accomplish
18	temporarili	felt	terrif
19	overcom	unfavor	drove
20	weak	formulari	disappoint
21	inabl	slow	weak
22	soft	3q	slowdown
23	unplan	temporari	soft
24	underperform	declin	weaker
25	pressur	feedback	delay
26	advers	anticip	shortfal
27	improv	impress	hope
28	sever	slip	lose
29	suspend	settl	slow
30	affect	combat	issu
31	lack	inabl	caus
32	momentum	action	deterior
33	unexpect	push	lost
34	exacerb	error	underperform
35	shutdown	impair	difficulti
36	declin	inflict	declin
37	setback	breach	coven
38	overrun	reset	awesom

Continued on next page

Table 5.17 – continued from previous page

Rank	Presentation section	Question section	Answer section
39	eros	frustrat	slower
40	stall	upset	unexpect
41	strong	react	challeng
42	pleas	imbal	wors
43	record	instabl	struggl
44	habit	slowdown	weaken
45	obes	materi	confus
46	quicksilv	lost	softer
47	decreas	sluggish	disconnect
48	encount	pronounc	deceler
49	runoff	coven	stronger
50	touch	urgenc	reacceler

5.2.4.5 Att-MIL Sentiment Dictionary

Section 5.2.4.4 illustrates that the att-MIL models can successfully rank the sentiment words regarding models' decisions. However, the word ranks also face limitations in distinguishing between positive and negative words. The observation from Table 5.17 shows that highly ranked words are a combination of positive and negative words. Thus, we manually divide the top 1,000 words in each section into positive and negative words. In the presentation section, we find 180 positive word roots (43 of them also found in the LM dictionary) and 248 negative word roots (138 of them also found in the LM dictionary); in the question section, we find 181 positive word roots (48 of them also found in the LM dictionary) and 191 negative word roots (91 of them also found the LM dictionary); and in the answer section, we find 139 positive word roots (26 of them also found in the LM dictionary) and 213 negative word roots (105 of them also found in the LM dictionary). The three sections above comprise our att-MIL sentiment dictionary for which every section of the conference calls has a customized corresponding sentiment word lists. Table 5.18 displays the top 50 sentiment word roots of each section in out att-MIL sentiment dictionaries.

Table 5.18: Top 50 word roots ranked by att-MIL models on different sections. We manually classify them into positive or negative words.

Rank	Presentation positive	Presentation negative	Answer positive	Answer negative	Question positive	Question negative
1	overcom	disappoint	resolv	delay	sustain	disappoint
2	improv	shortfal	settl	shortfal	strength	weak
3	strong	delay	pronounc	disappoint	impress	slowdown
4	pleas	impact	ramp	caus	awesom	soft
5	terrif	challeng	rebound	soft	nice	weaker
6	benefit	issu	address	paus	remark	delay
7	strength	slower	promot	issu	strong	shortfal
8	exceed	neg	lightn	impact	gain	hope
9	congratul	unfortun	valid	cancel	help	lose
10	profound	slow	recov	miss	outperform	slow
11	excel	caus	fix	blame	improv	drag
12	deliv	adasuv	deleverag	disrupt	upsid	issu
13	outperform	underestim	buildup	weak	benefit	caus
14	strongest	frustrat	epic	exacerb	congrat	deterior
15	correct	disrupt	acknowledg	lose	phenomen	lost
16	impress	slowdown	decis	unfavor	congratul	underperform
17	standout	weak	goodwil	formulari	excel	difficulti
18	remark	inabl	kickoff	slow	accomplish	declin
19	delight	soft	balloon	declin	terrif	slower
20	favor	unplan	desir	slip	drove	unexpect
21	address	underperform	activ	inabl	stronger	challeng
22	regain	pressur	expedit	action	reacceler	wors
23	resili	advers	experienc	push	confid	struggl
24	rais	sever	purpos	error	recoup	weaken
25	struggl	suspend	clear	impair	persist	confus
26	resolv	affect	undertak	inflict	increas	softer
27	stronger	lack	weaken	breach	pronounc	disconnect
28	upbeat	unexpect	smoothli	reset	epic	deceler
29	overperform	exacerb	allevi	frustrat	revis	disagre
30	achiev	shutdown	discoveri	upset	leap	sluggish
31	strengthen	declin	progress	imbal	reinsur	impact

Continued on next page

Table 5.18 – continued from previous page

Rank	Presentation positive	Presentation negative	Answer positive	Answer negative	Question positive	Question negative
32	handl	setback	deliber	instabl	ration	miss
33	nice	overrun	hope	slowdown	deleverag	cancel
34	gain	eros	upsel	materi	fix	ineffici
35	grew	stall	energ	lost	except	nonrecur
36	great	obes	instal	sluggish	amaz	burn
37	robust	decreas	understand	urgenc	clariti	slip
38	solidli	encount	warmer	litig	contribut	compens
39	help	runoff	wake	disput	reconcil	falloff
40	surpass	miss	explan	ineffici	special	pressur
41	recov	mistak	outright	bottom	fantast	push
42	pride	critic	logic	headwind	regain	disput
43	increas	obstacl	commit	apolog	permit	harp
44	enjoy	abus	environment	flowback	favor	linger
45	fantast	sluggish	confid	distract	top	overrun
46	posit	unaccept	correct	affect	sure	noncash
47	score	failur	complianc	challeng	wealth	mistaken
48	incred	loss	permit	termin	achiev	decreas
49	tailwind	problemat	programmat	hurt	honestli	incur
50	accomplish	outag	pass	softer	stabil	hurt

5.2.4.6 Att-MIL Dictionaries Excluding LM Dictionary

Table 5.19 displays the ranks of the top 50 word roots of the att-MIL Dictionaries, excluding the words in the LM dictionary. As discussed, the words in the LM dictionary are generated based on companies' financial reports. The new words generated are less formal when compared to the words in the LM dictionary due to the different word distributions between quarterly conference calls and the financial reports (i.e., conference calls consist of spoken English, while financial reports consist of written English). For example, new words such as "nice", "awesome", and "epic" are unusual for formal writing English, but they do carry sentimental meaning in spoken English when used in conference calls.

Table 5.19: Top 50 word roots of the att-MIL dictionaries excluding the word roots in the LM dictionary

Rank	Presentation positive	Presentation negative	Answer positive	Answer negative	Question positive	Question negative
1	overcom	shortfal	resolv	shortfal	sustain	soft
2	improv	impact	settl	caus	awesom	shortfal
3	terrif	challeng	pronounc	soft	nice	hope
4	exceed	issu	ramp	paus	remark	issu
5	congratul	neg	address	issu	help	caus
6	profound	unfortun	promot	impact	improv	deterior
7	excel	caus	lightn	blame	upsid	difficulti
8	deliv	adasuv	valid	exacerb	congrat	declin
9	correct	underestim	recov	unfavor	phenomen	unexpect
10	standout	frustrat	fix	formulari	congratul	challeng
11	remark	inabl	deleverag	declin	excel	wors
12	favor	soft	buildup	slip	terrif	struggl
13	address	unplan	epic	inabl	drove	confus
14	resili	pressur	acknowledg	action	reacceler	softer
15	rais	advers	decis	push	confid	disconnect
16	struggl	affect	goodwil	inflict	recoup	deceler
17	resolv	unexpect	kickoff	reset	increas	disagre
18	upbeat	exacerb	balloon	frustrat	pronounc	impact
19	overperform	declin	desir	imbal	epic	ineffici
20	achiev	eros	activ	instabl	revis	nonrecur
21	handl	stall	expedit	materi	leap	burn
22	nice	obes	experienc	urgenc	reinsur	slip
23	grew	decreas	purpos	litig	ration	compens
24	robust	encount	clear	disput	deleverag	falloff
25	solidli	runoff	undertak	ineffici	fix	pressur
26	help	mistak	smoothli	bottom	except	push
27	recov	critic	allevi	headwind	amaz	disput
28	pride	obstacl	discoveri	apolog	clariti	harp
29	increas	abus	deliber	flowback	contribut	linger
30	fantast	unaccept	hope	affect	reconcil	noncash
31	posit	failur	upsel	challeng	special	decreas

Continued on next page

Table 5.19 – continued from previous page

Rank	Presentation positive	Presentation negative	Answer positive	Answer negative	Question positive	Question negative
32	score	problemat	energ	termin	fantast	incur
33	incred	outag	instal	softer	permit	pushout
34	tailwind	hurdl	understand	pain	favor	whatsoev
35	underpin	tremend	warmer	downtick	top	forc
36	advoc	struggl	wake	compound	sure	postpon
37	enviabl	bump	explan	underestim	wealth	avoid
38	respond	protract	outright	adjust	achiev	unabl
39	brisk	devast	logic	correct	honestli	belabor
40	unriv	shortag	commit	wors	stabil	downgrad
41	keen	worri	environment	spillov	buyback	reiter
42	fortun	softer	confid	unfortun	save	lower
43	top	apolog	correct	rocki	tremend	reduc
44	steadili	postpon	complianc	postpon	solid	overcapac
45	subsid	undevelop	permit	hiccup	mutual	absent
46	except	modif	programmat	varianc	pass	unclear
47	propel	wors	pass	burst	respons	contract
48	healthi	unexpectedli	lift	burn	definit	uncertainti
49	prestig	lawsuit	elev	mute	exceed	redempt
50	energ	isol	respons	deterior	intrigu	apolog

5.2.4.7 Evaluation of att-MIL Sentiment Dictionaries

We follow the same steps in Section 5.1.7 to evaluate the Kendall's τ correlations (Section 5.1.4) between the polarity scores (Section 5.1.2) and the three-day returns (5.3) within same sets of sample groups (Section 5.1.7.2) on conference calls from 2010 to 2018. However, we removed the top-bottom 10%-return observations (All observations from the dataset in Section 5.2.4.1) from the evaluation dataset to avoid the self-attribution problem since we construct the att-MIL dictionary based on the three-day returns of the top-bottom 10%-return observations. We compare the **mean** and **median** of the correlation scores. T-test[70] and Mann–Whitney U-test[85] are also performed to determine if out-performances are statistically significant. Below, we only report statistically significant results.

5.2.4.7.1 All Samples

Question section: Table 5.20 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 3.35%, and median by 3.11%. The scores also outperform the extended LM's mean by 2.68%, and median by 2.40%. In addition, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 2.84%, and median by 2.99%. The scores also outperform extended LM's mean by 2.51%, and median 2.64%.

Table 5.20: Comparison of the correlations generated by three dictionaries in the question section for all samples

	neg_pcent	*posneg_diff*
LM mean (benchmark 1)	-2.85%	5.46%
Extended LM mean (benchmark 2)	-3.53%	5.79%
att-MIL mean	-6.21%	8.30%
Outperform percentage to LM	3.35%	2.84%
P value (t-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	2.68%	2.51%
P value (t-test) to extended LM	0.00%	0.00%
LM median (benchmark 1)	-2.71%	5.43%
Extended LM median (benchmark 2)	-3.42%	5.77%
att-MIL median	-5.82%	8.41%
Outperform percentage to LM	3.11%	2.99%
P value (u-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	2.40%	2.64%
P value (u-test) to extended LM	0.00%	0.00%

Presentation Section: Table 5.21 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 1.62%, and median by 1.73%. The scores also outperform extended LM's mean by 1.33%, and median by 1.44%. In addition, *pos_pcent* scores generated by the att-MIL dictionary outperform LM's mean by 1.79%, and median by 1.81%. The scores also outperform extended LM's mean by 2.43%, and median by 2.48%. Furthermore, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 2.63%, and median by 2.62%. The scores also outperform extended LM's mean by 3.28%, and median 3.40%.

Table 5.21: Comparison of the correlations generated by three dictionaries in the presentation section for all samples

	neg_pcent	*pos_pcent*	*posneg_diff*
LM mean (benchmark 1)	-4.70%	4.21%	5.92%
Extended LM mean (benchmark 2)	-4.99%	3.57%	5.26%
att-MIL mean	-6.32%	6.00%	8.55%
Outperform percentage to LM	1.62%	1.79%	2.63%
P value (t-test) to LM	0.09%	0.00%	0.00%
Outperform percentage to extended LM	1.33%	2.43%	3.28%
P value (t-test) to extended LM	0.54%	0.00%	0.00%
LM median (benchmark 1)	-4.32%	4.27%	5.97%
Extended LM median (benchmark 2)	-4.61%	3.60%	5.19%
att-MIL median	-6.05%	6.08%	8.59%
Outperform percentage to LM	1.73%	1.81%	2.62%
P value (u-test) to LM	0.00%	0.00%	0.00%
Outperform percentage to extended LM	1.44%	2.48%	3.40%
P value (u-test) to extended LM	0.00%	0.00%	0.00%

Answer Section: Table 5.22 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 3.04%, and median by 2.82%. The scores also outperform extended LM's mean by 3.21%, and median by 3.11%. In addition *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 2.75%, and median by 2.90%. The scores also outperform extended LM's mean by 3.16%, and median 2.85%.

Table 5.22: Comparison of the correlations generated by three dictionaries in the answer section for all samples

	neg_pcent	*posneg_diff*
LM mean (benchmark 1)	-2.52%	4.17%
Extended LM mean (benchmark 2)	-2.45%	3.97%
att-MIL mean	-4.11%	5.61%
Outperform percentage to LM	1.60%	1.45%
P value (t-test) to LM	0.11%	0.32%
Outperform percentage to extended LM	1.66%	1.64%
P value (t-test) to extended LM	0.08%	0.06%
LM median (benchmark 1)	-2.37%	3.94%
Extended LM median (benchmark 2)	-2.30%	3.57%
att-MIL median	3.99%	5.71%
Outperform percentage to LM	1.62%	1.76%
P value (u-test) to LM	0.00%	0.02%
Outperform percentage to extended LM	1.69%	2.14%
P value (u-test) to extended LM	0.00%	0.00%

5.2.4.7.2 Positive Samples

Question Section: Table 5.23 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 3.52%, and median by 3.15%. The scores also outperform extended LM's mean by 3.41%, and median by 3.35%. In addition, *pos_pcent* scores generated by the att-MIL dictionary outperform LM's mean by 2.05%, and median by 2.57%. The scores also outperform extended LM's mean by 1.09%, and median by 1.88%. Furthermore, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 4.29%, and median by 5.28%. The scores also outperform extended LM's mean by 3.33%, and median 3.58%.

Table 5.23: Comparison of the correlations generated by three dictionaries in the question section for positive samples

	neg_pcent	*pos_pcent*	*posneg_diff*
LM mean (benchmark 1)	0.66%	2.83%	2.72%
Extended LM mean (benchmark 2)	-0.78%	3.79%	3.67%
att-MIL mean	-4.18%	4.88%	7.01%
Outperform percentage to LM	3.52%	2.05%	4.29%
P value (t-test) to LM	0.00%	1.33%	0.00%
Outperform percentage to extended LM	3.41%	1.09%	3.33%
P value (t-test) to extended LM	0.00%	1.65%	0.00%
LM median (benchmark 1)	0.63%	3.12%	2.12%
Extended LM median (benchmark 2)	-0.42%	3.81%	3.82%
att-MIL median	-3.77%	5.69%	7.40%
Outperform percentage to LM	3.15%	2.57%	5.28%
P value (u-test) to LM	0.00%	0.06%	0.00%
Outperform percentage to extended LM	3.35%	1.88%	3.58%
P value (u-test) to extended LM	0.00%	1.50%	0.00%

Presentation Section: Table 5.24 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 1.71%, and median by 1.72%. The scores also outperform extended LM's mean by 1.32%, and median by 1.33%. In addition, *pos_pcent* scores generated by the att-MIL dictionary outperform LM's mean by 2.03% and median by 2.06%. The scores also outperform extended LM's mean by 2.70%, and median by 2.83%. Furthermore, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 2.72%, and median by 3.05%. The scores also outperform extended LM's mean by 3.50%, and median 3.82%.

Table 5.24: Comparison of the correlations generated by three dictionaries in the presentation section for positive samples

	neg_pcent	*pos_pcent*	*posneg_diff*
LM mean (benchmark 1)	-4.18%	3.62%	5.48%
Extended LM mean (benchmark 2)	-4.57%	2.95%	4.70%
att-MIL mean	-5.90%	5.65%	8.21%
Outperform percentage to LM	1.71%	2.03%	2.72%
P value (t-test) to LM	0.12%	0.00%	0.00%
Outperform percentage to extended LM	1.32%	2.70%	3.50%
P value (t-test) to extended LM	1.20%	0.00%	0.00%
LM median (benchmark 1)	-3.51%	3.50%	5.21%
Extended LM median (benchmark 2)	-3.89%	2.74%	4.45%
att-MIL median	-5.23%	5.56%	8.26%
Outperform percentage to LM	1.72%	2.06%	3.05%
P value (u-test) to LM	0.00%	0.00%	0.00%
Outperform percentage to extended LM	1.33%	2.83%	3.82%
P value (u-test) to extended LM	0.24%	0.00%	0.00%

Answer Section: Table 5.25 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 1.93%, and median by 2.19%. The scores also outperform extended LM's mean by 1.93%, and median by 2.05%. In addition, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 1.96%, and median by 2.05%. The scores also outperform extended LM's mean by 2.19%, and median 2.45%.

Table 5.25: Comparison of the correlations generated by three dictionaries in the answer section for positive samples

	neg_pcent	*posneg_diff*
LM mean (benchmark 1)	-1.86%	3.39%
Extended LM mean (benchmark 2)	-1.86%	2.15%
att-MIL mean	-3.79%	5.35%
Outperform percentage to LM	1.93%	1.96%
P value (t-test) to LM	0.02%	0.01%
Outperform percentage to extended LM	1.93%	2.19%
P value (t-test) to extended LM	0.03%	0.00%
LM median (benchmark 1)	-1.84%	3.39%
Extended LM median (benchmark 2)	-1.98%	2.99%
att-MIL median	-4.03%	5.44%
Outperform percentage to LM	2.19%	2.05%
P value (u-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	2.05%	2.45%
P value (u-test) to extended LM	0.00%	0.00%

5.2.4.7.3 Negative Samples

Question Section: Table 5.26 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 5.49%, and median by 4.86%. The scores also outperform extended LM's mean by 4.47%, and median by 3.87%. In addition, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 4.37% and median by 3.76%. The scores also outperform extended LM's mean by 3.88%, and median 3.67%.

Table 5.26: Comparison of the correlations generated by three dictionaries in the question section for negative samples

	neg_pcent	*posneg_diff*
LM mean (benchmark 1)	-0.31%	2.67%
Extended LM mean (benchmark 2)	-1.32%	3.16%
att-MIL mean	-5.80%	7.04%
Outperform percentage to LM	5.49%	4.37%
P value (t-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	4.47%	3.88%
P value (t-test) to extended LM	0.00%	0.00%
LM median (benchmark 1)	-0.18%	3.10%
Extended LM median (benchmark 2)	-1.17%	3.19%
att-MIL median	-5.04%	6.86%
Outperform percentage to LM	4.86%	3.76%
P value (u-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	3.87%	3.67%
P value (u-test) to extended LM	0.00%	0.00%

Presentation Section: Table 5.27 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 4.24%, and median by 4.93%. The scores also outperform extended LM's mean by 3.67%, and median by 4.87%. In addition, *pos_pcent* scores generated by the att-MIL dictionary outperform LM's mean by 3.09%, and median by 4.45%. The scores also outperform extended LM's mean by 3.80%, and median by 4.07%. Furthermore, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 6.75%, and median by 6.49%. The scores also outperform extended LM's mean by 7.11%, and median 6.77%.

Table 5.27: Comparison of the correlations generated by three dictionaries in the presentation section for negative samples

	neg_pcent	*pos_pcent*	*posneg_diff*
LM mean (benchmark 1)	-0.03%	0.00%	-0.11%
Extended LM mean (benchmark 2)	-0.61%	-0.72%	-0.47%
att-MIL mean	-4.28%	3.09%	6.64%
Outperform percentage to LM	4.24%	3.09%	6.75%
P value (t-test) to LM	0.00%	0.00%	0.00%
Outperform percentage to extended LM	3.67%	3.80%	7.11%
P value (t-test) to extended LM	0.00%	0.00%	0.00%
LM median (benchmark 1)	-0.66%	-0.89%	0.58%
Extended LM median (benchmark 2)	-0.72%	-0.50%	0.30%
att-MIL median	-5.59%	3.56%	7.07%
Outperform percentage to LM	4.93%	4.45%	6.49%
P value (u-test) to LM	0.00%	0.00%	0.00%
Outperform percentage to extended LM	4.87%	4.07%	6.77%
P value (u-test) to extended LM	0.00%	0.00%	0.00%

Answer Section: Table 5.28 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 4.40%, and median by 3.63%. The scores also outperform extended LM's mean by 3.71%, and median by 4.12%. In addition, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 3.14%, and median by 3.54%. The scores also outperform extended LM's mean by 1.95%, and median by 1.98%.

Table 5.28: Comparison of the correlations generated by three dictionaries in the answer section of for samples

	neg_pcent	*posneg_diff*
LM mean (benchmark 1)	0.14%	0.88%
Extended LM mean (benchmark 2)	-0.83%	2.08%
att-MIL mean	-4.54%	4.03%
Outperform percentage to LM	4.40%	3.14%
P value (t-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	3.71%	1.95%
P value (t-test) to extended LM	0.00%	0.00%
LM median (benchmark 1)	0.86%	0.24%
Extended LM median (benchmark 2)	0.37%	1.81%
att-MIL median	-4.49%	3.79%
Outperform percentage to LM	3.63%	3.54%
P value (u-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	4.12%	1.98%
P value (u-test) to extended LM	0.00%	0.00%

5.2.4.7.4 Top-bottom Samples

Question Section: Table 5.29 illustrates that *neg_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 5.59%, and median by 4.23%. The scores also outperform extended LM's mean by 5.10%, and median by 5.36%. In addition, *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 4.24%, and median by 3.68%. The scores also outperform extended LM's mean by 3.62%, and median 1.85%.

Table 5.29: Comparison of the correlations generated by three dictionaries in the question section for top-bottom samples

	neg_pcent	*posneg_diff*
LM mean (benchmark 1)	0.14%	1.55%
Extended LM mean (benchmark 2)	-0.63%	2.17%
att-MIL mean	-5.73%	5.79%
Outperform percentage to LM	5.59%	4.24%
P value (t-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	5.10%	3.62%
P value (t-test) to extended LM	0.00%	0.00%
LM median (benchmark 1)	1.17%	1.49%
Extended LM median (benchmark 2)	-0.03%	3.32%
att-MIL median	-5.40%	5.17%
Outperform percentage to LM	4.23%	3.68%
P value (u-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	5.36%	1.85%
P value (u-test) to extended LM	0.00%	0.00%

Presentation Section: Table 5.30 illustrates that *pos_pcent* scores generated by the att-MIL dictionary outperform the LM's mean by 2.64%, and median by 1.70%. The scores also outperform extended LM's mean by 2.89%, and median by 2.04%. In addition, *posneg_diff* scores generated by the att-MIL dictionary outperform LM's mean by 3.16%, and median by 2.97%. The scores also outperform extended LM's mean by 3.46%, and median by 4.76%.

Table 5.30: Comparison of the correlations generated by three dictionaries in the presentation section for top-bottom samples

	pos_pcent	*posneg_diff*
LM mean (benchmark 1)	2.49%	3.57%
Extended LM mean (benchmark 2)	2.24%	3.28%
att-MIL mean	5.13%	6.74%
Outperform percentage to LM	2.64%	3.16%
P value (t-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	2.89%	3.46%
P value (t-test) to extended LM	0.00%	0.00%
LM median (benchmark 1)	3.53%	4.55%
Extended LM median (benchmark 2)	3.19%	2.76%
att-MIL median	5.23%	7.52%
Outperform percentage to LM	1.70%	2.97%
P value (u-test) to LM	0.00%	0.00%
Outperform percentage to extended LM	2.04%	4.76%
P value (u-test) to extended LM	0.00%	0.00%

Answer Section: Table 5.31 illustrates that *posneg_diff* scores generated by the att-MIL dictionary outperform the LM's mean by 1.91%, and median by 2.67%. The scores also outperform extended LM's mean by 1.73%, and median by 2.00%.

Table 5.31: Comparison of the correlations generated by three dictionaries in the answer section for top-bottom samples

	$posneg\_diff$
LM mean (benchmark 1)	5.63%
Extended LM mean (benchmark 2)	5.80%
att-MIL mean	7.53%
Outperform percentage to LM	1.91%
P value (t-test) to LM	7.25%
Outperform percentage to extended LM	1.73%
P value (t-test) to extended LM	1.40%
LM median (benchmark 1)	5.64%
Extended LM median (benchmark 2)	6.21%
att-MIL median	8.21%
Outperform percentage to LM	2.67%
P value (u-test) to LM	5.99%
Outperform percentage to extended LM	2.00%
P value (u-test) to extended LM	0.01%

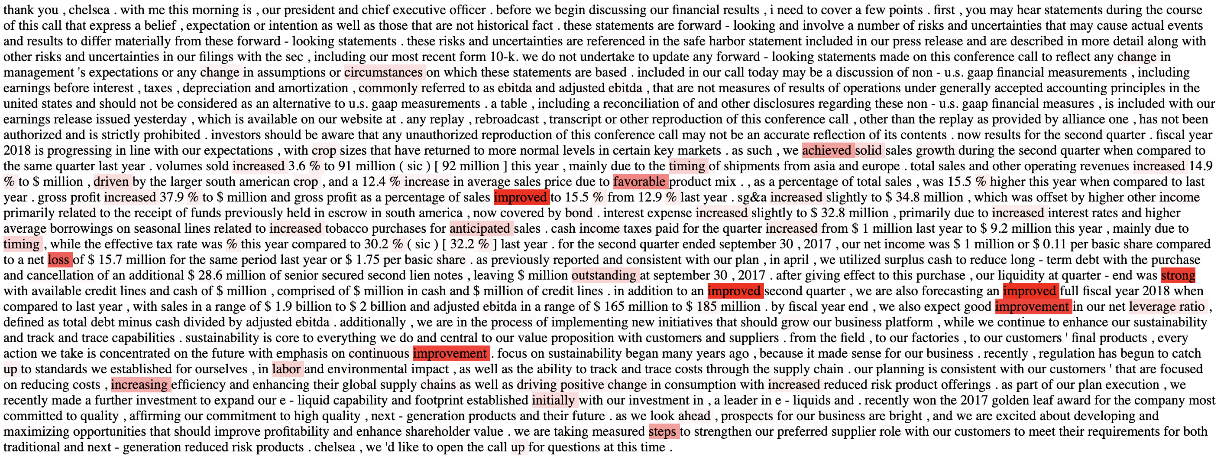
5.2.4.8 Discussion

The att-MIL sentiment dictionary shows noticeable advantages over both the LM dictionary and the extended LM dictionary in predicting the three-day returns based on conference calls in each sample group. It reaches its highest correlation in presentation sections of the all samples with an average 8.55% correlation between $posneg\_diff$ scores and three-day returns. It outperforms benchmarks the most on presentation sections of the negative samples with an average 6.64% correlation between $posneg\_diff$ scores and three-day returns while that of the LM dictionary is only -0.11% and that of the extended LM dicitonary is -0.47%. These results are also strong evidences that the att-MIL models can successfully pay more attentions to the important features highly related to the ground truths. In our case, the att-MIL models can rank the important sentiment words which are highly correlated to the three-day returns. This quality makes the model highly interpretable when most of the advanced deep learning models are black-box methods. Furthermore, the att-MIL sentiment dictionary illustrates this model can be effectively used to perform corpus-based sentiment lexicon induction.

5.2.5 Visualization for Conference Calls

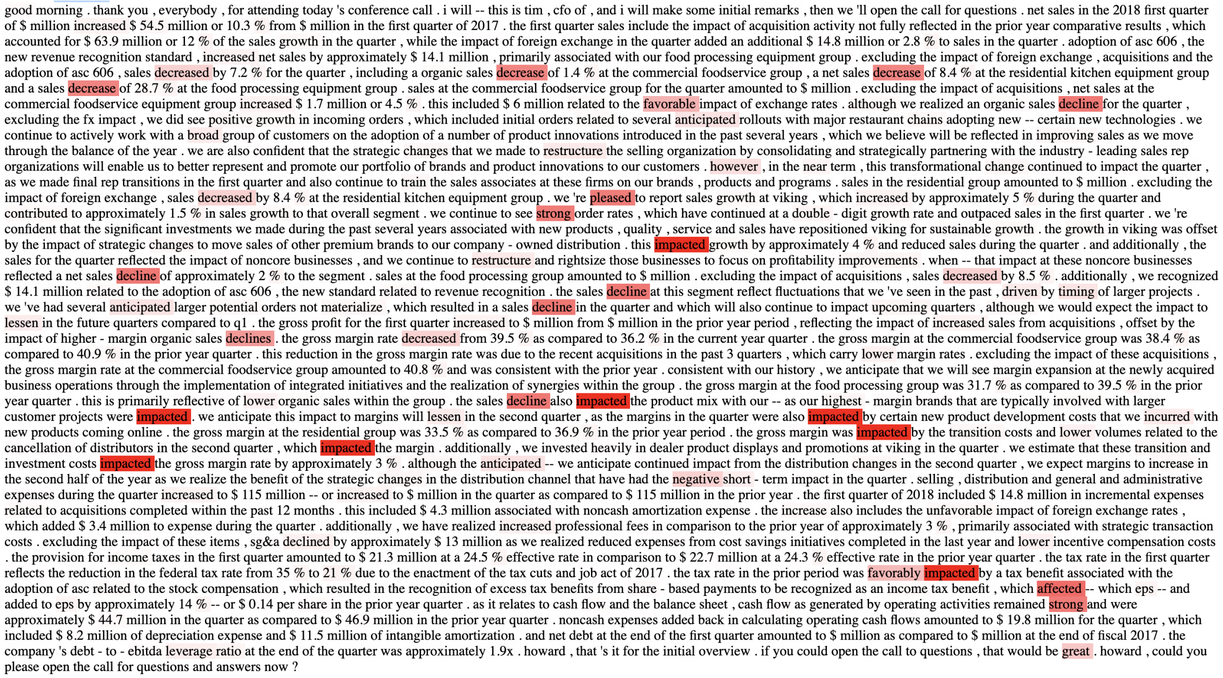
Applying the same att-MIL visualization method in Section 5.2.3, we use the attention scores to explain the model's decision by highlighting the important words in the sentence. We visualize sentences using red highlights based on the scores. Higher scores result in a darker highlight, and vice versa. This section provides examples of visualizations from different sections in conference calls, namely (1) the results of a presentation section are given in Figure 5.4; (2) the results of a question section are given in Figure 5.5; and (3) the results of an answer section are given in Figure 5.6.

The results of both positive examples (see Figures 5.4a, 5.5a and 5.6a) and negative examples (see Figures 5.4b, 5.5b and 5.6b) demonstrate that att-MIL model can assign higher attention scores to sentiment words that contribute to the decision of the model. This level of interpretability can help investors know the rationales behind the algorithm-based prediction when they make financial investment decisions based on conference calls.



thank you , chelsea . with me this morning is , our president and chief executive officer . before we begin discussing our financial results , i need to cover a few points . first , you may hear statements during the course of this call that express a belief , expectation or intention as well as those that are not historical fact . these statements are forward - looking and involve a number of risks and uncertainties that may cause actual events and results to differ materially from these forward - looking statements . these risks and uncertainties are referenced in the safe harbor statement included in our press release and are described in more detail along with other risks and uncertainties in our filings with the sec , including our most recent form 10-k. we do not undertake to update any forward - looking statements made on this conference call to reflect any change in management 's expectations or any change in assumptions or circumstances on which these statements are based . included in our call today may be a discussion of non - u.s. gaap financial measurements , including earnings before interest , taxes , depreciation and amortization , commonly referred to as ebitda and adjusted ebitda , that are not measures of results of operations under generally accepted accounting principles in the united states and should not be considered as an alternative to u.s. gaap measurements . a table , including a reconciliation of and other disclosures regarding these non - u.s. gaap financial measures , is included with our earnings release issued yesterday , which is available on our website at . any replay , rebroadcast , transcript or other reproduction of this conference call , other than the replay as provided by alliance one , has not been authorized and is strictly prohibited . investors should be aware that any unauthorized reproduction of this conference call may not be an accurate reflection of its contents . now results for the second quarter . fiscal year 2018 is progressing in line with our expectations , with crop sizes that have returned to more normal levels in certain key markets . as such , we achieved solid sales growth during the second quarter when compared to the same quarter last year . volumes sold increased 3.6 % to 91 million (sic) [92 million] this year , mainly due to the timing of shipments from asia and europe . total sales and other operating revenues increased 14.9 % to $ million , driven by the larger south american crop , and a 12.4 % increase in average sales price due to favorable product mix . , as a percentage of total sales , was 15.5 % higher this year when compared to last year . gross profit increased 37.9 % to $ million and gross profit as a percentage of sales improved to 15.5 % from 12.9 % last year . sg&a increased slightly to $ 34.8 million , which was offset by higher other income primarily related to the receipt of funds previously held in escrow in south america , now covered by bond . interest expense increased slightly to $ 32.8 million , primarily due to increased interest rates and higher average borrowings on seasonal lines related to increased tobacco purchases for anticipated sales . cash income taxes paid for the quarter increased from $ 1 million last year to $ 9.2 million this year , mainly due to timing , while the effective tax rate was % this year compared to 30.2 % (sic) [32.2 %] last year . for the second quarter ended september 30 , 2017 , our net income was $ 1 million or $ 0.11 per basic share compared to a net loss of $ 15.7 million for the same period last year or $ 1.75 per basic share . as previously reported and consistent with our plan , in april , we utilized surplus cash to reduce long - term debt with the purchase and cancellation of an additional $ 28.6 million of senior secured second lien notes , leaving $ million outstanding at september 30 , 2017 . after giving effect to this purchase , our liquidity at quarter - end was strong with available credit lines and cash of $ million , comprised of $ million in cash and $ million of credit lines . in addition to an improved second quarter , we are also forecasting an improved full fiscal year 2018 when compared to last year , with sales in a range of $ 1.9 billion to $ 2 billion and adjusted ebitda in a range of $ 165 million to $ 185 million . by fiscal year end , we also expect good improvement in our net leverage ratio , defined as total debt minus cash divided by adjusted ebitda . additionally , we are in the process of implementing new initiatives that should grow our business platform , while we continue to enhance our sustainability and track and trace capabilities . sustainability is core to everything we do and central to our value proposition with customers and suppliers . from the field , to our factories , to our customers ' final products , every action we take is concentrated on the future with emphasis on continuous improvement . focus on sustainability began many years ago , because it made sense for our business . recently , regulation has begun to catch up to standards we established for ourselves , in labor and environmental impact , as well as the ability to track and trace costs through the supply chain . our planning is consistent with our customers ' that are focused on reducing costs , increasing efficiency and enhancing their global supply chains as well as driving positive change in consumption with increased reduced risk product offerings . as part of our plan execution , we recently made a further investment to expand our e - liquid capability and footprint established initially with our investment in , a leader in e - liquids and . recently won the 2017 golden leaf award for the company most committed to quality , affirming our commitment to high quality , next - generation products and their future . as we look ahead , prospects for our business are bright , and we are excited about developing and maximizing opportunities that should improve profitability and enhance shareholder value . we are taking measured steps to strengthen our preferred supplier role with our customers to meet their requirements for both traditional and next - generation reduced risk products . chelsea , we 'd like to open the call up for questions at this time .

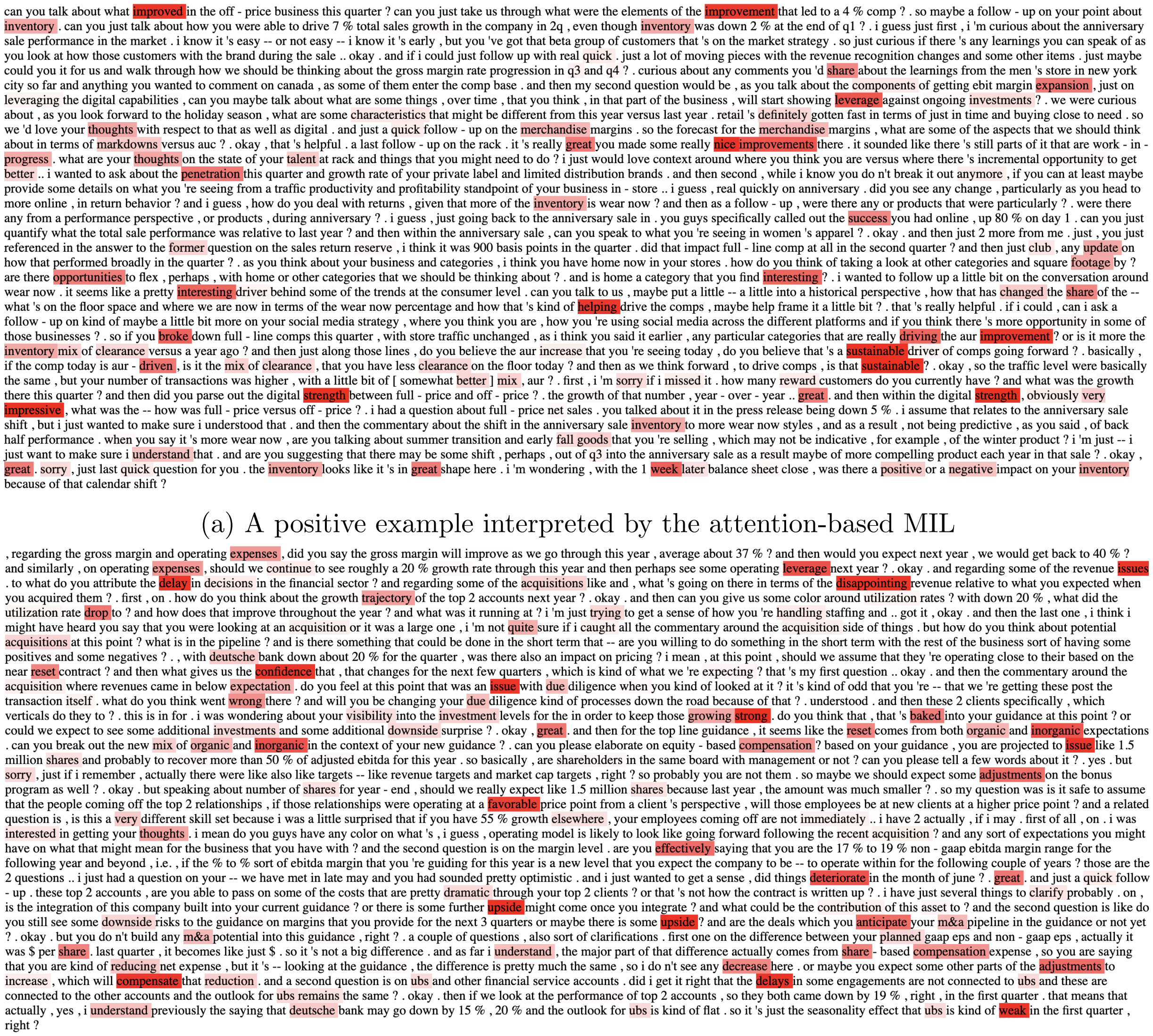
(a) A positive example interpreted by the attention-based MIL



good morning . thank you , everybody , for attending today 's conference call . i will -- this is tim , cfo of , and i will make some initial remarks , then we 'll open the call for questions . net sales in the 2018 first quarter of $ million increased $ 54.5 million or 10.3 % from $ million in the first quarter of 2017 . the first quarter sales include the impact of acquisition activity not fully reflected in the prior year comparative results , which accounted for $ 63.9 million or 12 % of the sales growth in the quarter , while the impact of foreign exchange in the quarter added an additional $ 14.8 million or 2.8 % to sales in the quarter . adoption of asc 606 , the new revenue recognition standard , increased net sales by approximately $ 14.1 million , primarily associated with our food processing equipment group . excluding the impact of foreign exchange , acquisitions and the adoption of asc 606 , sales decreased by 7.2 % for the quarter , including a organic sales decrease of 1.4 % at the commercial foodservice group , a net sales decrease of 8.4 % at the residential kitchen equipment group and a sales decrease of 28.7 % at the food processing equipment group . sales at the commercial foodservice group for the quarter amounted to $ million . excluding the impact of acquisitions , net sales at the commercial foodservice equipment group increased $ 1.7 million or 4.5 % . this included $ 6 million related to the favorable impact of exchange rates . although we realized an organic sales decline for the quarter , excluding the fx impact , we did see positive growth in incoming orders , which included initial orders related to several anticipated rollouts with major restaurant chains adopting new -- certain new technologies . we continue to actively work with a broad group of customers on the adoption of a number of product innovations introduced in the past several years , which we believe will be reflected in improving sales as we move through the balance of the year . we are also confident that the strategic changes that we made to restructure the selling organization by consolidating and strategically partnering with the industry - leading sales rep organizations will enable us to better represent and promote our portfolio of brands and product innovations to our customers . however , in the near term , this transformational change continued to impact the quarter , as we made final rep transitions in the first quarter and also continue to train the sales associates at these firms on our brands , products and programs . sales in the residential group amounted to $ million . excluding the impact of foreign exchange , sales decreased by 8.4 % at the residential kitchen equipment group . we 're pleased to report sales growth at viking , which increased by approximately 5 % during the quarter and contributed to approximately 1.5 % in sales growth to that overall segment . we continue to see strong order rates , which have continued at a double - digit growth rate and outpaced sales in the first quarter . we 're confident that the significant investments we made during the past several years associated with new products , quality , service and sales have repositioned viking for sustainable growth . the growth in viking was offset by the impact of strategic changes to move sales of other premium brands to our company - owned distribution . this impacted growth by approximately 4 % and reduced sales during the quarter . and additionally , the sales for the quarter reflected the impact of noncore businesses , and we continue to restructure and rightsize those businesses to focus on profitability improvements . when -- that impact at these noncore businesses reflected a net sales decline of approximately 2 % to the segment . sales at the food processing group amounted to $ million . excluding the impact of acquisitions , sales decreased by 8.5 % . additionally , we recognized $ 14.1 million related to the adoption of asc 606 , the new standard related to revenue recognition . the sales decline at this segment reflect fluctuations that we 've seen in the past , driven by timing of larger projects . we 've had several anticipated larger potential orders not materialize , which resulted in a sales decline in the quarter and which will also continue to impact upcoming quarters , although we would expect the impact to lessen in the future quarters compared to q1 . the gross profit for the first quarter increased to $ million from $ million in the prior year period , reflecting the impact of increased sales from acquisitions , offset by the impact of higher - margin organic sales declines . the gross margin rate decreased from 39.5 % as compared to 36.2 % in the current year quarter . the gross margin at the commercial foodservice group was 38.4 % as compared to 40.9 % in the prior year quarter . this reduction in the gross margin rate was due to the recent acquisitions in the past 3 quarters , which carry lower margin rates . excluding the impact of these acquisitions , the gross margin rate at the commercial foodservice group amounted to 40.8 % and was consistent with the prior year . consistent with our history , we anticipate that we will see margin expansion at the newly acquired business operations through the implementation of integrated initiatives and the realization of synergies within the group . the gross margin at the food processing group was 31.7 % as compared to 39.5 % in the prior year quarter . this is primarily reflective of lower organic sales within the group . the sales decline also impacted the product mix with our -- as our highest - margin brands that are typically involved with larger customer projects were impacted . we anticipate this impact to margins will lessen in the second quarter , as the margins in the quarter were also impacted by certain new product development costs that we incurred with new products coming online . the gross margin at the residential group was 33.5 % as compared to 36.9 % in the prior year period . the gross margin was impacted by the transition costs and lower volumes related to the cancellation of distributors in the second quarter , which impacted the margin . additionally , we invested heavily in dealer product displays and promotions at viking in the quarter . we estimate that these transition and investment costs impacted the gross margin rate by approximately 3 % . although the anticipated -- we anticipate continued impact from the distribution changes in the second quarter , we expect margins to increase in the second half of the year as we realize the benefit of the strategic changes in the distribution channel that have had the negative short - term impact in the quarter . selling , distribution and general and administrative expenses during the quarter increased to $ 115 million -- or increased to $ million in the quarter as compared to $ 115 million in the prior year . the first quarter of 2018 included $ 14.8 million in incremental expenses related to acquisitions completed within the past 12 months . this included $ 4.3 million associated with noncash amortization expense . the increase also includes the unfavorable impact of foreign exchange rates , which added $ 3.4 million to expense during the quarter . additionally , we have realized increased professional fees in comparison to the prior year of approximately 3 % , primarily associated with strategic transaction costs . excluding the impact of these items , sg&a declined by approximately $ 13 million as we realized reduced expenses from cost savings initiatives completed in the last year and lower incentive compensation costs . the provision for income taxes in the first quarter amounted to $ 21.3 million at a 24.5 % effective rate in comparison to $ 22.7 million at a 24.3 % effective rate in the prior year quarter . the tax rate in the first quarter reflects the reduction in the federal tax rate from 35 % to 21 % due to the enactment of the tax cuts and job act of 2017 . the tax rate in the prior period was favorably impacted by a tax benefit associated with the adoption of asc related to the stock compensation , which resulted in the recognition of excess tax benefits from share - based payments to be recognized as an income tax benefit , which affected -- which eps -- and added to eps by approximately 14 % -- or $ 0.14 per share in the prior year quarter . as it relates to cash flow and the balance sheet , cash flow as generated by operating activities remained strong and were approximately $ 44.7 million in the quarter as compared to $ 46.9 million in the prior year quarter . noncash expenses added back in calculating operating cash flows amounted to $ 19.8 million for the quarter , which included $ 8.2 million of depreciation expense and $ 11.5 million of intangible amortization . and net debt at the end of the first quarter amounted to $ million as compared to $ million at the end of fiscal 2017 . the company 's debt - to - ebitda leverage ratio at the end of the quarter was approximately 1.9x . howard , that 's it for the initial overview . if you could open the call to questions , that would be great . howard , could you please open the call for questions and answers now ?

(b) A negative example interpreted by the attention-based MIL

Figure 5.4: Examples of the presentation section interpreted by attention scores; each score is rescaled as $s_k = (s_k - \min(S))/(\max(S) - \min(S))$



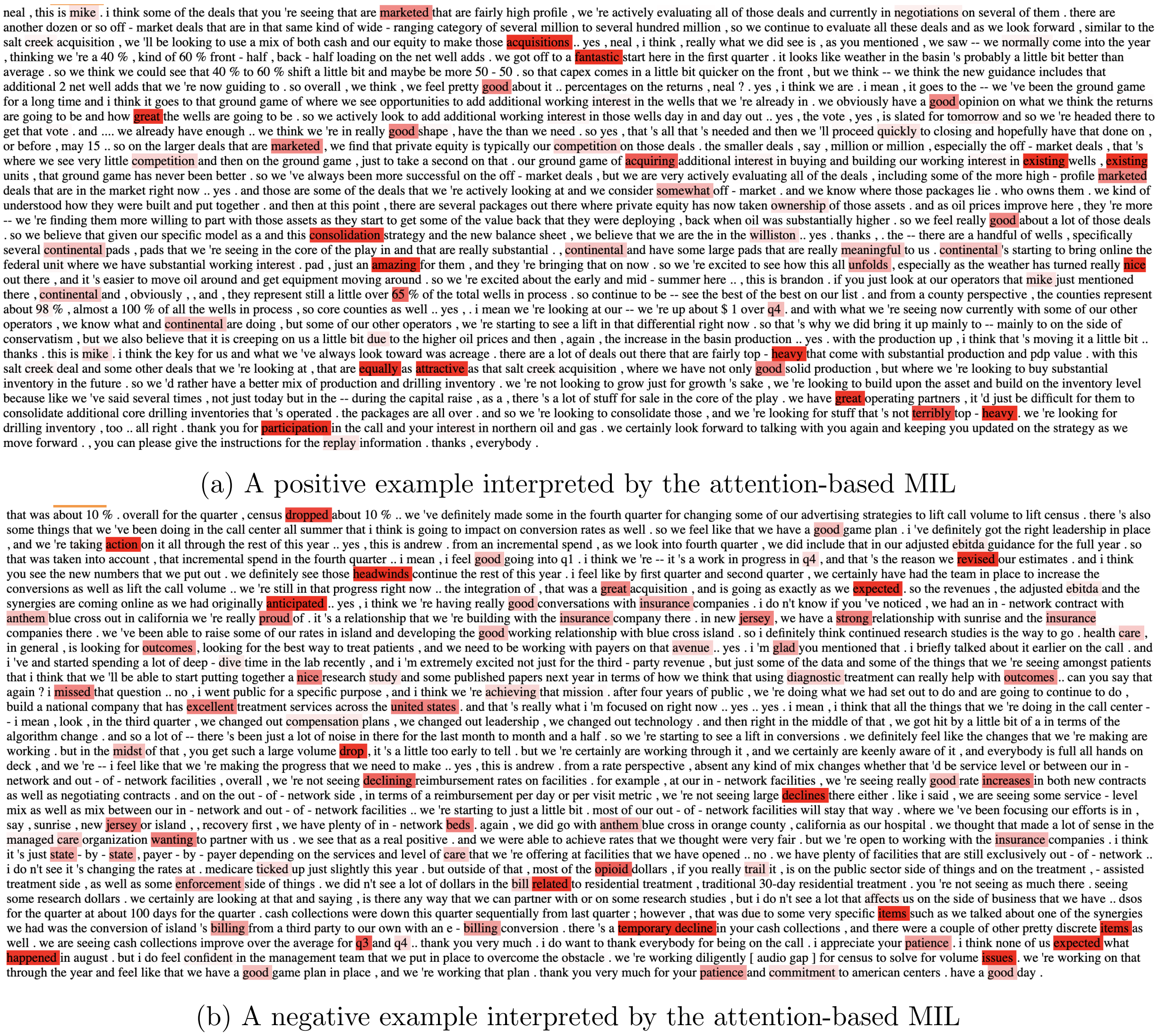
can you talk about what improved in the off - price business this quarter ? can you just take us through what were the elements of the improvement that led to a 4 % comp ? . so maybe a follow - up on your point about inventory . can you just talk about how you were able to drive 7 % total sales growth in the company in 2q , even though inventory was down 2 % at the end of q1 ? . i guess just first , i 'm curious about the anniversary sale performance in the market . i know it 's easy -- or not easy -- i know it 's early , but you 've got that beta group of customers that 's on the market strategy . so just curious if there 's any learnings you can speak of as you look at how those customers with the brand during the sale .. okay . and if i could just follow up with real quick . just a lot of moving pieces with the revenue recognition changes and some other items . just maybe could you it for us and walk through how we should be thinking about the gross margin rate progression in q3 and q4 ? . curious about any comments you 'd share about the learnings from the men 's store in new york city so far and anything you wanted to comment on canada , as some of them enter the comp base . and then my second question would be , as you talk about the components of getting ebit margin expansion , just on leveraging the digital capabilities , can you maybe talk about what are some things , over time , that you think , in that part of the business , will start showing leverage against ongoing investments ? . we were curious about , as you look forward to the holiday season , what are some characteristics that might be different from this year versus last year . retail 's definitely gotten fast in terms of just in time and buying close to need . so we 'd love your thoughts with respect to that as well as digital . and just a quick follow - up on the merchandise margins . so the forecast for the merchandise margins , what are some of the aspects that we should think about in terms of markdowns versus auc ? . okay , that 's helpful . a last follow - up on the rack . it 's really great you made some really nice improvements there . it sounded like there 's still parts of it that are work - in - progress . what are your thoughts on the state of your talent at rack and things that you might need to do ? i just would love context around where you think you are versus where there 's incremental opportunity to get better .. i wanted to ask about the penetration this quarter and growth rate of your private label and limited distribution brands . and then second , while i know you do n't break it out anymore , if you can at least maybe provide some details on what you 're seeing from a traffic productivity and profitability standpoint of your business in - store .. i guess , real quickly on anniversary . did you see any change , particularly as you head to more online , in return behavior ? and i guess , how do you deal with returns , given that more of the inventory is wear now ? and then as a follow - up , were there any or products that were particularly ? . were there any from a performance perspective , or products , during anniversary ? . i guess , just going back to the anniversary sale in . you guys specifically called out the success you had online , up 80 % on day 1 . can you just quantify what the total sale performance was relative to last year ? and then within the anniversary sale , can you speak to what you 're seeing in women 's apparel ? . okay . and then just 2 more from me . just , you just referenced in the answer to the former question on the sales return reserve , i think it was 900 basis points in the quarter . did that impact full - line comp at all in the second quarter ? and then just club , any update on how that performed broadly in the quarter ? . as you think about your business and categories , i think you have home now in your stores . how do you think of taking a look at other categories and square footage by ? are there opportunities to flex , perhaps , with home or other categories that we should be thinking about ? . and is home a category that you find interesting ? . i wanted to follow up a little bit on the conversation around wear now . it seems like a pretty interesting driver behind some of the trends at the consumer level . can you talk to us , maybe put a little -- a little into a historical perspective , how that has changed the share of the -- what 's on the floor space and where we are now in terms of the wear now percentage and how that 's kind of helping drive the comps , maybe help frame it a little bit ? . that 's really helpful . if i could , can i ask a follow - up on kind of maybe a little bit more on your social media strategy , where you think you are , how you 're using social media across the different platforms and if you think there 's more opportunity in some of those businesses ? . so if you broke down full - line comps this quarter , with store traffic unchanged , as i think you said it earlier , any particular categories that are really driving the aur improvement ? or is it more the inventory mix of clearance versus a year ago ? and then just along those lines , do you believe the aur increase that you 're seeing today , do you believe that 's a sustainable driver of comps going forward ? . basically , if the comp today is aur - driven , is it the mix of clearance , that you have less clearance on the floor today ? and then as we think forward , to drive comps , is that sustainable ? . okay , so the traffic level were basically the same , but your number of transactions was higher , with a little bit of [somewhat better] mix , aur ? . first , i 'm sorry if i missed it . how many reward customers do you currently have ? and what was the growth there this quarter ? and then did you parse out the digital strength between full - price and off - price ? . the growth of that number , year - over - year .. great . and then within the digital strength , obviously very impressive , what was the -- how was full - price versus off - price ? . i had a question about full - price net sales . you talked about it in the press release being down 5 % . i assume that relates to the anniversary sale shift , but i just wanted to make sure i understood that . and then the commentary about the shift in the anniversary sale inventory to more wear now styles , and as a result , not being predictive , as you said , of back half performance . when you say it 's more wear now , are you talking about summer transition and early fall goods that you 're selling , which may not be indicative , for example , of the winter product ? i 'm just -- i just want to make sure i understand that . and are you suggesting that there may be some shift , perhaps , out of q3 into the anniversary sale as a result maybe of more compelling product each year in that sale ? . okay , great . sorry , just last quick question for you . the inventory looks like it 's in great shape here . i 'm wondering , with the 1 week later balance sheet close , was there a positive or a negative impact on your inventory because of that calendar shift ?

(a) A positive example interpreted by the attention-based MIL

, regarding the gross margin and operating expenses , did you say the gross margin will improve as we go through this year , average about 37 % ? and then would you expect next year , we would get back to 40 % ? and similarly , on operating expenses , should we continue to see roughly a 20 % growth rate through this year and then perhaps see some operating leverage next year ? . okay . and regarding some of the revenue issues . to what do you attribute the delay in decisions in the financial sector ? and regarding some of the acquisitions like and , what 's going on there in terms of the disappointing revenue relative to what you expected when you acquired them ? . first , on . how do you think about the growth trajectory of the top 2 accounts next year ? . okay . and then can you give us some color around utilization rates ? with down 20 % , what did the utilization rate drop to ? and how does that improve throughout the year ? and what was it running at ? i 'm just trying to get a sense of how you 're handling staffing and .. got it , okay . and then the last one , i think i might have heard you say that you were looking at an acquisition or it was a large one , i 'm not quite sure if i caught all the commentary around the acquisition side of things . but how do you think about potential acquisitions at this point ? what is in the pipeline ? and is there something that could be done in the short term that -- are you willing to do something in the short term with the rest of the business sort of having some positives and some negatives ? . , with deutsche bank down about 20 % for the quarter , was there also an impact on pricing ? i mean , at this point , should we assume that they 're operating close to their based on the near reset contract ? and then what gives us the confidence that , that changes for the next few quarters , which is kind of what we 're expecting ? that 's my first question .. okay . and then the commentary around the acquisition where revenues came in below expectation . do you feel at this point that was an issue with due diligence when you kind of looked at it ? it 's kind of odd that you 're -- that we 're getting these post the transaction itself . what do you think went wrong there ? and will you be changing your due diligence kind of processes down the road because of that ? . understood . and then these 2 clients specifically , which verticals do they to ? . this is in for . i was wondering about your visibility into the investment levels for the in order to keep those growing strong . do you think that , that 's baked into your guidance at this point ? or could we expect to see some additional investments and some additional downside surprise ? . okay , great . and then for the top line guidance , it seems like the reset comes from both organic and inorganic expectations . can you break out the new mix of organic and inorganic in the context of your new guidance ? . can you please elaborate on equity - based compensation ? based on your guidance , you are projected to issue like 1.5 million shares and probably to recover more than 50 % of adjusted ebitda for this year . so basically , are shareholders in the same board with management or not ? can you please tell a few words about it ? . yes . but sorry , just if i remember , actually there were like also like targets -- like revenue targets and market cap targets , right ? so probably you are not them . so maybe we should expect some adjustments on the bonus program as well ? . okay . but speaking about number of shares for year - end , should we really expect like 1.5 million shares because last year , the amount was much smaller ? . so my question was is it safe to assume that the people coming off the top 2 relationships , if those relationships were operating at a favorable price point from a client 's perspective , will those employees be at new clients at a higher price point ? and a related question is , is this a very different skill set because i was a little surprised that if you have 55 % growth elsewhere , your employees coming off are not immediately .. i have 2 actually , if i may . first of all , on . i was interested in getting your thoughts . i mean do you guys have any color on what 's , i guess , operating model is likely to look like going forward following the recent acquisition ? and any sort of expectations you might have on what that might mean for the business that you have with ? and the second question is on the margin level . are you effectively saying that you are the 17 % to 19 % non - gaap ebitda margin range for the following year and beyond , i.e. , if the % to % sort of ebitda margin that you 're guiding for this year is a new level that you expect the company to be -- to operate within for the following couple of years ? those are the 2 questions .. i just had a question on your -- we have met in late may and you had sounded pretty optimistic . and i just wanted to get a sense , did things deteriorate in the month of june ? . great . and just a quick follow - up . these top 2 accounts , are you able to pass on some of the costs that are pretty dramatic through your top 2 clients ? or that 's not how the contract is written up ? . i have just several things to clarify probably . on , is the integration of this company built into your current guidance ? or there is some further upside might come once you integrate ? and what could be the contribution of this asset to ? and the second question is like do you still see some downside risks to the guidance on margins that you provide for the next 3 quarters or maybe there is some upside ? and are the deals which you anticipate your m&a pipeline in the guidance or not yet ? . okay . but you do n't build any m&a potential into this guidance , right ? . a couple of questions , also sort of clarifications . first one on the difference between your planned gaap eps and non - gaap eps , actually it was $ per share . last quarter , it becomes like just $. so it 's not a big difference . and as far i understand , the major part of that difference actually comes from share - based compensation expense , so you are saying that you are kind of reducing net expense , but it 's -- looking at the guidance , the difference is pretty much the same , so i do n't see any decrease here . or maybe you expect some other parts of the adjustments to increase , which will compensate that reduction . and a second question is on ubs and other financial service accounts . did i get it right that the delays in some engagements are not connected to ubs and these are connected to the other accounts and the outlook for ubs remains the same ? . okay . then if we look at the performance of top 2 accounts , so they both came down by 19 % , right , in the first quarter . that means that actually , yes , i understand previously the saying that deutsche bank may go down by 15 % , 20 % and the outlook for ubs is kind of flat . so it 's just the seasonality effect that ubs is kind of weak in the first quarter , right ?

(b) A negative example interpreted by the attention-based MIL

Figure 5.5: Examples of the question section interpreted by attention scores; each score is rescaled as $s_k = (s_k - \min(S))/(\max(S) - \min(S))$



neal , this is mike . i think some of the deals that you 're seeing that are marketed that are fairly high profile , we 're actively evaluating all of those deals and currently in negotiations on several of them . there are another dozen or so off - market deals that are in that same kind of wide - ranging category of several million to several hundred million , so we continue to evaluate all these deals and as we look forward , similar to the salt creek acquisition , we 'll be looking to use a mix of both cash and our equity to make those acquisitions .. yes , neal , i think , really what we did see is , as you mentioned , we saw -- we normally come into the year , thinking we 're a 40 % , kind of 60 % front - half , back - half loading on the net well adds . we got off to a fantastic start here in the first quarter . it looks like weather in the basin 's probably a little bit better than average . so we think we could see that 40 % to 60 % shift a little bit and maybe be more 50 - 50 . so that capex comes in a little bit quicker on the front , but we think -- we think the new guidance includes that additional 2 net well adds that we 're now guiding to . so overall , we think , we feel pretty good about it .. percentages on the returns , neal ? . yes , i think we are . i mean , it goes to the -- we 've been the ground game for a long time and i think it goes to that ground game of where we see opportunities to add additional working interest in the wells that we 're already in . we obviously have a good opinion on what we think the returns are going to be and how great the wells are going to be . so we actively look to add additional working interest in those wells day in and day out .. yes , the vote , yes , is slated for tomorrow and so we 're headed there to get that vote . and .... we already have enough .. we think we 're in really good shape , have the than we need . so yes , that 's all that 's needed and then we 'll proceed quickly to closing and hopefully have that done on , or before , may 15 .. so on the larger deals that are marketed , we find that private equity is typically our competition on those deals . the smaller deals , say , million or million , especially the off - market deals , that 's where we see very little competition and then on the ground game , just to take a second on that . our ground game of acquiring additional interest in buying and building our working interest in existing wells , existing units , that ground game has never been better . so we 've always been more successful on the off - market deals , but we are very actively evaluating all of the deals , including some of the more high - profile marketed deals that are in the market right now .. yes . and those are some of the deals that we 're actively looking at and we consider somewhat off - market . and we know where those packages lie . who owns them . we kind of understood how they were built and put together . and then at this point , there are several packages out there where private equity has now taken ownership of those assets . and as oil prices improve here , they 're more -- we 're finding them more willing to part with those assets as they start to get some of the value back that they were deploying , back when oil was substantially higher . so we feel really good about a lot of those deals . so we believe that given our specific model as a and this consolidation strategy and the new balance sheet , we believe that we are the in the williston .. yes . thanks , . the -- there are a handful of wells , specifically several continental pads , pads that we 're seeing in the core of the play in and that are really substantial . , continental and have some large pads that are really meaningful to us . continental 's starting to bring online the federal unit where we have substantial working interest . pad , just an amazing for them , and they 're bringing that on now . so we 're excited to see how this all unfolds , especially as the weather has turned really nice out there , and it 's easier to move oil around and get equipment moving around . so we 're excited about the early and mid - summer here .. , this is brandon . if you just look at our operators that mike just mentioned there , continental and , obviously , , and , they represent still a little over 65 % of the total wells in process . so continue to be -- see the best of the best on our list . and from a county perspective , the counties represent about 98 % , almost a 100 % of all the wells in process , so core counties as well .. yes , . i mean we 're looking at our -- we 're up about $ 1 over q4 . and with what we 're seeing now currently with some of our other operators , we know what and continental are doing , but some of our other operators , we 're starting to see a lift in that differential right now . so that 's why we did bring it up mainly to -- mainly to on the side of conservatism , but we also believe that it is creeping on us a little bit due to the higher oil prices and then , again , the increase in the basin production .. yes . with the production up , i think that 's moving it a little bit .. thanks . this is mike . i think the key for us and what we 've always look toward was acreage . there are a lot of deals out there that are fairly top - heavy that come with substantial production and pdp value . with this salt creek deal and some other deals that we 're looking at , that are equally as attractive as that salt creek acquisition , where we have not only good solid production , but where we 're looking to buy substantial inventory in the future . so we 'd rather have a better mix of production and drilling inventory . we 're not looking to grow just for growth 's sake , we 're looking to build upon the asset and build on the inventory level because like we 've said several times , not just today but in the -- during the capital raise , as a , there 's a lot of stuff for sale in the core of the play . we have great operating partners , it 'd just be difficult for them to consolidate additional core drilling inventories that 's operated . the packages are all over . and so we 're looking to consolidate those , and we 're looking for stuff that 's not terribly top - heavy . we 're looking for drilling inventory , too .. all right . thank you for participation in the call and your interest in northern oil and gas . we certainly look forward to talking with you again and keeping you updated on the strategy as we move forward . , you can please give the instructions for the replay information . thanks , everybody .

(a) A positive example interpreted by the attention-based MIL

that was about 10 % . overall for the quarter , census dropped about 10 % .. we 've definitely made some in the fourth quarter for changing some of our advertising strategies to lift call volume to lift census . there 's also some things that we 've been doing in the call center all summer that i think is going to impact on conversion rates as well . so we feel like that we have a good game plan . i 've definitely got the right leadership in place , and we 're taking action on it all through the rest of this year .. yes , this is andrew . from an incremental spend , as we look into fourth quarter , we did include that in our adjusted ebitda guidance for the full year . so that was taken into account , that incremental spend in the fourth quarter .. i mean , i feel good going into q1 . i think we 're -- it 's a work in progress in q4 , and that 's the reason we revised our estimates . and i think you see the new numbers that we put out . we definitely see those headwinds continue the rest of this year . i feel like by first quarter and second quarter , we certainly have had the team in place to increase the conversions as well as lift the call volume .. we 're still in that progress right now .. the integration of , that was a great acquisition , and is going as exactly as we expected . so the revenues , the adjusted ebitda and the synergies are coming online as we had originally anticipated .. yes , i think we 're having really good conversations with insurance companies . i do n't know if you 've noticed , we had an in - network contract with anthem blue cross out in california we 're really proud of . it 's a relationship that we 're building with the insurance company there . in new jersey , we have a strong relationship with sunrise and the insurance companies there . we 've been able to raise some of our rates in island and developing the good working relationship with blue cross island . so i definitely think continued research studies is the way to go . health care , in general , is looking for outcomes , looking for the best way to treat patients , and we need to be working with payers on that avenue .. yes . i 'm glad you mentioned that . i briefly talked about it earlier on the call . and i 've and started spending a lot of deep - dive time in the lab recently , and i 'm extremely excited not just for the third - party revenue , but just some of the data and some of the things that we 're seeing amongst patients that i think that we 'll be able to start putting together a nice research study and some published papers next year in terms of how we think that using diagnostic treatment can really help with outcomes .. can you say that again ? i missed that question .. no , i went public for a specific purpose , and i think we 're achieving that mission . after four years of public , we 're doing what we had set out to do and are going to continue to do , build a national company that has excellent treatment services across the united states . and that 's really what i 'm focused on right now .. yes .. yes . i mean , i think that all the things that we 're doing in the call center - - i mean , look , in the third quarter , we changed out compensation plans , we changed out leadership , we changed out technology . and then right in the middle of that , we got hit by a little bit of a in terms of the algorithm change . and so a lot of -- there 's been just a lot of noise in there for the last month to month and a half . so we 're starting to see a lift in conversions . we definitely feel like the changes that we 're making are working . but in the midst of that , you get such a large volume drop , it 's a little too early to tell . but we 're certainly are working through it , and we certainly are keenly aware of it , and everybody is full all hands on deck , and we 're -- i feel like that we 're making the progress that we need to make .. yes , this is andrew . from a rate perspective , absent any kind of mix changes whether that 'd be service level or between our in - network and out - of - network facilities , overall , we 're not seeing declining reimbursement rates on facilities . for example , at our in - network facilities , we 're seeing really good rate increases in both new contracts as well as negotiating contracts . and on the out - of - network side , in terms of a reimbursement per day or per visit metric , we 're not seeing large declines there either . like i said , we are seeing some service - level mix as well as mix between our in - network and out - of - network facilities .. we 're starting to just a little bit . most of our out - of - network facilities will stay that way . where we 've been focusing our efforts is in , say , sunrise , new jersey or island , , recovery first , we have plenty of in - network beds . again , we did go with anthem blue cross in orange county , california as our hospital . we thought that made a lot of sense in the managed care organization wanting to partner with us . we see that as a real positive . and we were able to achieve rates that we thought were very fair . but we 're open to working with the insurance companies . i think it 's just state - by - state , payer - by - payer depending on the services and level of care that we 're offering at facilities that we have opened .. no . we have plenty of facilities that are still exclusively out - of - network .. i do n't see it 's changing the rates at . medicare ticked up just slightly this year . but outside of that , most of the opioid dollars , if you really trail it , is on the public sector side of things and on the treatment , - assisted treatment side , as well as some enforcement side of things . we did n't see a lot of dollars in the bill related to residential treatment , traditional 30-day residential treatment . you 're not seeing as much there . seeing some research dollars . we certainly are looking at that and saying , is there any way that we can partner with or on some research studies , but i do n't see a lot that affects us on the side of business that we have .. dsos for the quarter at about 100 days for the quarter . cash collections were down this quarter sequentially from last quarter ; however , that was due to some very specific items such as we talked about one of the synergies we had was the conversion of island 's billing from a third party to our own with an e - billing conversion . there 's a temporary decline in your cash collections , and there were a couple of other pretty discrete items as well . we are seeing cash collections improve over the average for q3 and q4 .. thank you very much . i do want to thank everybody for being on the call . i appreciate your patience . i think none of us expected what happened in august . but i do feel confident in the management team that we put in place to overcome the obstacle . we 're working diligently [audio gap] for census to solve for volume issues . we 're working on that through the year and feel like that we have a good game plan in place , and we 're working that plan . thank you very much for your patience and commitment to american centers . have a good day .

(b) A negative example interpreted by the attention-based MIL

Figure 5.6: Examples of the answer section interpreted by attention scores; each score is rescaled as $s_k = (s_k - \min(S))/(\max(S) - \min(S))$

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Sentiment analysis is a highly active research field with numerous valuable applications in different domains. With the recent rapid progress of deep learning, more advanced methodologies have been successfully developed to improve sentiment analysis performance. However, in the finance domain, the industry still primarily uses the dictionary-based method for sentiment analysis for purposes of efficiency and interpretability. The sentiment dictionary method is highly domain-specific and expensive to manually create. The dictionary-based method's performance is also inferior to the performance of deep learning models. However, deep learning models are characterized by poor interpretability, whereby they typically fail to explain algorithm-based decision making. This thesis is an attempt to resolve the two problems in the finance domain.

We first trained the word2vec on company quarterly earning calls to learn the rich semantic relations between words in the corpus. We added sentiment information using an existing sentiment dictionary. This approach allowed us to extend the original dictionary to a more extensive one that is tailored to company quarterly earning calls. We also proposed a highly interpretable deep learning model to classify sentiment of documents with attentions associated with each word as the importance of the word for the decision making.

In the evaluation, we demonstrated that the sentiment polarity scores of the corporate quarterly conference calls calculated by the extended new dictionary have a higher correlation with three-day returns than the scores generated by the Loughran-McDonald dictionary. The att-MIL model proposed in Section 4.4 also exhibits a better performance in terms of sentiment classifications. Moreover, it successfully highlights the essential words in the sentences that help to explain the decision of the model. The att-MIL model can likewise generate a rank of the essential words in terms of the classification task. The new att-MIL sentiment dictionary generated based on the att-MIL model outperforms both the LM dictionary and the extended LM dictionary significantly on every sample group regarding three-day returns.

6.2 Potential Future Work

This work can be further developed in various directions. For the extended sentiment dictionary, only the word2vec embedding and one additional sentiment dictionary are used. Different word embedding algorithms and sentiment dictionaries can be used for generating broader sentiment lists. For the att-MIL model, we assumed that every instance in the bag is independent. However, a more reasonable assumption is that the words in a sentence depend on each other (the assumption of the language model [73]). Thus, future research can combine attention-based multiple-instance learning with the assumption of the dependent instances in the bag [95].

The new dictionaries in this thesis can be further evaluated. We can construct a monthly re-balanced long-short portfolio based on the polarity scores and the performance can be measured by backtesting. The results can show how beneficial the new dictionaries are to the real-world investment.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in Neural Information Processing Systems*, 15:561–568, 01 2002.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014.

[3] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[4] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Research-paper recommender systems : a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, 2016.

[5] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

[6] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . , 1999.

[7] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480, 1992.

[8] Sneha Chaudhari, Gungor Polatkan, Rohan Ramanath, and Varun Mithal. An attentive survey of attention models. *arXiv preprint arXiv:1904.02874*, 2019.

[9] Yixin Chen, Jinbo Bi, and James Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE transactions on pattern analysis and machine intelligence*, 28:1931–47, 01 2007.

[10] Veronika Cheplygina, Lauge Sørensen, David M. J. Tax, Marleen de Bruijne, and Marco Loog. Label stability in multiple instance learning. *CoRR*, abs/1703.04986, 2017.

[11] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[12] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. *Lexical acquisition: exploiting on-line resources to build a lexicon*, 115:164, 1991.

[13] Dan Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. pages 1237–1242, 07 2011.

[14] Dan C. Ciresan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *CoRR*, abs/1202.2745, 2012.

[15] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[16] Elizabeth Demers, Clara Vega, et al. Soft information in earnings announcements: News or noise? 2008.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[18] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1–2):31–71, January 1997.

[19] Joseph E Engelberg and Christopher A Parsons. The causal impact of media in financial markets. *The Journal of Finance*, 66(1):67–97, 2011.

[20] Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding topic signals in large-scale text. *CoRR*, abs/1602.06979, 2016.

[21] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

[22] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[23] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[24] Gene Golub and William Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224, 1965.

[25] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[26] Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 545–552. Curran Associates, Inc., 2009.

[27] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[28] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[29] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, July 1997. Association for Computational Linguistics.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[31] Jeremy Howard and Sebastian Ruder. Fine-tuned language models for text classification. *arXiv preprint arXiv:1801.06146*, page 194, 2018.

[32] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *CoRR*, abs/1802.04712, 2018.

[33] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[34] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[35] Jaap Kamps, Maarten Marx, Robert J Mokken, Maarten De Rijke, et al. Using wordnet to measure semantic orientations of adjectives. In *LREC*, volume 4, pages 1115–1118. Citeseer, 2004.

[36] Melih Kandemir, Chong Zhang, and Fred A Hamprecht. Empowering multiple instance histopathology cancer diagnosis by cell graphs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 228–235. Springer, 2014.

[37] Siavash Kazemian, Shunan Zhao, and Gerald Penn. Evaluating sentiment analysis evaluation: A case study in securities trading. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 119–127, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[38] James D. Keeler, David E. Rumelhart, and Wee Kheng Leow. Integrated segmentation and recognition of hand-printed numerals. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 557–563. Morgan-Kaufmann, 1991.

[39] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[41] Svetlana Kiritchenko, Xiaodan Zhu, and Saif Mohammad. Sentiment analysis of short informal text. *The Journal of Artificial Intelligence Research (JAIR)*, 50, 08 2014.

[42] Sabino P Kothari, Xu Li, and James E Short. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, 84(5):1639–1670, 2009.

[43] Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606, 2015.

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[45] Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. 1997.

[46] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.

[47] Feng Li. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102, 2010.

[48] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[49] Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. Key instance detection in multi-instance learning. 2012.

[50] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[51] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[52] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998.

[53] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998.

[54] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural networks : the official journal of the International Neural Network Society*, 16:555–9, 06 2003.

[55] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. pages 1–12, 01 2013.

[56] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of wordsd and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[57] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[58] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 01 2008.

[59] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR*, abs/1211.5063, 2012.

[60] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[61] Llukan Puka. Kendall's tau. In *International Encyclopedia of Statistical Science*, pages 713–715. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[62] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016.

[63] Jan Ramon and Luc De Raedt. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pages 53–60, 2000.

[64] Delip Rao and Deepak Ravichandran. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, page 675–682, USA, 2009. Association for Computational Linguistics.

[65] Vikas C Raykar, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, and R Bharat Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proceedings of the 25th international conference on Machine learning*, pages 808–815, 2008.

[66] Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Dür, and Linda Anderson. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. *arXiv preprint arXiv:1702.01978*, 2017.

[67] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ẅhy should i trust you?̈: Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[68] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[69] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, page 318–362. MIT Press, Cambridge, MA, USA, 1986.

[70] Oscar Sheynin. Helmert's work in the theory of errors. *Archive for history of exact sciences*, 49(1):73–104, 1995.

[71] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[72] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017.

[73] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

[74] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[75] Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. Sentiment embeddings with applications to sentiment analysis. *IEEE transactions on knowledge and data Engineering*, 28(2):496–509, 2015.

[76] Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[77] Tan Thongtan and Tanasanee Phienthrakul. Sentiment classification using document embeddings trained with cosine similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 407–414, Florence, Italy, July 2019. Association for Computational Linguistics.

[78] Peter D Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*, 2002.

[79] Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[80] Maria V. Valueva, Nikolay N. Nagornov, Pave A. Lyakhov, Georgiy V. Valuev, and Nikolay I. Chervyakov. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation*, 2020.

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[82] G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

[83] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. Allennlp interpret: A framework for explaining predictions of nlp models. *ArXiv*, abs/1909.09251, 2019.

[84] Amy Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45, 02 2013.

[85] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[86] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[87] Wei Xu, Alan Ritter, Chris Callison-Burch, William Dolan, and Yangfeng Ji. Extracting lexically divergent paraphrases from twitter. *Transactions of the Association for Computational Linguistics*, 2:435–448, 12 2014.

[88] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019.

[89] 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.

[90] Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems*, pages 887–898, 2018.

[91] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[92] Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[93] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.

[94] Wei Zhang, Kazuyoshi Itoh, Jun Tanida, and Yoshiki Ichioka. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. *Appl. Opt.*, 29(32):4790–4797, Nov 1990.

[95] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256, 2009.

APPENDICES

Appendix A

Statistics of the Extended Sentiment Lists

A.1 Word counts for lists of different thresholds

Table A.1: Word counts for the new lists with different threshold. Each list contains only the extra words(words that does not appear in the LM Dictionary)

Threshold	Positive List	Negative List
0	0	0
0.1	0	0
0.2	0	0
0.3	0	0
0.4	0	0
0.5	0	0
0.6	0	1
0.7	1	3
0.8	2	4
0.9	3	7
1.0	7	11
1.1	9	16
1.2	14	18

Continued on next page

Table A.1 – continued from previous page

Threshold	Positive List	Negative List
1.3	19	21
1.4	29	27
1.5	44	34
1.6	57	37
1.7	71	43
1.8	84	52
1.9	103	67
2.0	121	85
2.1	136	92
2.2	149	105
2.3	169	112
2.4	188	124
2.5	206	135
2.6	226	152
2.7	250	164
2.8	275	181
2.9	301	200
3.0	336	225
3.1	353	236
3.2	368	253
3.3	405	287
3.4	421	298
3.5	447	324
3.6	457	341
3.7	477	367
3.8	496	397
3.9	509	446
4.0	517	465

A.2 Extended dictionary

A.2.1 Extra Positive Words

This section will displays all the extra positive words in the new extended dictionary. All words in this section does not include the words in the current LM dictionary.

Index	Word Root	Word Example
1	fun	fun
2	joy	joy
3	honest	honest
4	comedi	comedy
5	bonu	bonus
6	prize	prize
7	entertain	entertaining
8	knowledg	knowledgeable
9	talent	talented
10	award	award
11	awesom	awesome
12	celebr	celebrate
13	fabul	fabulous
14	relationship	relationship
15	excit	excite
16	thank	thankful
17	healthi	healthy
18	amaz	amazing
19	pretti	pretty
20	save	savings
21	comfort	comfortable
22	courag	courage
23	prosper	prosper
24	well	wellness
25	pleas	pleasing
26	love	loved
27	care	care
28	mom	mom
29	inexpens	inexpensive

Index	Word Root	Word Example
30	intellig	intelligent
31	beauti	beauty
32	energet	energetic
33	glad	glad
34	harmoni	harmony
35	passion	passion
36	independ	independence
37	bacon	bacon
38	geniu	genius
39	warm	warm
40	bless	blessed
41	grate	grateful
42	educ	educate
43	respect	respectful
44	extraordinari	extraordinary
45	hope	hopeful
46	fresh	freshness
47	help	helpful
48	gener	generous
49	accur	accurate
50	sexi	sexy
51	inspir	inspire
52	wonder	wonderful
53	adventur	adventure
54	optimum	optimum
55	dedic	dedication
56	brave	brave
57	stimul	stimulate
58	fascin	fascinating
59	product	productive
60	clever	clever
61	fortun	fortunate
62	fulfil	fulfill
63	congratul	congratulations
64	champion	champion
65	fashion	fashionable

Index	Word Root	Word Example
66	uniqu	unique
67	reliabl	reliable
68	ambit	ambition
69	gift	gift
70	clariti	clarity
71	thought	thoughtful
72	relax	relaxed
73	reliev	relieve
74	luxuri	luxury
75	admir	admired
76	famili	family
77	movi	movie
78	optim	optimism
79	imagin	imagine
80	discov	discover
81	loyalti	loyalty
82	remark	remarkable
83	simpl	simple
84	encourag	encourage
85	conveni	convenient
86	safeti	safety
87	nurtur	nurture
88	appreci	appreciative
89	upbeat	upbeat
90	terrif	terrific
91	applaus	applause
92	fruit	fruitful
93	phenomen	phenomenal
94	partner	partner
95	rose	rose
96	upgrad	upgrade
97	teach	teach
98	simplic	simplicity
99	mentor	mentor
100	desir	desire
101	invit	invitation

Index	Word Root	Word Example
102	incent	incentive
103	spirit	spirit
104	superb	superb
105	donat	donation
106	earn	earn
107	eleg	elegant
108	embrac	embrace
109	storytel	storytelling
110	abil	ability
111	intimaci	intimacy
112	commend	commend
113	proud	proud
114	jewelri	jewelry
115	support	supportive
116	credibl	credible
117	kudo	kudos
118	profit	profit
119	promis	promising
120	magnific	magnificent
121	nice	nice
122	neat	neat
123	son	son
124	custom	customized
125	authent	authentic
126	replenish	replenish
127	flexibl	flexibility
128	surviv	survive
129	effect	effectiveness
130	father	father
131	legendari	legendary
132	potenti	potential
133	sincer	sincere
134	qualiti	quality
135	chariti	charity
136	mileston	milestone
137	creation	creation

Index	Word Root	Word Example
138	festiv	festival
139	import	important
140	cool	cool
141	thrive	thrive
142	friend	friend
143	vision	vision
144	interest	interesting
145	engag	engaged
146	instrument	instrumental
147	insight	insight
148	agreement	agreement
149	organ	organize
150	soft	softness
151	expert	expert
152	championship	championship
153	keen	keen
154	sophist	sophistication
155	confid	confidence
156	wife	wife
157	worthwhil	worthwhile
158	complement	complement
159	applaud	applaud
160	steadfast	steadfast
161	adopt	adoption
162	gratitud	gratitude
163	steadi	steady
164	fragranc	fragrance
165	gratifi	gratifying
166	profound	profound
167	readi	ready
168	capabl	capability
169	relief	relief
170	cooper	cooperative
171	remedi	remedy
172	emot	emotion
173	connect	connect

Index	**Word Root**	**Word Example**
174	visionari	visionary
175	signific	significant
176	accept	acceptable
177	contribut	contribution
178	produc	produce
179	soften	soften
180	agil	agility
181	promot	promote
182	reassur	reassure
183	articul	articulate
184	possibl	possibility
185	intrigu	intriguing
186	open	openness
187	club	club
188	drive	drive
189	pride	pride
190	partnership	partnership
191	score	score
192	kid	kids
193	perform	perform
194	activ	active
195	reput	reputable
196	graduat	graduation
197	power	powerful
198	awar	awareness
199	realiz	realize
200	tribut	tribute
201	bubbl	bubble
202	footwear	footwear
203	acknowledg	acknowledge
204	consist	consistent
205	attent	attentive
206	recoveri	recovery
207	elev	elevate
208	recov	recover
209	savvi	savvy

Index	Word Root	Word Example
210	expans	expansion
211	broaden	broaden
212	certainti	certainty
213	rebuild	rebuild
214	attract	attract
215	reviv	revival
216	eager	eager
217	wealthi	wealthy
218	orderli	orderly
219	lean	lean
220	accommod	accommodate
221	simplifi	simplify
222	sensit	sensitive
223	holist	holistic
224	build	build
225	classic	classic
226	recogniz	recognizable
227	vital	vitality
228	solv	solve
229	showcas	showcase
230	maintain	maintain
231	grow	grow
232	cultur	culture
233	commun	communicate
234	fellow	fellow
235	decent	decent
236	instruct	instructive
237	essenti	essential
238	prefer	preferable
239	teamwork	teamwork
240	retain	retain
241	nimbl	nimble
242	allevi	alleviate
243	seek	seek
244	motiv	motivation
245	stun	stunning

Index	Word Root	Word Example
246	magazin	magazine
247	leader	leader
248	astut	astute
249	revolutionari	revolutionary
250	athlet	athlete
251	bridal	bridal
252	maxim	maximize
253	beneficiari	beneficiary
254	acceler	acceleration
255	familiar	familiar
256	brother	brother
257	guarante	guarantee
258	profession	professionalism
259	invest	invest
260	disciplin	disciplined
261	compet	competence
262	charact	character
263	adept	adept
264	advanc	advanced
265	divers	diversity
266	transform	transformation
267	time	timely
268	substanti	substantial
269	entrepreneur	entrepreneur
270	straightforward	straightforward
271	comeback	comeback
272	reestablish	reestablish
273	person	personality
274	fix	fix
275	address	address

A.2.2 Extra Negative Words

This section will displays all the extra negative words in the new extended dictionary. All words in this section does not include the words in the current LM dictionary.

Index	**Word Root**	**Word Example**
1	terror	terrorism
2	diseas	disease
3	nausea	nausea
4	viru	virus
5	kill	kill
6	unhappi	unhappy
7	death	death
8	infect	infection
9	pollut	pollution
10	fatal	fatal
11	victim	victim
12	epidem	epidemic
13	diarrhea	diarrhea
14	terribl	terrible
15	sad	sad
16	steal	steal
17	breakup	breakup
18	asbesto	asbestos
19	afraid	afraid
20	sick	sick
21	obes	obese
22	tsunami	tsunami
23	horribl	horrible
24	terrorist	terrorist
25	toxic	toxic
26	anxieti	anxiety
27	mean	mean
28	depress	depression
29	asthma	asthma
30	constip	constipation
31	ugli	ugly
32	outbreak	outbreak
33	hangov	hangover
34	hell	hell
35	pain	painful

Index	Word Root	Word Example
36	theft	theft
37	wast	waste
38	fake	fake
39	pneumonia	pneumonia
40	handicap	handicap
41	disabl	disabled
42	uncomfort	uncomfortable
43	suppress	suppress
44	emerg	emergency
45	trash	trash
46	neg	negativity
47	inflamm	inflammation
48	flood	flood
49	scare	scared
50	overwhelm	overwhelmed
51	audit	audit
52	cathet	catheter
53	stupid	stupid
54	agit	agitation
55	chao	chaos
56	hypertens	hypertension
57	lawsuit	lawsuit
58	monsoon	monsoon
59	missil	missile
60	dementia	dementia
61	fee	fee
62	gambl	gambling
63	crash	crash
64	orphan	orphan
65	relaps	relapse
66	inflat	inflation
67	blame	blame
68	meaningless	meaningless
69	awkward	awkward
70	combat	combat
71	nasti	nasty

Index	**Word Root**	**Word Example**
72	creep	creep
73	lesion	lesion
74	cigarett	cigarette
75	debilit	debilitating
76	struggl	struggle
77	scari	scary
78	foolish	foolish
79	messi	messy
80	allerg	allergic
81	dire	dire
82	earthquak	earthquake
83	nicotin	nicotine
84	meltdown	meltdown
85	anemia	anemia
86	seizur	seizure
87	brutal	brutal
88	aggress	aggressive
89	contamin	contamination
90	rumor	rumor
91	warfar	warfare
92	overweight	overweight
93	lousi	lousy
94	tear	tear
95	expir	expire
96	hurrican	hurricane
97	explos	explosion
98	desper	desperate
99	lockup	lockup
100	old	old
101	deadlin	deadline
102	crowd	crowded
103	noisi	noisy
104	bureaucraci	bureaucracy
105	congest	congestion
106	fibrosi	fibrosis
107	worri	worried

Index	Word Root	Word Example
108	uncertainti	uncertainty
109	bacteria	bacteria
110	radiat	radiation
111	inject	inject
112	casualti	casualty
113	symptom	symptom
114	dropout	dropout
115	irrelev	irrelevant
116	expens	expensive
117	nois	noise
118	tobacco	tobacco
119	scarc	scarce
120	plaqu	plaque
121	ban	ban
122	regret	regret
123	mortgag	mortgage
124	oncologist	oncologist
125	liabil	liability
126	anem	anemic
127	bug	bug
128	regress	regression
129	hesit	hesitant
130	restrict	restrict
131	irrat	irrational
132	addict	addiction
133	leakag	leakage
134	alarm	alarming
135	court	court
136	dent	dent
137	fight	fight
138	legisl	legislation
139	legislatur	legislature
140	nervou	nervous
141	pressur	pressure
142	skeptic	skeptical
143	subdu	subdued

Index	Word Root	Word Example
144	tension	tension
145	oversight	oversight
146	uncertain	uncertain
147	steroid	steroid
148	mute	muted
149	tornado	tornado
150	insignific	insignificant
151	economi	economy
152	strict	strict
153	gout	gout
154	ridicul	ridiculous
155	unclear	unclear
156	bacteri	bacterial
157	takeov	takeover
158	depend	dependency
159	blackout	blackout
160	dump	dump
161	surrend	surrender
162	cardiac	cardiac
163	forget	forget
164	sellout	sellout
165	resist	resistance
166	lumpi	lumpy
167	disord	disorder
168	empti	empty
169	rundown	rundown
170	pend	pending
171	drain	drain
172	anxiou	anxious
173	offens	offensive
174	neglig	negligible
175	unlik	unlikely
176	hefti	hefty
177	handout	handout
178	extract	extraction
179	redund	redundancy

Index	Word Root	Word Example
180	elimin	elimination
181	chronic	chronic
182	edema	edema
183	nonexist	nonexistent
184	sever	severance
185	drainag	drainage
186	judg	judge
187	temper	temper
188	shock	shock
189	undergo	undergo
190	pessimist	pessimistic
191	execut	execute
192	issu	issue
193	hit	hit
194	residu	residue
195	slump	slump
196	overlap	overlap
197	dissect	dissect
198	symptomat	symptomatic
199	liabl	liable
200	polit	politics
201	leak	leak
202	differenti	differential
203	fabric	fabrication
204	judici	judicial
205	presum	presume
206	compress	compressed
207	friction	friction
208	shrink	shrink
209	hook	hook
210	immigr	immigration
211	unsolicit	unsolicited
212	dispos	disposal
213	reloc	relocate
214	rush	rush
215	departur	departure

Index	Word Root	Word Example
216	remedi	remedial
217	lull	lull
218	repetit	repetitive
219	merger	merger
220	arbitr	arbitration
221	absent	absent
222	chunki	chunky
223	withdraw	withdraw
224	alcohol	alcohol
225	trivial	trivial
226	collater	collateral
227	drug	drug
228	outpati	outpatient
229	mutat	mutation
230	intervent	intervention
231	unfamiliar	unfamiliar
232	arbitrari	arbitrary
233	regul	regulation
234	decreas	decrease
235	turbul	turbulent
236	teas	tease
237	trough	trough
238	stroke	stroke
239	solicit	solicitation
240	undertak	undertake
241	complex	complex
242	rule	ruling
243	spike	spike
244	juri	jury
245	drop	drop
246	exhaust	exhaustive
247	inact	inactive
248	transmit	transmit
249	dilemma	dilemma
250	compli	comply
251	intens	intensive

Index	Word Root	Word Example
252	republican	republican
253	preemptiv	preemptive
254	remov	removal
255	supervis	supervision
256	collagen	collagen
257	indiffer	indifferent
258	redirect	redirect
259	thyroid	thyroid
260	remiss	remiss
261	vagu	vague
262	spotti	spotty
263	prohibit	prohibit
264	congression	congressional
265	tire	tired
266	intraven	intravenous
267	substanc	substance
268	assumpt	assumption
269	bumpi	bumpy

Appendix B

Additional Results for Attention-based Multiple Instance Learning

B.1 The top 100 words ranked by the att-MIL models

Rank	att-MIL	gated-att-MIL
1	forgettable	unwatchable
2	unfunny	forgettable
3	unwatchable	unfunny
4	insipid	4/10
5	4/10	uninspired
6	uninspired	disappointing
7	disappointing	tedious
8	lackluster	lackluster
9	1/10	worst
10	tedious	underwhelming
11	awful	2/10
12	mediocre	waste
13	uninteresting	unoriginal
14	5/10	1/10
15	lousy	poorly

Rank	**att-MIL**	**gated-att-MIL**
16	poorly	uninspiring
17	uninspiring	insipid
18	worst	lousy
19	underwhelming	disappointment
20	disjointed	amateurish
21	dreadful	mildly
22	unappealing	monotonous
23	unimaginative	unimaginative
24	disappointment	dull
25	bland	unremarkable
26	flimsy	ineffective
27	pointless	flimsy
28	monotonous	unappealing
29	atrocious	abysmal
30	laughable	pointless
31	whiny	mediocre
32	amateurish	awful
33	incoherent	bland
34	unconvincing	uninteresting
35	lacklustre	atrocious
36	lifeless	disjointed
37	dull	dreadful
38	0/10	appalling
39	unoriginal	unconvincing
40	unremarkable	wasted
41	overlong	horrid
42	horrid	laughable
43	horrible	tiresome
44	overpriced	shoddy
45	unexciting	pathetic
46	godawful	unpleasant
47	pathetic	unimpressive
48	drivel	dreary
49	uncreative	embarrassment
50	tiresome	muddled
51	trite	woeful

Rank	**att-MIL**	**gated-att-MIL**
52	appallingly	sluggish
53	miscast	drivel
54	woeful	5/10
55	waste	embarrassing
56	overcooked	worthless
57	unmemorable	lifeless
58	embarrassing	wretched
59	mess	vacuous
60	ineffective	unmemorable
61	mildly	overlong
62	abysmal	unexciting
63	listless	wasting
64	appalling	pitiful
65	unimpressive	lukewarm
66	anemic	dismal
67	subpar	lacks
68	abysmally	tasteless
69	woefully	tripe
70	wretched	woefully
71	badly	inane
72	incompetent	horrible
73	cloying	uneventful
74	overused	lethargic
75	sucky	mess
76	pitiful	0/10
77	craptastic	uninvolving
78	clunky	embarrassingly
79	tripe	terrible
80	shoddy	downright
81	uneventful	dreck
82	terrible	redeeming
83	yawn	incoherent
84	ludicrous	unprofessional
85	inane	godawful
86	interminable	badly
87	miserably	nauseating

Rank	**att-MIL**	**gated-att-MIL**
88	inadequate	lacklustre
89	irritating	lamest
90	wasting	substandard
91	worse	trite
92	embarrassment	clunky
93	risible	irritating
94	talentless	inept
95	unlikeable	useless
96	dreary	incompetent
97	dreck	worse
98	inept	boring
99	slipshod	appallingly
100	gimmicky	miscast