Read Like A Human: The Role of Quantified Knowledge in Market Predictions

Tiancheng Wang

A thesis presented for the degree of Master of Research
Supervisors: Dr Jing Shi and Dr Terry Pan

Department of Applied Finance
Macquarie Business School
Macquarie University

18 May 2020

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my supervisors, Dr Jing Shi and Dr Terry Pan, for their exceptional support, expertise and patience during the undertaking of this thesis.

I would also like to thank Macquarie University for their financial and IT support.

Professional editors, Elite Editing, provided copyediting and proofreading services in accordance with, and restricted to Standards D and E of the Australian Standards for Editing Practice.

Finally I would like to thank my girlfriend, for every meal she cooked in the long nights when I was overwhelmed by codes and bugs, as well as my parents, for their support during the hard time.

Abstract

In this study, I propose a novel entity-specific sentiment index to examine how massive general knowledge can be quantified and used to extract better financial inferences from media outlets following a similar reasoning process as human news readers. With the advent of graph representation techniques, external knowledge can be represented by knowledge graphs, and then quantified through graph embedding processes. By modifying traditional sentiment analysis using quantified knowledge, I find that the introduction of external knowledge significantly and consistently improves the predictive power of sentiment indexes as indicators of stock market activity.

Keywords: Media outlets; Market activity predictions; Knowledge Graphs; Sentiment Analysis; Natural Language Processing; Neural Networks

Chapter 1: Introduction

In the last three decades a significant amount of research investigating the impact of media outlets on the financial market has been published. Researchers held diverging views on the effects that media could cast on the stock market. As modelled by De Long, Shleifer, Summers, and Waldmann (1990), high media sentiment might lead to excessively optimistic estimation of the future prospective of the market, resulting noise investors overvaluing the stocks. An example for this theory is the relationship between speculative bubbles and media publication. According to Shiller (2000), speculative bubbles started to be documented roughly alongside with the advent of newspapers. In *Irrational Exuberance* (2000) he states that news publishers tend to act as demagogues that exaggerate the effects of past events. Recent good performance of the financial market could motivate the publishers emitting overly optimistic sentiment through their publication, which reshapes the public sentiment into a surge of excessive exuberance that further elevates the stock price from its fundamental value, snowballing into a even higher future media sentiment. The 'Amplification Mechanism', as named by Shiller (2000), is considered as a natural cause in Ponzi Processes. His work provides evidence for the earlier establishments from Merton (1987), who claims that the media exposure could play an important role in the progression of speculative bubbles.

In this thesis I propose a novel knowledge-aware entity-specific media sentiment index. Following the same definition in machine learning literature, *Entity* in this thesis refer to all independent existence in the real world (e.g. Martin_Luther_King_jr, ExxonMobil and Ipad etc.). *Knowledge*, on the other hand, means the relationship between entities (e.g. Bill_Gates → Founder → Microsoft). By incorporating Knowledge Graph techniques into the decision process, the proposed index is able to extract media information that is most relevant to our interested entity, which is the Dow Jones Industrial Average (DJIA) in most parts of this thesis.

This thesis contributes to the existing literature in the following ways. First, as to my knowledge, this report is one of the earliest attempts examining how external knowledge could be incorporated in asset pricing literature. Second, comparing to traditional sentiment analysis, the proposed method could filter out irrelevant sentiment expressions by assigning minimum attention to the sentiment scores when the news articles are less relevant, resulting a more resilient model against noise. Third, the spillover effects are retained. By factoring in the cross-entity relations, it is possible for the index to capture the nuances in news articles that have no direct mention of the research targets. Last but not least, for every topic of interest, the proposed method provides an all-in-one entity-specific index, resulting less exposure to potential issues caused by multicollinearity, and better cooperation with other predictors.

With the advent of computational textual analysis, a vast literature is able to examine the relation between media outlets and the financial market on a much larger scale (Tetlock, 2007; Tetlock, Saar-Tsechansky and Macskassy, 2008; Loughran and

McDonald, 2011; Garcia, 2012). Most of the works find evidence supporting the existence of media effects on the financial market. These findings challenge the classic Efficient Market Hypothesis (Fama, 1970), which believes that fresh information should spread over the market in a timely fashion, resulting in short-lived arbitrage opportunities.

A number of popular textual analysis algorithms extract and quantify the semantic meaning of texts in a quasi-humanly fashion, by counting sentiment keywords (Loughran and McDonald, 2011), or spotting semantic topics (Blei, Ng, and Jordan, 2003). Generally speaking, news content is highly condensed, and full of named entities (such as Barack_Obama and Apple_Inc). In order to draw financial inferences from news articles, human readers will first extract the entities they spot in the context, and then interpret the news article based on their personal knowledge about the entities as well as the contexts. The quality of interpretation is affected by both the quality of the news article, and the knowledge base of the reader.

In 2011, Griffin, Hirschey and Kelly proposed a number of hypotheses about media influence over the financial market around the world, including hypothesizing the quality of media outlets as a potential factor affecting the financial forecastability of news articles. They propose an index measuring how relevant the news articles are to the firms' value-related events, as a proxy for the quality of the media. In the experiments they find that news articles from countries with higher media quality have significantly higher financial forecastability comparing to their low-quality counterparts. In typical sentiment analysis, the data are sourced from large sets of news articles that cover a wide range of topics, which are not always relevant to the focus of the research. In the context of the finance literature, a news article titled "Wall Street Asks When, Not if, the Fed Will Cut Interest Rates"[1] is most likely more relevant than one titled "Nazis Killed Her Father. Then She Fell in Love With One,"[2] although both article are archived in the "Business" section of *The New York Times*. News articles with lower quality might distract news consumers with overwhelming amount of irrelevant information. An intuitive inference from the findings from Griffin et al. (2011) is that it is optimal to pay more attention to media sentiments when the news is more relevant to the research topic of interest. Being able to identify the topics in news articles, human readers can decide how relevant the articles are. However, without being integrated with filter mechanics, traditional sentiment models might produce questionable indexes that poorly reflect market activities, as I discuss in Subsection 4.3.

To remedy this issue, Van de Kauter, Breesch and Hoste (2015) propose a company-specific sentiment analysis based on newswires. During the data pre-processing all the news articles fail to mention at least one of their target companies (KBC, Delhaize, AB InBev and Belgacom) will be filtered out from the data set. Remaining news articles will be assigned to different groups based on their mentions of corresponding firms.

[1] Source: *The New York Times*, https://www.nytimes.com/2019/06/09/business/wall-street-federal-reserve-interest-rates.html?searchResultPosition=86.

[2] Source: *The New York Times*, https://www.nytimes.com/2019/06/14/business/reimann-jab-nazi-keurig-krispy-kreme.html?searchResultPosition=41.

Company-specific sentiment indices are calculated based on the filtered data sets, and compared against benchmark sentiment analysis without filtering mechanics. The experiment results show that the filter-based sentiment index significantly outperforms its baselines. By filtering out irrelevant sentiment expressions and detecting explicit and implicit sentiment, they manage to yield stronger predictive power from newswires.

Sharing similarities to their work, a number of studies incorporate topic-based filter mechanics to improve the quality of the data set (Fang and Press, 2009). Despite the encouraging improvements in the prediction quality, the black-or-white filter mechanics might introduce the issue of overfiltering. For instance, an article titled "Taiwanese chip maker plans to raise prices"[3] reports a pending increase in chip prices from the Taiwan Semiconductor Manufacturing Co. (TSMC), the dominant manufacturer in the semiconductor market and major supplier of tech giants such as Intel and Apple. In most scenarios, this article will be overlooked by topic models since it does not focus on typical targets in the finance literature, such as the Dow Jones Industrial Average (DJIA) or its components. However, this news article might actually render significant effects on DJIA components such as AAPL in at least two possible ways. First, according to the Category-Based Comovement Theory (Barberis, Shleifer and Wurgler, 2005), parts of noise investors tend to group securities based on their natural characteristics (e.g. same industry) before making investment decisions. When

[3] Source: *The New York Times*, https://www.nytimes.com/2008/05/27/business/worldbusiness/27iht-chip.1.13234862.html?searchResultPosition=5.

experiencing a shock in the investor sentiment, the category-based investors will channel in(out) funds to(from) ***all*** the stocks within the category. In our example, if TSMC as a manufacturer of semiconductor is considered by investors to be part of the IT industry, the category-based investors might short TSMC as well as AAPL since they pick up the structural profit pressure TSMC are facing after reading the actual news. Second, human readers with sufficient knowledge tend to pay more attention to this news article since they realize the relationship between TSMC and AAPL. They might consider the future price increase from a major supplier as a negative signal for the prospective of AAPL, resulting short position in AAPL. In conclusion, the negative sentiment from the example news might have spillover effects on other firms that related to TSMC, although not being mentioned in the actual context.

Ideally, an accurate interpretation of the news articles should be aware of both the entity information in the news articles and the relationship between the entities as well. Being equipped with proper knowledge, experienced human readers could discriminate the relevance of news articles on a deeper knowledge level, rather than a semantic level where most traditional textual analysis techniques are trapped in. Motivated by the lack of research on the use of external knowledge in asset pricing literature, I attempt to improve the quality of financial interpretation from media outlets by capturing the inter-entity relationship with Knowledge Graph techniques.

With the advent of Knowledge Graph techniques, it is possible to incorporate this structural information into the sentiment indexes. As shown in Figure 2, a typical knowledge graph is a directional graph whose nodes denote real world entities and the edges denote different relationships. Intrinsically, a knowledge graph is a semantic network that captures the relations between knowledge entities (Xu et al., 2016). In its simplest form, a knowledge graph can be reduced to a series of Ontologies, triplets consisting of Class, Attribute and Relationship, as shown in Figure 1. Billions of ontologies are stored inside databases dedicated to structured data. The application of knowledge representation generally involves downloading a subgraph from one of these databases. In this thesis, I adopt DBpedia, one of the big four open projects in knowledge representation (Freebase, Wikidata, DBpedia, and YAGO), as my source of external knowledge. To incorporate the external information into the model, the knowledge will go through an embedding process. The idea behind embedding is to transform information that is not in numeric form into low-dimensional vector representations in continuous real space. In addition, the vector representation should carry all the information stored in the original form. A detailed discussion about the embedding methods is presented in Chapter 3.

Figure 1. An Example of Ontology

Figure 1 shows a typical ontology triplet containing two entities and one relationship. The direction of the relation should also be captured by the embedding process.

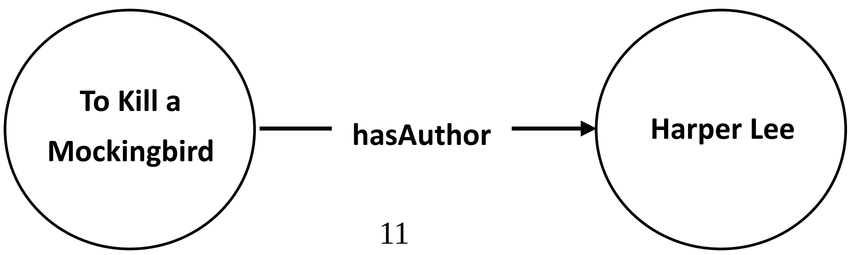
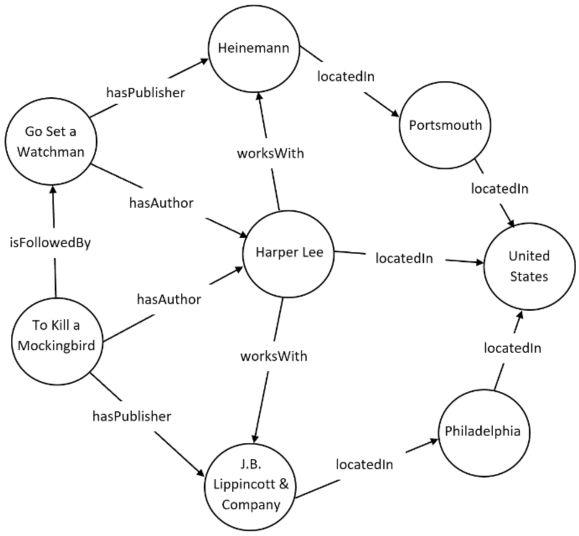


Figure 2. An Example of a Knowledge Graph

Figure 2 shows a typical knowledge graph capturing the entities (denoted by nodes) from various topics. The nodes are linked by arrows, which show their relationships.



Research on feeding standardised knowledge to intelligent systems dates back to the 1980s (Russell and Norvig, 1995). The rapid generation and accumulation of data has led to the formation and widespread use of ontologies. The concept of representing knowledge in graphs has been popularised by the inception of Google's Knowledge Graph in 2012. Word Embedding algorithms (Mikolov et al., 2013) and deep neural networks in machine learning are considered key breakthroughs in knowledge representations, which make quantifying general knowledge significantly easier and more efficient. Although knowledge graphs have been widely adopted for tasks such as news recommendation systems (Wang et al., 2018) and fraud detection (Zhang and Yin, 2018), its application in asset pricing is still largely unknown.

In this thesis, I propose a novel index capturing the entity-specific media sentiment that is aware of external knowledge. The indices are calculated by synthesizing information from more than 200,000 news articles that appeared between January 2007 and December 2016 and archived by the *NYT*. Both the sentiment tone and the topic mentions of news articles are extracted using two separate machine learning process. The monthly media sentiment is then modified based on the relations between the topics mentioned in the articles and my target topic, which is Dow Jones Industrial Average (DJIA). More attention will be paid to the media sentiment when the news articles are more relevant to DJIA. A knowledge graph (KG) that contains more than 2,400,000 entities and 6,600,000 triplets is created as an external knowledge source to mimic the human judgement process. Entities and cross-entity relationships in the knowledge graph are embedded into vector representations using a deep learning process, namely TransD (Ji et al., 2016). The continuous low-dimensional vector representation paves the way for the quantification of the relationships between entities, which is an essential part of the entity-specific sentiment index.

In this thesis the result from sentiment analysis following a similar methodology as Tetlock (2007) is used as one of the primary benchmarks, namely '*Sent_news*'. The proposed index (*Sent_DJIA*) is essentially *Sent_news* modified by the KG. The modification process simply goes by the name **KG modification**. By comparing the *Sent_DJIA* against *Sent_news*, I find evidence for the improvements brought by the inclusion of external knowledge. The most noticeable effect is that KG modification shows significant correction power, resulting in a sizeable increase in predictive power. In the in-sample test I conduct, the KG modified sentiment index produces an R-squared of 7.3%, which is 55.3% higher than the original one, which has an R-squared of 4.7%. More intriguingly, the KG modified sentiment index consistently outcompetes both the original index and the historical average, regardless of the time window. On average, the KG modified sentiment index obtains an out-of-sample R-squared of 2.7%, while the original one obtains an average R-squared of -4.0%. A selection of widely accepted macro indicators (Welch and Goyal, 2008) and sentiment indexes (Baker and Wurgler, 2006) are tested as benchmarks. The test results indicate that the KG modified sentiment index has the strongest predictive power among all the predictors. Furthermore, the results are robust to different parameter settings and regression setups. *Sent_DJIA* also shows key characteristics of typical sentiment indices (Brown and Cliff, 2005; Tetlock, 2007; Tetlock, 2011), indicating that the index is an valid measurement for media sentiment.

The remainder of this thesis is structured as follow. Previous literature covering the application of textual analysis in finance and economics is presented in Chapter 2. In Chapter 3 the data source adopted in this thesis is acknowledged. The overall model design is first illustrated in the beginning part of Chapter 4, then be divided into four major components and thoroughly discussed in each individual subsection. Since the methodology proposed in this thesis is reasonably complex, it is crucial to choose the most appropriate approach for every step, and a series of examinations is conducted in

Chapter 4. Chapter 5 focuses mainly on the in-sample regression tests. The in-sample experiments provide us with a direct observation of the general effects of the external knowledge on sentiment indexes. The findings are further extended to and consolidated by the out-of-sample regression tests presented in Chapter 6. In Chapter 7, I conclude the findings in previous sections, and briefly introduce the limitation of this framework, as well as the future potential of this topic.

Chapter 2: Literature Review

In recent two decades an increasing number of studies incorporating textual analysis has been witnessed in finance and economics literature. Based on the adopted methods two strands of research relate the most to this thesis.

2.1 Sentiment Analysis

One strand of research examines how media sentiment influences financial markets. As one of the pioneers, Tetlock (2007) applies the Harvard IV-4 (HIV-4) psychosocial dictionary to extract the sentiment tone of the *Wall Street Journal*'s "Abreast of the Market" column on the Dow Jones Industrial Average (DJIA). He finds that high levels of media pessimism indicate downward pressure on market prices. In addition, unusual fluctuations of media pessimism might temporarily increase market trading. In contrast, Loughran and McDonald (2011) claim that almost three-fourths of the negative words in the HIV-4 dictionary do not typically have negative meanings in financial contexts. They suggest that utilizing the HIV-4 dictionary for finance research might lead to inaccurate conclusions. Instead, they propose a tailor-made dictionary that is fine-tuned for financial contexts, namely the Loughran-McDonald (LM) dictionary. Building on the findings, Garcia (2012) applies the LM dictionary to news articles and examines how media sentiment affects stock returns. She finds that media sentiments not only could help predicting daily stock returns. the predictive power of media sentiment is also more prominent during recessions. Furthermore, this effect is especially significant on Mondays and on days after holidays. Outside of empirical asset pricing, Liu and McConnell (2013) find that the level of media attention and the sentiment tone have significant effects on managers' decisions to abandon value-reducing acquisition attempts.

2.2 Topical Analysis

In another strand of research, topical analysis is included in the methodology. Outside the finance literature, Baker, Bloom, and Davis (2016) develop a new index to measure economic policy uncertainty (EPU). They manually scan articles in the media and count the number of articles that contain a trio of predefined keywords pertaining to uncertainty, the economy, and policy. The media-based index is widely accepted (Pástor and Veronesi, 2013; Bachmann, Elstner and Sims, 2013) as indicator of EPU, and they show that it is negatively related to stock returns. Recently, machine learning methods have begun to be used for asset pricing. Calomiris and Mamaysky (2019) use text mining techniques on news articles to examine their various effects on the stock market's risk and return in 51 countries from 1996 to 2015. They extract the textual topics using the Louvain method (Blondel et al., 2008) and construct topic-specific sentiment indexes. The similarity between the rationale behind the topic-specific sentiment indexes and the EPU approach enables them to use the EPU index as a benchmark. They suggest that the algorithmic approach has greater explanatory power than the EPU index when predicting stock movements. Another work that is closely related to mine is Fang and Press (2009). They develop an index based on the relevance score to examine the relation between stock returns and media coverage. LexisNexis provides relevance scores to capture how likely a news article is related to a particular topic. The authors count the number of news articles with a relevance score of 90% or above to obtain a firm-specific time series of interest. They find that media exposure is negatively related to stock returns in the United States. Firms with low media coverage significantly outperform those frequently featured in the media by over 0.20% per month on average.

Chapter 3: Data Source

Every news article appearing in the "Business" and "U.S." sections of the *New York Times* (*NYT*) between January 2007 and December 2016 is collected as the primary source of information. Theoretically, external knowledge should help redirect user attention to more relevant topics. One goal of the paper is to examine the effects of external knowledge. By not manually selecting news articles, unreliable effects due to subjectivity are mitigated. I downloaded more than 6,600,000 ontologies from DBpedia as source materials for the knowledge graph. I use the latest stable version, which was released in October 2016; thus, the knowledge graph characterizes the information of the world in 2016. The dynamic nature of knowledge constantly erodes the legitimacy of the knowledge graph over time. Any results from before 2007, which marks the start of Global Financial Crisis (GFC), are highly questionable. I chose the experiment period based on this consideration. Below I discuss how false information could potentially affect the predictions.

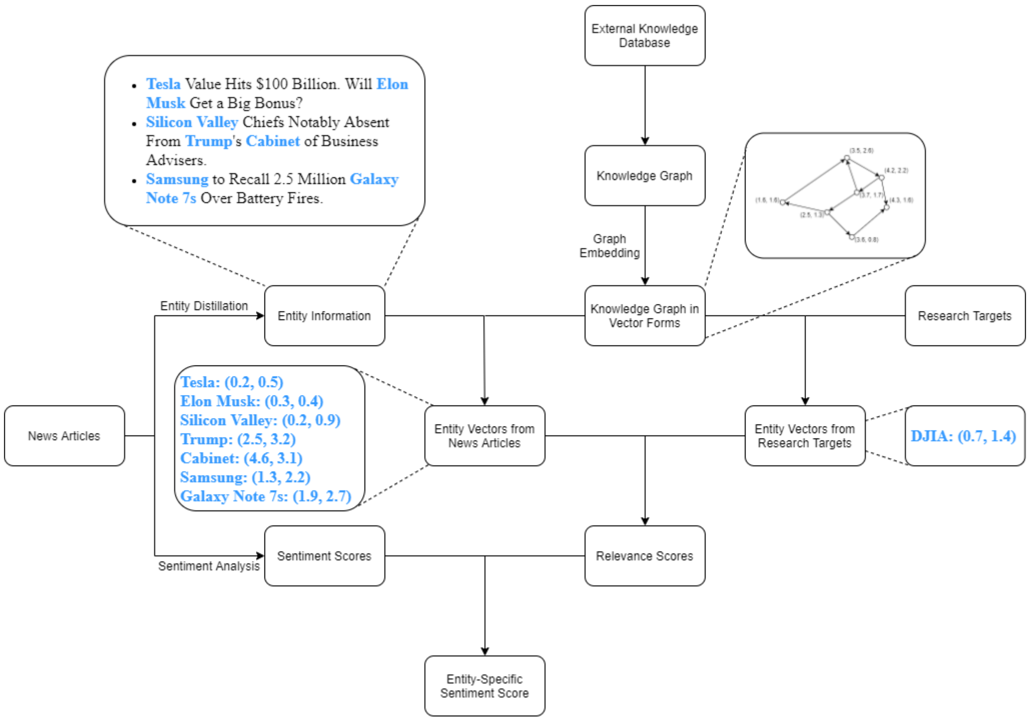
Entity information in the news articles is extracted and linked to the data entries from DBpedia by DBpedia Spotlight (Daiber et al., 2013), a Named Entity Recognition (NER) and Entity Linking (EL) (Lehmann et al., 2015) service provided by DBpedia.

Chapter 4: Methodology

In this chapter I first illustrate the overall design of the entity-specific sentiment index. Key steps are thoroughly examined and discussed in the following subsections. Similar to the human reasoning process, the entity-specific sentiment index recognizes media outlets as input information and processes the information based on external knowledge. Sentiment information is modified based on how relevant the news articles are to the topic of interest, which is DJIA. I assume that the topics of news articles are determined by the entities discussed within the articles. And both entity information and cross-entity relationships are captured by a knowledge graph. After going through an embedding process, the entity information from the media outlets and the research targets can be quantified into a low-dimensional vector representation, which enables me to characterize the relevance in numeric forms.

Figure 3. Overall Structure of the Entity-Specific Sentiment Index

Figure 3 shows the overall design of the entity-specific sentiment index. The whole index mainly consists of two processes. Sentiment scores of news articles shall first be measured through a bag-of-word (BoW) sentiment analysis, then the scores are modified by external knowledge through knowledge representation.



4.1 Entity Distillation

I follow the same process to create a deep knowledge-aware network (DKN) as in Wang et al. (2018). Before constructing the knowledge graph, DKN requires a Entity Distillation (ED) process, a combination of Named Entity Recognition (NER), Disambiguation, and Entity Linking (EL). Named entities (e.g., Apple_Inc.) are first identified and extracted from news articles. To mitigate misclassification related to polysemy (e.g., Apple and apple might refer to different entities), the identified entities go through a disambiguation process based on their surrounding textual content. Disambiguated entities are then linked to their corresponding entries in the database.

All valid ontologies subject to identified entities from the distillation process were downloaded from DBpedia. Depending on the tail entity it links to, an ontology is classified as either carrying general information (e.g., {*Barack_Obama, 1961-08-04,* *birthDate*}), knowledge information (e.g., {*Barack_Obama, Nancy_Moritz, isAppointerOf*}), or miscellaneous information (e.g., {*Barack_Obama, 534366, wikiPageID*}). Only knowledge information whose tail entity is another valid entry in the database is retained as a building block for the knowledge graph.

Table 1. Summary Statistics of the Knowledge Graph

Table 1 shows a simple summary of the basic statistics of the knowledge graph. '#' denotes 'the number of'.

# entities	# relations	# triples	avg. # entities per article	avg. # mentions per entity
2,450,536	7,867	6,608,995	10.738	2.697

4.2 Graph Embedding

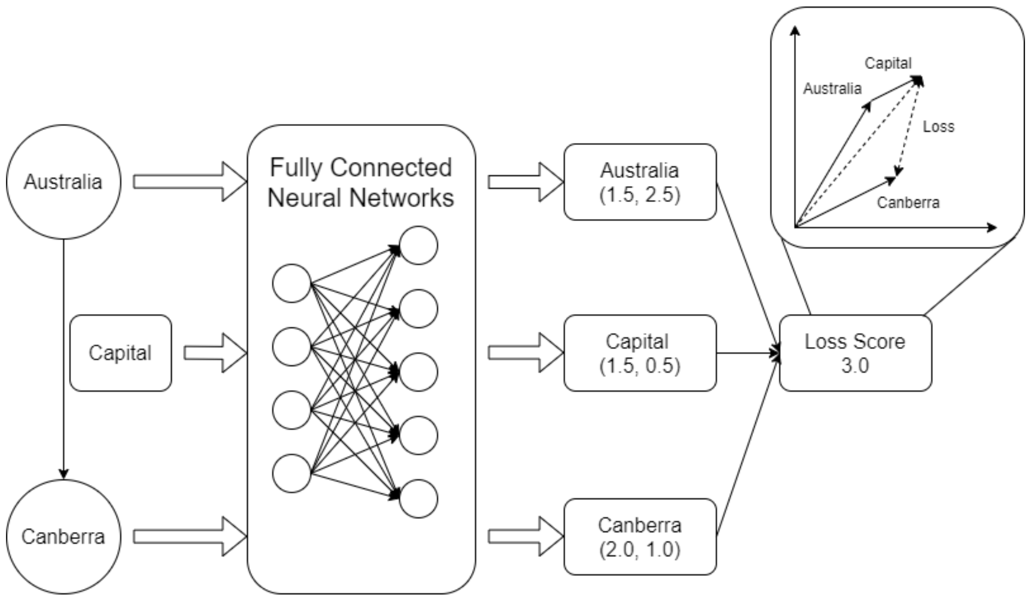
Modern methods of graph embedding are largely inspired by Word2vec by Mikolov (Mikolov et al., 2013). TransE is the most representative distance model (Bordes et al., 2013). The idea behind the TransE model is that, for every ontology triplet {head, tail, relation} assigning one vector representation $v_i \in R^d$, where $i \in \{head, tail, relation\}$to each entity and relationship in the triplet. Ideally, the vector representation will carry the information contained in the triplet, which translates to $v_{head} + v_{relation} = v_{tail}$ (e.g., $v_{Australia} + v_{Capital} = v_{Canberra}$). In practice, the full knowledge graph is first striped down to millions of ontologies as input data sets. Through a fully connected neural network, every entity and relation occurred in the input data will be assigned with a vector representation. Then the quality of the vector representation is estimated by a loss function. For TransE, the loss function is defined as

$$L(h, t, r) = max(0, d_{positive} - d_{negative} + \text{margin}) \quad (1)$$

where $d = \|v_h + v_r - v_t\|_i$ where $i \in \{1, 2\}$is the L1 or L2 norm capturing the distance, indicating how far the sum of head and relation vector is away from its tail vector. $d_{positive}$ and $d_{negative}$ are the individual distance measurements of positive/negative samplings. In positive samples the triplets are correct (e.g., {Australia, Canberra, Capital}), we expect the distance $d_{positive}$ to be as close as possible. On the contrary, in negative samples the triplets are intentionally incorrect (e.g., {Australia, Star_Wars, Capital}), we expect the distance $d_{negative}$ as far away as possible. Thus, the optimization goal is to minimize the loss function. The *margin* is a control parameter, restricting the positive/negative spread within a certain limit.

Figure 4. Basic Example of TransE Model

Figure 4 illustrates the basic concept behind the TransE model. The loss function in the example is the L1 norm defined by: $L(h,t,r) = max(0, \|v_h + v_r - v_t\|_1 - d_{negative} + \text{margin})$. For visual simplicity both $d_{negative} = 0$ and margin $= 0$.

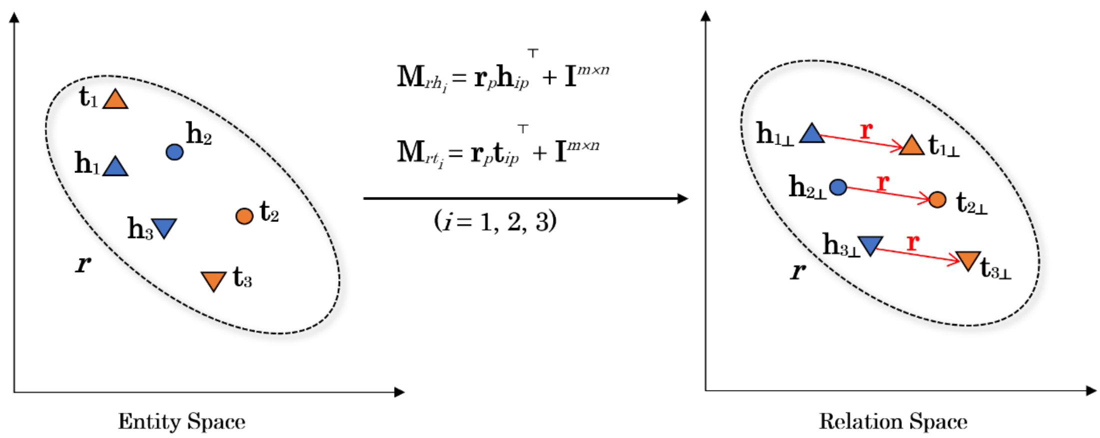


Although the concise model is widely praised as a classic, its intrinsic simplicity makes it less effective when tackling larger knowledge graphs where linguistic ambiguity prevails (Wang et al., 2014; Lin et al., 2015; Ji et al., 2016). Retaining the core idea of the TransE model, the TransD model (Ji et al., 2016) assumes that entity vectors $\{v_h, v_t\}$ and relation vector $\{v_r\}$ reside in different spaces $\{R^d_{entity}, R^d_{relation}\}$.

Mapping matrices are applied to entity vectors for projections onto $R^d_{relation}$, where projections $v_{h\perp}, v_{t\perp}$ along with v_r are optimized in a manner similar to TransE. I use TransD in this study as the embedding method since it provides me a better solution for 1-to-N, N-to-1 and N-to-N[4] relations, which can be commonly found in the large knowledge graph adopted in this thesis, while still being relatively restrained in terms of computational complexity.

Figure 5. Simple Illustration of TransD Model

Figure 4 is a simple illustration of a TransD model. Each circle and triangle represents an entity pair appearing in a triplet of relation. M_{rh} and M_{rt} are mapping matrices of h and t, respectively. h_{ip}, t_{ip}, where$i \in \{1,2,3\}$, and r_p are projection vectors. $h_{i\perp}$ and $t_{i\perp}$, where $i \in \{1,2,3\}$ are projected vectors of entities. The projected vectors satisfy $h_{i\perp} + r \approx t_{i\perp}$.



By applying TransD to the knowledge graph, every entity (node) is assigned a vector representation $v_{entity} \in R^d$, where R^d is a d-dimensional real space. Theoretically TransD, as a distance model, will quantify the distinctions between two entities by the vector distance: the longer the distance, the larger the difference between the two

[4] 1-to-N means one head entity and one relation may refers to multiple different tail entities. For instance both {Michael_Jackson, Billie_Jean, is Artist of} and {Michael_Jackson, Thriller, is Artist of} are valid triplets. The same rationale goes for N-to-1 and N-to-N.

entities. This is the key feature I exploit below.

4.3 Sentiment Analysis.

The sentiment measures involve a standard bag-of-words (BoW) procedure. A predefined dictionary associating words with sentiment scores (positive or negative) is used to determine the overall tone of the news articles based on the annotated text. Despite the burgeoning application of deep learning in sentiment analysis (Zhang et al., 2018), the performance of deep learning algorithms is highly related to specific tasks and datasets, making it less convincing when being generalized to other tasks. The BoW method is prominently used in financial research (Kearney et al., 2014) and is praised for its objectivity, large dataset compatibility, and cross-task generality (Loughran and McDonald, 2016), making it an ideal model since my study involves comparing my KG modified sentiment index against traditional unmodified sentiment indexes. The different choices of the BoW dictionary have significant impact on the results of sentiment analysis. A natural curiosity is how much extra edge, if any, a finance-specific library could provide as compared to general libraries in empirical asset pricing tasks. I examine both Loughran and McDonald (LM) (Loughran and McDonald, 2011), a tailor-made library tuned for financial documents, and Harvard GI/IV-4 (HIV-4), a collection of sub-libraries designed for general documents.

The number of news articles in each section varies each month, which affects the total sentiment score. A strand of study deems the number of news articles as an important factor in media-related topics (Fang and Press, 2009; Baker, Bloom, and Davis, 2016). During times when serious events happen or political uncertainty is high, publishers tend to increase the number of articles covering these topics. Based on similar rationale, the raw sentiment scores without factoring out the the number of news articles is used as the results of sentiment analysis. With these concerns in mind, I apply LM and HIV-4 to news articles from both 'Business' and 'U.S.' section. The sentiment score is estimated by polarity, defined by:

$$polarity = \frac{\#pos - \#neg}{\#total\ words} \tag{2}$$

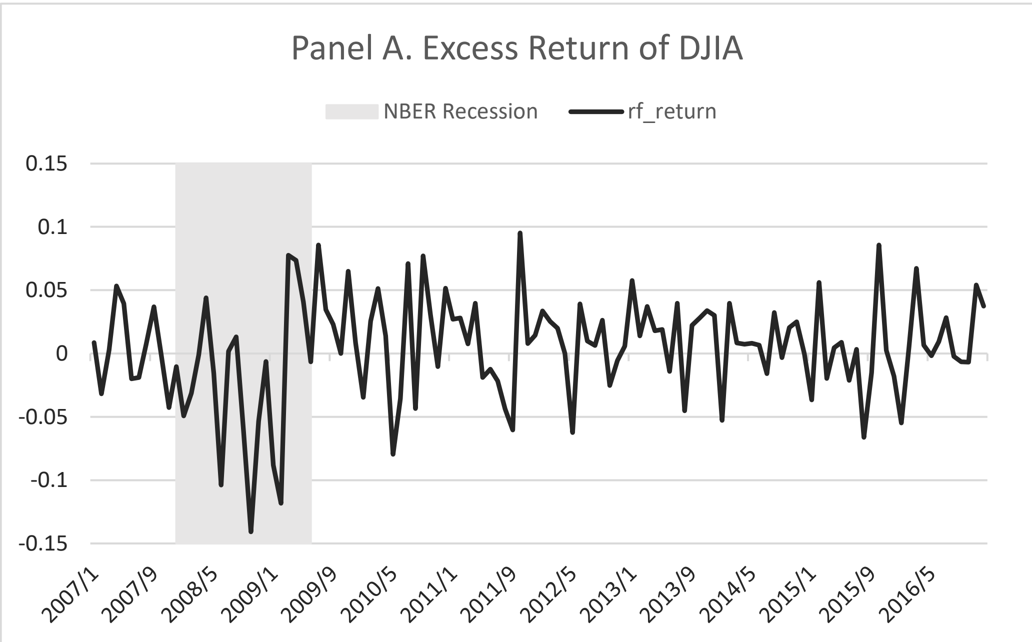
Where '#' denotes the number of.

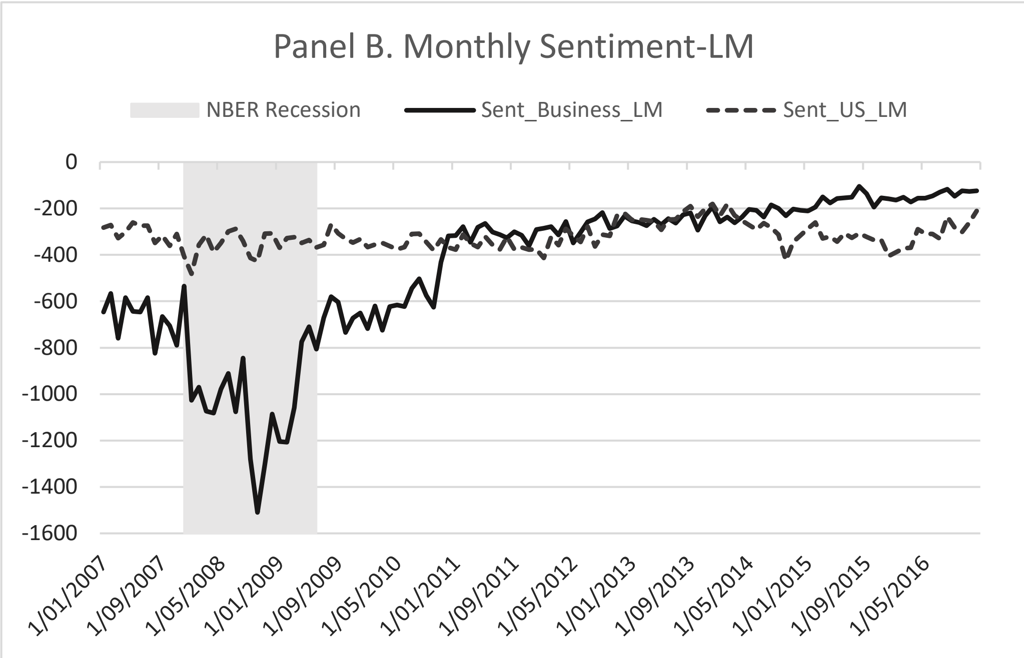
The result of sentiment analysis is presented in Figure 6. In Panel B, the sentiment index derived from the 'Business' section, namely *Sent_Business_LM* shows much higher variances than its 'U.S.' counterpart *Sent_US_LM*. It suggests that the LM library is more sensitive towards business related topics. Another noticeable fact is that *Sent_Business_LM* shows higher correlation with the DJIA return series *ret_DJIA*. This correlation is particularly prominent during NBER recession periods, where a deep valley is observed around the full swing of Global Financial Crisis (GFC). In contrast, the explanatory power of *Sent_US_LM* is highly questionable. Throughout the sample period *Sent_US_LM* remains relatively stable and no significant correlation between *Sent_US_LM* and *ret_DJIA* is observed. This intuitive result hints that the choice of media source has great impact on the quality of the index. The noisy news articles from the ‘U.S.’ section is inferior to the ones from the ‘Business’ section.

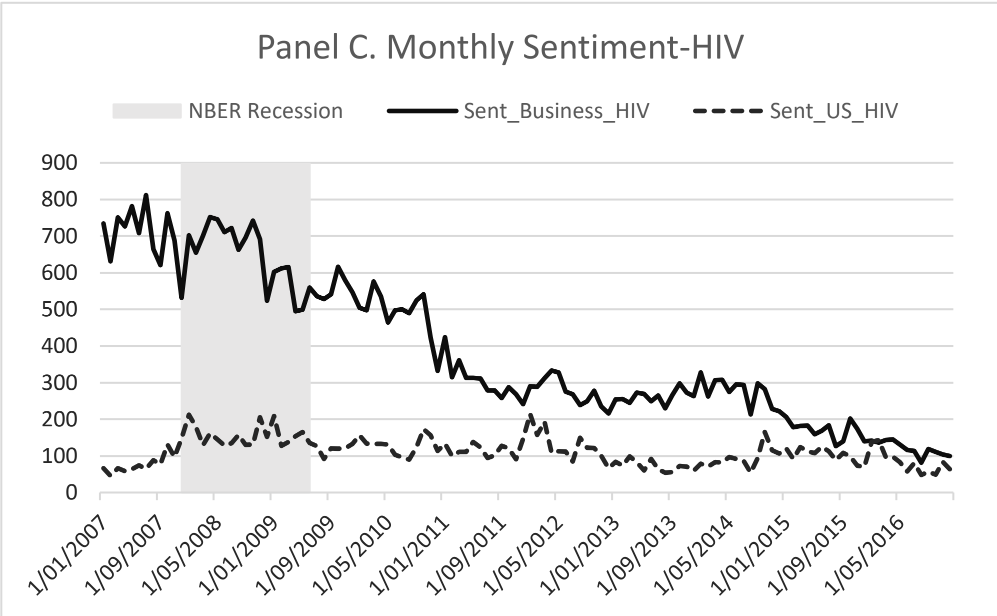
As shown in Panel C of Figure 6, the sentiment results from HIV-4 library tells a completely contradicted story. Most strikingly, the HIV-4 index generates a more optimistic sentiment result compared to the LM index when a vast majority of reports are negative. Furthermore, the HIV-4 library reports a surge of optimism during the GFC period, which is highly questionable.

Figure 6. Illustration of the Data Series

Figure 6 shows the full data series of the sentiment indexes throughout the sample period. Panel A is the excess return over risk-free rate of the DJIA. Panel B shows the total monthly sentiment indexes deducted from the LM dictionary, where *Sent_Business_LM* depicts the index from news articles in the “Business” section of the *New York Times*, while *Sent_US_LM* depicts the index from news articles in the “U.S.” section. Panel C shows the results from the HIV-4 library following the same process. Since principle component analysis is not applicable to the benchmark LM library, as only positive and negative subclasses are used in the index. NBER recession periods are highlighted.







To further consolidate the findings, I run an in-sample univariate ordinary least squares (OLS) regression using the monthly log excess return series of the DJIA. Both sentiment indexes are regressed against the DJIA return series in the current month to test their reportive power, as well as the return series next month to test their predictive power. The formula is:

$$\text{T0: } R_t^{DJIA} = \hat{\alpha} + \hat{\beta} \times S_t^{section} + \varepsilon_t \quad (3)$$

$$\text{T1: } R_{t+1}^{DJIA} = \hat{\alpha} + \hat{\beta} \times S_t^{section} + \varepsilon_{t+1} \quad (4)$$

Where $t \in T, section \in [Business, U.S.]$. R_t^{DJIA} is the log excess return of DJIA at time *t* and $Sent_t^{section}$ is the sentiment index generated by applying the LM library to the corresponding sections. $\hat{\alpha}, \hat{\beta}$, and ε_t are model parameters. Newey and West (1987) standard errors are used to ensure the results are robust to heteroskedasticity and autocorrelation.

Table 2. In-Sample Regression Results for LM-based Sentiment Indexes

Table 2 shows the regression results of the monthly log excess return of the DJIA against both sentiment indexes *Sent_Business_LM* and *Sent_US_LM*. Both indexes are regressed against the DJIA return series in the current month to test their reportive power, as well as the return series next month to test their predictive power. β and R^2 are the beta coefficient and R-squared score for the regression. Numbers in parenthesis report the p-statistic of β. *, **, and *** denote the two-tailed statistical significance at 10%, 5%, and 1% levels, respectively.

	T0		**T1**	
	β	R^2	β	R^2
Panel A. Full Data Set				
Sent_Business_LM	0.307**	0.094	0.218	0.047
	(0.019)		(0.115)	
Sent_US_LM	0.157	0.024	0.153	0.023
	(0.107)		(0.116)	
Panel B. Recession Period				
Sent_Business_LM	1.575***	0.409	0.432**	0.061
	(0.000)		(0.017)	
Sent_US_LM	0.858	0.107	0.626	0.113
	(0.178)		(0.063)	

As shown in Panel A of Table 2, throughout the sample period, *Sent_Business_LM* shows a significantly positive correlation with the current market returns, with beta coefficient $\beta = 0.307$ and $R^2_{T0} = 9.4\%$. Although the results show that the sentiment index still predicts around $R^2_{T1} = 4.6\%$ of the return in the next month, the predictive power is statistically insignificant. The results provide support for the findings in Tetlock (2007). However, during NBER-defined recessions, the media sentiment demonstrates much stronger predictive power, with a statistically significant R-squared of $R^2_{T1} = 6.1\%$. Considering that the number of articles in the Business section also peaks in 2008, the *NYT* likely concentrates on printing business-related topics during recessions, and the increase in focus is positively related to the predictive power. In contrast, The explanatory power of *Sent_US_LM* is statistically insignificant in both T0 and T1 tests. The results suggest that *Sent_Business_LM* has higher market return predictability than *Sent_US_LM*. The result provides support for above discussion that news articles in the Business section of the *NYT* demonstrate larger correlation to market returns. Thus, I mainly focus on this set of news below.

For comparison, I also construct another sentiment index based on the HIV-4 library. To begin with, I check how reactive the LM and HIV-4 is when working on business related documents. Using the results from the U.S. section as a baseline, it is possible to compare the sensitivity of both libraries when detecting sentiment on business-related documents. Business sensitivity is defined as:

$$Sens_{lib}^{\text{Business}} = \frac{Variance_{lib}^{\text{Business}}}{Variance_{lib}^{\text{US}}} \quad (5)$$

Where $lib \in [LM, HIV-4]$, $Sens_{lib}^{\text{Business}}$ is the business document sensitivity of a library, $Variance_{lib}^{\text{Business}}$ and $Variance_{lib}^{\text{US}}$ are the variance in sentiment scores under both sections from both libraries. The result of the positive and negative word counts and the sensitivity checks are presented in Table 3. It suggest that LM library is a more sensitive library when working on business-related topics as compared to the HIV-4 positive/negative subclass.

Table 3. Basic Statistics and Sensitivity Check

Panels A and B show the basic statistics of sentiment measures under the LM and HIV4 libraries. # denotes "the number of." Panel C shows the sensitivity test results. Higher score means more variance in the Business section is captured than the in the U.S. section, indicating a higher sensitivity to business-related topics.

	#Positive	#Negative	#Neutral	Max	Min	Mean	Stdev
Panel A. U.S.							
LM	6609	67327	3821	0.999	-0.999	-0.489	0.392
HIV4	58602	16996	2159	0.999	-0.999	0.169	0.277
Panel B. Business							
LM	24737	114958	10751	0.999	-0.999	-0.359	0.463
HIV4	133012	14663	2771	0.999	-0.999	0.307	0.258
Panel C. Sensitivity							
LM	1.395						
HIV4	0.868						

Next I put both sentiment indexes under an in-sample regression test following the same definition as formula (3) and (4). The in-sample test results in Table 4 suggest that the LM index significantly outperforms the HIV-4 index in almost every possible way. The LM index not only shows stronger predictive power with higher R-squared values, it is also statistically more significant during both expansions and recessions. The LM index is proven to be not only more sensitive to business related news articles, but also more relevant to the market activities in this study[5]. Thus, *Sent_Business_LM* is chosen as the sentiment index.

Table 4. In-Sample Regression Results Comparing LM and HIV-4 Library

Table 4 shows the regression results of the monthly log excess return of the DJIA against both LM and HIV-4 indexes. Both sentiment indexes are regressed against the DJIA return series in the current month to test their reportive power, as well as the return series next month to test their predictive power. The formula is defined as: T0: $R_t^{DJIA} = \hat{\alpha} + \hat{\beta} \times Sent_t^{lib} + \varepsilon_t$ T1: $R_{t+1}^{DJIA} = \hat{\alpha} + \hat{\beta} \times Sent_t^{lib} + \varepsilon_{t+1}$ where $t \in T$ and $lib \in$ [LM, HIV-4]. R_t^{DJIA} is the log excess return of DJIA at time *t* and $Sent_t^{lib}$ is the sentiment index generated by applying the LM and HIV-4 to the Business section. $\hat{\alpha}, \hat{\beta}$, and ε_t are model parameters. β and R^2 are the beta coefficient and R-squared score for the regression. Numbers in parenthesis report the p-statistic of β. Newey and West (1987) standard errors are used to ensure the results are robust to heteroskedasticity and autocorrelation. '*', '**' and '***' denotes significant at 10%, 5% and 1% level respectively.

	T0		T1	
	β	R^2	β	R^2
Panel A. Full Data Set				
Sent_Business_LM	0.307**	0.094	0.218	0.046
	(0.019)		(0.115)	
Sent_Business_HIV	-0.169	0.028	-0.164	0.027
	(0.128)		(0.127)	
Panel B. Recession Period				
Sent_Business_LM	1.575***	0.409	0.432**	0.061
	(0.000)		(0.017)	
Sent_Business_HIV	1.399	0.137	-0.967	0.058
	(0.106)		(0.256)	

[5] According to Loughran and McDonald (2011), the HIV-4 as a general library has a tendency to be overly sensitive to negative words that do not have pessimistic meanings in financial contexts. Their argument contradicts the result of this study where HIV-4 is overly optimistic. Since the performance of different BoW libraries is not the main focus of this study and LM is working decently for my data sets, I chose LM as the library and leave more in-depth examination between LM and HIV-4 for future discussion.

4.4 Combining Sentiment Indexes with a Knowledge Graph

Upon completion of above processes, a news article is decomposed into two components:

1. All the entities mentioned in the news article, and the corresponding vector representations $v_{entity} \in R^d$ based on the knowledge graph.

2. The sentiment score of the news article.

I assume that every news article proportionally reflect and impacts the market, and the impact on different targets varies. For instance, a news article mentioning “Microsoft” over 30 times might have a bigger impact on Microsoft than United Therapeutics. The topics of the news articles are captured by the entities, which are quantified by the embedding process. Since DJIA and its components are my targets of interest, the vector representation of the DJIA will be extracted from the knowledge graph as well. With all the available information, the entity-specific sentiment is formulized by:

$$Sent_{DJIA} = Sent_{news} \times e^{(-\lambda \times d_{DJIA}^{news})} \tag{6}$$

Where $Sent_{DJIA}$ is the effective sentiment of the news articles on the DJIA. $Sent_{news}$ is the original sentiment score of the news articles. d_{DJIA}^{news} is the vector distance between DJIA and the entities in the news articles. λ is a strictly positive control parameter.

The design of the sentiment index requires me to centralize the sentiment scores before modification by the impact factor $e^{(-\lambda \times d_{DJIA}^{news})}$. Most newspapers have a list of preferred words and phrases that are not generally neutral in most BoW libraries. The publication's general sentiment attitude and political slant could also reflect the viewpoint of the publisher. Furthermore, the choice of BoW library also affects the general tone. As shown in Subsection 4.3, the monthly sentiment from LM is strictly negative while the one from HIV-4 is purely positive. Here I assume that the textual sentiment of a news article is modeled by $Sent_{article} = Sent_{base} + Sent_{relative}$, where $Sent_{base}$ is the viewpoint of the publisher and $Sent_{relative}$ is the real sentiment that impacts the market. When using the raw sentiment scores derived from sentiment analysis as the original index, the base tone will also be amplified by the modifier. In the scenario of a strictly negative sentiment analysis, optimistic articles that frequently mention financial topics might yield a similar score as a pessimistic report on a trending movie star, which is problematic. Thus $Sent_{news}$ here is the centralized sentiment score that captures the relative tone of the collected articles.

To avoid overfitting, the KG modification factor is simply modeled by $e^{(-\lambda \times d_{DJIA}^{news})}$. d_{DJIA}^{news} is the standardized vector distance between $news$ and $DJIA$. The larger the distinction between the news and the target entities, the longer the distance, resulting in a diminishing KG modification factor that assigns less attention to the media sentiment. λ is a positive control parameter to ensure that the impact factor is consistent with different measures of distance.

By extending the idea to longer periods of time and larger sets of news articles, an entity-specific sentiment index can be generated. I set up multiple scenarios to test the performance of the knowledge graph. To find out how external knowledge could help improving the predictive performance of the sentiment index, I inserted $Sent_{DJIA}$ and $Sent_{news}$ back to back in an OLS regression following the same principle as stated in the Subsection 4.3. I test a selection of well-known (Qiu and Welch, 2004) sentiment measures, including the Baker-Wurgler sentiment index (Baker and Wurgler, 2006), as benchmarks.[6] Welch and Goyal (2008) comprehensively re-examined a list of highly influential macroeconomic and financial indicators. I adopted all 11 macroeconomic indicators as additional benchmarks.[7]

Chapter 5: In-Sample Regression Tests

In this chapter I focus on the in-sample performance of KG in predicting the return of

[6]Source: Malcolm Baker and Jeffery Wurgler, Investor Sentiment in the Stock Market, http://people.stern.nyu.edu/jwurgler/.

[7]Source: Ivo Welch and Amit Goyal, A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, http://www.hec.unil.ch/agoyal/.

DJIA. To maximize it comparability against other benchmarks, the KG modified index is calculated on a monthly basis which defined by:

$$Sent_{month_t}^{DJIA} = Sent_{month_t}^{news} \times e^{(-\lambda \times d_{month_t}^{news,DJIA})} \tag{7}$$

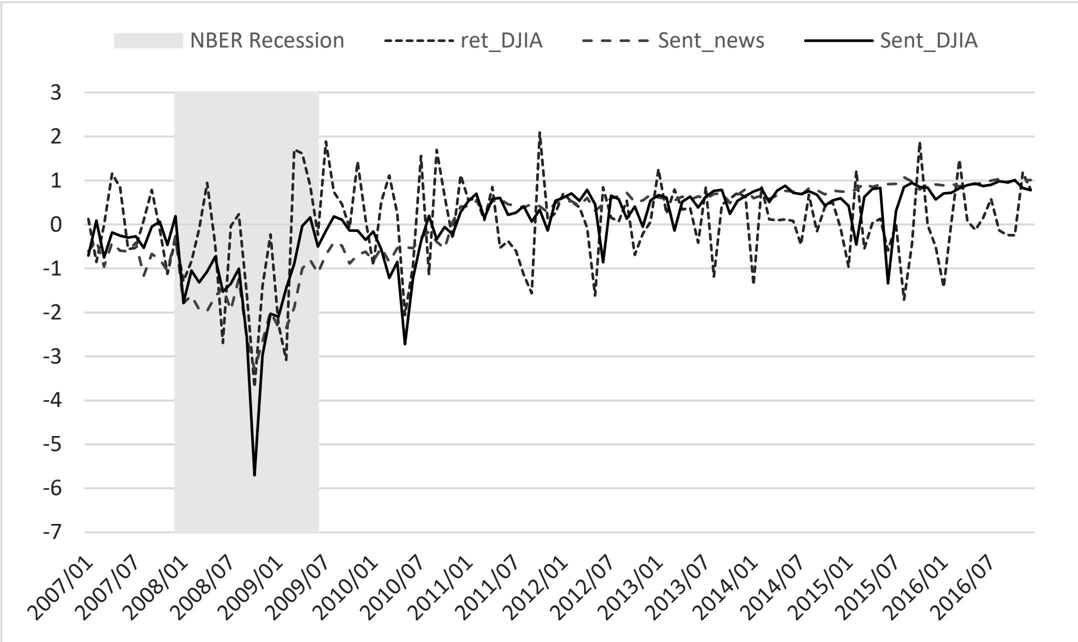
Where $t \in T$. $Sent_{month_t}^{DJIA}$ is the monthly KG modified sentiment, $Sent_{month_t}^{news}$ is the monthly unmodified sentiment score derived directly from applying LM to the Business section of the *NYT*. $d_{month_t}^{news,DJIA}$ is the monthly average distance between DJIA and all the news article.

5.1 General Effects of KG

The KG modified index *Sent_DJIA* is presented in Figure 7, alongside with the unmodified index *Sent_news* as well as the log excess return of DJIA *ret_DJIA*. All three data series are standardized to mean 0 and standard deviation 1 to ensure that they are comparable on a statistic level. There are two types of major effects resulting from the modified sentiment index. First, the KG modification tends to amplify the movements of the original sentiment index. I calculated the Lag 1 differential series $D_t = Sent_{t+1} - Sent_t$ for both *Sent_DJIA* and *Sent_news*. The standard deviation of the DJIA is 0.405, slightly more than the amount of the original series whose standard deviation is 0.324.

Figure 7. Illustration of Media Sentiment Indices: 2007 to 2016

Figure 7 depicts the monthly media sentiment indices from 2007 to 2016 (120 months). ret_DJIA is the monthly log excess return of DJIA. *Sent_news* is the unmodified sentiment index from the Business section of the *New York Times*, *Sent_DJIA* is the KG modified sentiment index. All data series have been standardized to mean 0 and standard deviation 1 for visual clarity. Both *Sent_news* and *Sent_DJIA* are mostly negative in the original series throughout the testing period. NBER recession periods are highlighted.



Second, the inclusion of external knowledge impels the original index towards a better convergence with the return series. To capture the correction effects, I count the number of times the KG modified sentiment index impels the original index towards the return series. Let r_t, $Sent_t^{news}$, and $Sent_t^{DJIA}$ denote the log excess return, unmodified sentiment score, and KG modified sentiment index score in month t, respectively. A modification is defined as $M_t = Sent_t^{DJIA} - Sent_t^{news}$, and an actual deviation is defined as $D_t = r_t - Sent_t^{news}$. A modification is considered correct when both M_t and D_t are positive (true positive, TP), or negative (true negative, TN). This setup enables me to conduct a confusion matrix, as shown in Table 5, to examine the general effects brought by external knowledge. The test yields a recall (true positive rate) of 0.692 and specificity (true negative rate) of 0.515, indicating that ~61% of the time the KG modification will improve the correlation between the unmodified sentiment index and the market return. Moreover, the modification shows a greater corrective power when returns are lower, with a negative predictive value (NPV) of 0.686, meaning that ~69% of negative corrections suggested by a KG modified sentiment index will improve the correlation. Although sentiment indexes and market returns are not directly comparable and any statistical inference from this test is unreliable outside return forecasting tasks, the confusion matrix provides a general intuition about the effects of the KG modification index.

Table 5. Confusion Matrix of the Correction Effects

Table 5 is the confusion matrix capturing the correction effects. To ensure that sentiment indexes and stock returns are directly comparable, all data series have been standardized to mean 0 and standard deviation 1. r_t, $Sent_t^{news}$, and $Sent_t^{DJIA}$ denote the log excess return, unmodified sentiment score, and KG modified sentiment index score in month t, respectively. A modification is defined as $M_t = Sent_t^{DJIA} - Sent_t^{news}$, and an actual deviation is defined as $D_t = r_t - Sent_t^{news}$.

		Modification M_t	
		Positive	Negative
Deviation D_t	Positive	36	16
	Negative	33	35

To formally test whether the knowledge graph will provide additional insight beyond known sources of predictability, I run a standard univariate OLS regression for both *Sent_DJIA* and *Sent_news*, as well as a selection of benchmarks sentiment indexes and marcoeconomic predictors. Table A.1. and A.2.[8] show the basic summary of statistics and correlation coefficient among 18 predictors in total.

5.2 In-Sample Regression Results

The in-sample univerate OLS regression is defined as:

$$\text{T0: } R_t^{DJIA} = \hat{\alpha} + \hat{\beta} \times P_t^i + \varepsilon_t \tag{8}$$

$$\text{T1: } R_{t+1}^{DJIA} = \hat{\alpha} + \hat{\beta} \times P_t^i + \varepsilon_{t+1} \tag{9}$$

Where $t \in T,\ i \in \{1,\ 2,\ \ldots,\ 18\}$. R_t^{DJIA} is the log excess return of DJIA at month *t*. P_t^i is the i^{th} predictor at month *t*. $\hat{\alpha}$, $\widehat{\beta}$, and ε_t are model parameters. Newey and West (1987) standard errors are used to ensure the results are robust to heteroskedasticity and autocorrelation.

As shown in Table 6, both *Sent_DJIA* and *Sent_news* are strongly correlated with the monthly returns. On average, 9.4% of the change in monthly returns can be explained by the unmodified media sentiment in the same month, while using the KG modified sentiment index boosts the R-squared to an impressive 13.7%. Moreover, the KG modified sentiment index is statistically more significant than the original one at the 1% level. The results of the T1 next month predictability test show that the average impact of a one standard deviation increment in the unmodified index on the stock return next month is 0.218 standard deviation, which translates into 91.6 basis points, considerably

[8] Table A.1. and Table A.2. can be found in the Appendix.

larger than the unconditional mean of the DJIA returns (30 basis points). Although the predictive power of the unmodified index holds up to be one of the highest with R-squared of 4.7%, it is no longer significant at the 1% level, dropping to a two-tail *p*-value of 0.115. Similarly, the explanatory power of the KG modified index decreases dramatically when moving from reporting the market activities this month to predicting the stock returns next month. However, the performance improvements brought by KG seizes its momentum, the KG modified index is outcompeting its unmodified counterpart in every aspect. Outside the feudal competition, the KG modified index stands out with the strongest predictive power with T1 R-squared of 7.3%. The only potential competitor is the stock variance (sum of squared daily returns, denoted by *SVAR*), with an unparalleled T0 R-squared of 18.8%. Overall the KG modification noticeably increases the in-sample explanatory power of the media sentiment index in the full data set. Although the improvement shows striking similarities with the correction effects discussed in Subsection 5.1, the insufficient testing data prevents me from consolidating the theory. When comparing against other benchmarks, the KG modified index outperforms almost all competing predictors by a sizeable margin.

Table 6. In-Sample Regression Results: 2007 to 2016

Table 6 shows the in-sample regression results for the 2007-2016 period. All regressions are conducted on standardized data series with a mean of 0 and a standard deviation of 1. Both *Sent_DJIA* and *Sent_news* are regressed against the DJIA return series in the current month to test their reportive power, as well as the return series next month to test their predictive power. β and R^2 are the beta coefficient and R-squared score for the regression. Numbers in parenthesis report the p-statistic of β. *, ** and *** denotes significant at 10%, 5% and 1% level respectively.

	T0		T1	
	β	R^2	β	R^2
Sent_news	0.307**	0.094	0.218	0.047
	(0.019)		(0.115)	
Sent_DJIA	0.372***	0.137	0.272	0.073
	(0.004)		(0.098)	
SENT	-0.107	0.011	-0.126	0.016
	(0.277)		(0.190)	
PDND	-0.169	0.028	-0.145	0.021
	(0.235)		(0.331)	
NIPO	0.126	0.016	0.061	0.004
	(0.270)		(0.587)	
CEFD	0.026	0.001	0.090	0.008
	(0.813)		(0.396)	
EQTI	-0.051	0.003	-0.029	0.001
	(0.665)		(0.801)	
D12	0.011	0.000	0.012	0.000
	(0.890)		(0.880)	
E12	0.102	0.010	0.063	0.004
	(0.502)		(0.663)	
BM	-0.191	0.036	0.085	0.007
	(0.09)		(0.407)	
Rf	-0.084	0.007	-0.108	0.012
	(0.290)		(0.223)	
DFY	0.197	0.038	0.121	0.014
	(0.242)		(0.455)	
LTY	-0.089	0.008	-0.165	0.027
	(0.345)		(0.067)	
NTIS	0.212	0.045	0.179	0.032
	(0.086)		(0.137)	
INFL	0.086	0.007	0.132	0.017
	(0.528)		(0.129)	
LTR	-0.266*	0.070	0.097	0.009
	(0.036)		(0.362)	
CORPR	0.082	0.007	0.159	0.025
	(0.500)		(0.307)	
SVAR	-0.435***	0.188	-0.230*	0.052
	(0.000)		(0.027)	

Table 7 presents the regression results during the NBER recession periods. A drastic increase in the predictability for both *Sent_DJIA* and *Sent_news* is observed, which has been well documented by previous literature (e.g., Rapach, Strauss, and Zhou, 2010, 2013; Henkel, Martin, and Nardari, 2011; Dangl and Halling, 2012; Garcia, 2013; Adämmer and Schüssler, 2020). Although these studies employ different forecasting models, data sets, and evaluation periods, they all report much stronger predictive power during economic downturns.

Table 7. In-Sample Regression Results During NBER Recessions

Table 7 shows the in-sample regression results for recessions. Note that since the data I use from NBER recessions is an unstandardized slice of the full data, the beta coefficients are no longer consistent with the R-squared values. Both sentiment indices are regressed against the DJIA return series in the current month to test their reportive power, as well as the return series next month to test their predictive power. β and R^2 are the beta coefficient and R-squared score for the regression. Numbers in parenthesis report the p-statistic of β. *, ** and *** denotes significant at 10%, 5% and 1% level respectively.

	T0		**T1**	
	β	R^2	β	R^2
Sent_news	1.575***	0.409	0.858**	0.107
	(0.000)		(0.017)	
Sent_DJIA	0.760**	0.235	0.492	0.086
	(0.006)		(0.144)	
SENT	-0.189	0.025	-0.367	0.082
	(0.521)		(0.168)	
PDND	-0.377	0.053	-0.622	0.127
	(0.202)		(0.054)	
NIPO	1.189	0.075	1.315	0.081
	(0.259)		(0.255)	
CEFD	-0.056	0.003	0.195	0.026
	(0.827)		(0.428)	
EQTI	-0.181	0.012	-0.100	0.003
	(0.598)		(0.773)	
D12	-4.799*	0.159	-7.148***	0.309
	(0.036)		(0.000)	

E12	-0.244	0.019	-0.510	0.071
	(0.587)		(0.269)	
BM	-0.776***	0.319	-0.066	0.002
	(0.000)		(0.856)	
Rf	-0.270	0.012	-0.457	0.031
	(0.595)		(0.418)	
DFY	0.024	0.000	-0.049	0.002
	(0.926)		(0.867)	
LTY	-0.636	0.039	-0.295	0.007
	(0.324)		(0.765)	
NTIS	0.498	0.037	0.523	0.036
	(0.501)		(0.483)	
INFL	0.139	0.029	0.253	0.083
	(0.401)		(0.098)	
LTR	0.081	0.007	0.145	0.020
	(0.674)		(0.376)	
CORPR	0.272	0.109	0.186	0.045
	(0.059)		(0.421)	
SVAR	-0.331***	0.194	-0.171	0.045
	(0.002)		(0.074)	

More intriguingly, the outperformance brought by the KG modification no longer gains momentum during NBER recession periods. In terms of both T0 reportive power and T1 predictive power, *Sent_DJIA* is outcompeted by *Sent_news* in every criterion. Beyond common sources of statistical distractions, the chief suspect for this observation is the degenerating quality of the external knowledge. Real world knowledge is intrinsically time-sensitive. The knowledge graph adopted in this study is constructed based on the world in 2016, which is considerably distinct from its pre-GFC counterpart. The results suggest that incorrect external knowledge might result in misleading modifications that taint the predictions.

I conduct a Diebold-Mariano test (Diebold and Mariano, 1995) to examine whether or not the KG modification improves the prediction of stock returns. To mitigate this test's overreaction in small samples, the HLN modification (Harvey, Leybourne, and Newbold, 1997) is introduced to the DM test. I find that the DM statistic is 2.207 with a two-tail p-value of 0.029. This suggests that although the statistical evidence is not sufficiently strong to confirm that the outperformance of the KG modified sentiment index is consistent at the 5% level, the index is still reasonably distinct from and better than the unmodified one. I conclude that the KG modification increased the performance of the unmodified media sentiment index, which is an impressive predictor itself.

The unmodified sentiment index based on the *NYT* articles outperforms the benchmark sentiment indices and macroeconomics factors. This outperformance is further strengthened by the inclusion of knowledge graph, apart from a few exceptions (e.g., SVAR in T0 regression, suggesting a stronger reporting power for the stock returns in the current month).

5.3 Theory Behind the Sentiment Index

As expected, the correlation coefficients of both sentiment indexes are predominately positive as shown in Table 6. This positive relation between media sentiment and short-term stock returns provides support for findings from other studies (e.g., Brown and Cliff, 2005; Tetlock, 2007; Ni, Wang, and Xue, 2015). The positive correlations suggest that an optimistic sentiment might drive market valuations beyond their intrinsic value. In addition, Brown and Cliff (2005) claim that the over optimism can only temporally elevate market valuations, which will be corrected in the long term. They find a clear descending trend in beta coefficients when performing OLS regressions to longer lags; the optimistic sentiment only prevails for 6-24 months. When the optimism wears off, the market starts to correct the over valuation, resulting a long-term negative coefficient.

Following the same rationale, I test both *Sent_DJIA* and *Sent_news* over longer periods. The univariate OLS is defined by:

$$R_{t+h}^{DJIA} = \hat{\alpha} + \hat{\beta} \times P_t^i + \varepsilon_{t+h} \tag{10}$$

Where $h > 1$, $i \in [Sent\_news,\ Sent\_DJIA]$, and R_{t+h}^{DJIA} is the DJIA log excess return at month *t+h*. $\hat{\alpha}$, $\widehat{\beta}$, and ε_{t+h} are model parameters. Newey and West (1987) standard errors are used to ensure the results are robust to heteroskedasticity and autocorrelation.

Table 8. In-Sample Regression Results for Longer Time Horizon

Table 8 shows the correlation coefficients in the regression results for longer time periods. None of the beta coefficient proven to be statistically significant.

	Sentiment Indexes	
Time Horizon	Sent_news	Sent_DJIA
1 month	0.220	0.272
2 months	0.192	0.170
3 months	0.219	0.211
4 months	0.182	0.178
5 months	0.116	0.086

6 months	0.056	-0.016
12 months	0.018	-0.054
18 months	0.006	-0.068
24 months	-0.097	-0.113
30 months	-0.019	0.030

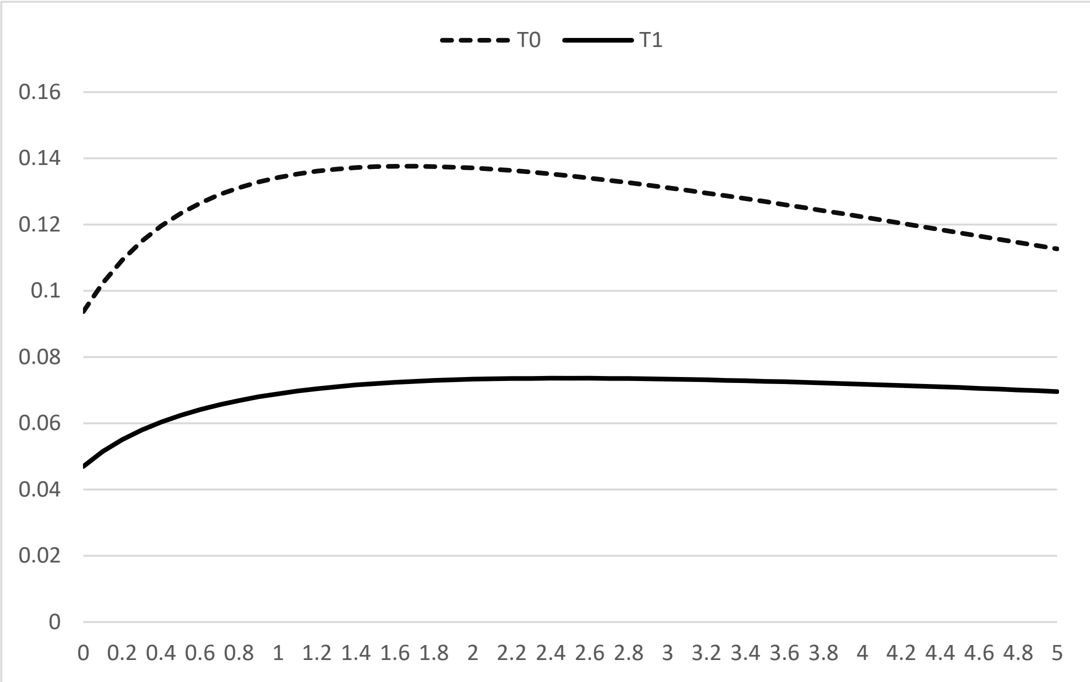
The results in Table 8 are consistent with the findings in Brown and Cliff (2005). I find that both sentiment indices are positively related to the return series up to a certain point. For the unmodified sentiment index, it starts to negatively impact the market return during the 19th month, while the KG modified sentiment index starts to correct in the 6th month. None of the sentiment indexes demonstrate statistical significance over longer time horizons. This could be due to unobserved factors, which are not examined in this study. In conclusion, both indices possess typical features of sentiment estimations documented in previous studies. Furthermore, he KG modification increases the sensitivity of the original sentiment index, correcting more quickly to its intrinsic value than the unmodified one.

5.4 In-Sample Robustness Check

In this subsection, I conduct in-sample robustness test to examine the improvements that the KG modification index demonstrates above. Characterized by formula (7), by design the prediction results are affected by the choice of the control parameter λ. Because of the independent nature between the sentiment index $Sent_{month_i}^{news}$ and the distance measure $d_{month_i}^{news,DJI}$, changes in λ could potentially alter the overall trend of the entity-specific index $Sent^{DJIA}_{month_i}$. For instance, assume that $Sent^{news}_{month1} = 0.8$ and $Sent^{news}_{month2} = 0.2$, $d^{news,DJIA}_{month1} = 0.9$ and $d^{news,DJIA}_{month2} = 0.1$, and $Sent^{DJIA}_{month1} > Sent^{DJIA}_{month2}$ when $\lambda = 1.7$ while $Sent^{DJIA}_{month1} < Sent^{DJIA}_{month2}$ when $\lambda = 1.8$. In the case when the KG modification is not actually correcting the regression, the resulting goodness-of-fit might fluctuate along with different choices of λ, in which case it is possible that the seemingly promising results are achieved by a lucky hit on λ. To test the robustness of the index on the choice of parameter, I first perform a series of in-sample OLS regressions with different choices of λ then check the goodness-of-fit, as shown in Figure 8.

Figure 8. In-Sample Robustness Test

Figure 8 shows the R-squared scores of 51 in-sample OLS regressions with different choices of λ. The x-axis is the value of the control parameter λ, while the y-axis is the R-squared score of the goodness-of-fit measure. The T0 curve depicts the R-squared value in the T0 regressions, while the T1 curve depicts the R-squared value in T1 predictions. The choice of λ ranges from 0 to 5.0, with intervals of 0.1. Since the impact factor grows exponentially with λ, any number larger than 5.0 is not likely to be picked for prediction. When $\lambda = 0$, the entity-specific sentiment index $Sent^{DJIA}_{month_i}$ is equivalent to the original sentiment index.



The results are encouraging in Figure 8. When the results are random and the knowledge graph is not actually improving the prediction, there will be a zigzag discrete R-squared curve when switching between different λ values. However, the smooth continuous R-squared curve proves that the noise disturbance is minimal compared to the overall trend. I start the test from $\lambda = 0$, as shown on the left of Figure 8. By definition, $Sent_{month}^{DJIA}$ is equivalent to $Sent_{month}^{news}$ when $\lambda = 0$. It appears that regardless the choice of λ, $Sent_{month}^{DJIA}$ yields better predictions than $Sent_{month}^{news}$, as shown by the very left point also being the lowest point. Moreover, the result is consistent with both the reportive power of T0 and the predictive power of T1. Both curves show similar trends and peaks. The robustness test results suggest that the KG modification consistently improves the performance of the original sentiment index, regardless of the choice of the control parameter λ for both the T0 reportive tasks and the T1 predictive tasks.

Chapter 6: Out-of-Sample Regression Tests

In typical stock return prediction tasks, in-sample superiority does not necessarily translate to better performance in out-of-sample predictions (Welch and Goyal, 2008). A major concern here is to what extent does the KG modified sentiment index retain its positive effects out-of-sample, and more importantly, how consistent it is.

6.1 Model Description and Results

The simple historical mean has been well documented (Welch and Goyal, 2008) as a stringent baseline in out-of-sample prediction tasks. In the following tests, I use the historical mean $\widehat{r_{t+1}^{HM}} = \overline{r_{1:t}}$ as a benchmark to put all the predictors into context. I employ the widely accepted out-of-sample R-squared (Campbell and Thompson, 2008) as the main evaluation standard to test how well the predictors perform as compared to the historical mean. The out-of-sample R-squared is calculated as:

$$R_{OS}^2 = 1 - \frac{\sum(r_t - \widehat{r_t})}{\sum(r_t - \overline{r_t})} \qquad (11)$$

Where R_{OS}^2 is the out-of-sample R-squared. r_t is the stock return at time *t*. $\widehat{r_t}$ is the model prediction of the return at time *t*. $\overline{r_t}$ is the historical mean of return up until time *t*. Diebold-Mariano statistics (Diebold and Mariano, 1995) with Harvey-Leybourne-Newbold modification (Harvey, Leybourne, and Newbold, 1997) and Clark-West test (Clark and West, 2007) are used as auxiliaries in addition to R_{OS}^2. The results are presented in Table 9.

Table 9. Summary of the out-of-sample test

Table 9 provides the out-of-sample prediction results for all competing predictors. Predictions are generated by an univariate regression based on all historical data with the recursive window starting from the 56th month. The recursive prediction function is defined as: $\hat{R_{t+1}} = \hat{\alpha_i} + \hat{\beta_i} \times P_t^i$, where P^i is the i^{th} predictor, and $\hat{\alpha_i}$ and $\hat{\beta_i}$ are the corresponding parameter estimations in $R_t = \hat{\alpha_i} + \hat{\beta_i} \times P_{t-1}^i + \varepsilon_t^i$ trained from all historical data (from 1 to *t*-1). R_{OS}^2 is the out-of-sample R-squared statistics (Campbell and Thompson, 2008) capturing the percentage reduction in mean squared error (MSE) as compared to the historical mean predictions. DM_HLN is the Diebold-Mariano (DM) statistics (Diebold and Mariano, 1995) of every predictor over the historical mean. I adjust the DM test using the Harvey-Leybourne-Newbold (HLN) modification (Harvey, Leybourne, and Newbold, 1997) to improve the compatibility with small samples. CW is the *t*-statistics in the Clark-West test (Clark and West, 2007). Both DM_HLN and CW tests are used as auxiliaries in addition to R_{OS}^2 to indicate the level of significance. Newey and West (1987) standard errors are used to ensure the results are robust to heteroskedasticity and autocorrelation. '*', '**' and '***' denotes significant at 5%, 2% and 1% level respectively.

	RMSE	R_{OS}^2	DM_HLN	CW
Sent_news	0.763	0.029	0.288	1.939**
Sent_DJIA	0.744	0.075	1.121	2.315**
SENT	0.775	-0.003	-0.141	0.217
PDND	0.790	-0.044	-1.114	-0.379
NIPO	0.790	-0.043	-1.316	-0.868
CEFD	0.767	0.017	1.235	1.431*
EQTI	0.776	-0.006	-1.868	-1.795
D12	0.880	-0.294	-3.304	-2.349
E12	0.777	-0.009	-0.538	-0.338
BM	0.769	0.012	1.254	1.456*
Rf	0.764	0.025	1.400	1.891*
DFY	0.765	0.023	1.153	1.911*
LTY	0.758	0.040	0.487	2.139**
NTIS	0.778	-0.012	-0.225	0.624
INFL	0.800	-0.071	-1.763	-1.052
LTR	0.770	0.009	0.324	0.719
CORPR	0.770	0.009	0.176	0.894
SVAR	0.768	0.014	0.465	1.203

The KG modified sentiment index (*Sent_DJIA*) overwhelms the unmodified index (*Sent_news*) in all instances; it not only outperforms the historical mean by a larger margin (R_{OS}^2 = 0.075 compared to 0.029), but is also more statistically significant in both the DM_HLN and CW tests. In addition, the KG modified sentiment index is the most effective predictor among all available benchmarks. It dwarfs all competitors with an unparalleled R_{OS}^2 of 7.5%, and is triples the R_{OS}^2 of its closest rival outside the two major sentiment indices, which is the risk-free rate (*Rf*) with and R_{OS}^2 of 2.5%. The DM and CW test results suggest that this outperformance is one of the most statistically significant inferences that can be drawn from the prediction results.

6.2 Out-of-Sample Robustness Check

The prediction outcomes are closely related to the recursive window I use. When switching between various setups of regressions with different recursive window size, the performance of the predictors changes dramatically. In the discussion above, I start the prediction from the 56^{th} month, which yields the best results for both *Sent_news* and *Sent_DJIA*. In practice, there is no feasible way to obtain this information in advance, and the choice of training and testing periods is highly subjective and situational. Naturally it raises the question as to how consistent can we expect the outperformance of the KG modified sentiment index to be.

I conduct a series of prediction experiments as robustness tests. A point in time t is chosen as the starting month. All 18 competing predictors are then trained to predict the next month's excess return following the same univariate prediction function $\widehat{R_{t+1}} = \hat{\alpha} + \hat{\beta} \times P_t^i$ as defined above. The R_{OS}^2 from each predictor is collected in a corresponding 18-dimensional vector S_t. The choice of t ranges from ¼ to ¾ the total number of months (30 to 90). The experiments exhaust all reasonable choices of time periods and the R_{OS}^2 statistics are given in a matrix $M \in R^{18\times 61}$, summarized in Table 10.

Table 10. Summary of the Out-of-Sample Tests.

Table 10 provides the R_{OS}^2 results. "Mean" is the average R_{OS}^2 from all 61 choices of t for each predictor. As the chief criterion for robustness tests, "Mean" indicates the average level of outperformance from each predictor over the historical mean; the higher the better. "Max" and "Min" are the maximum and minimum level of R_{OS}^2, capturing the range of variation. "Stdev" is the standard deviation of R_{OS}^2, reflecting how stable a predictor performs.

	Mean	Max	Min	Stdev
Sent_news	-0.040	0.029	-0.108	0.031
Sent_DJIA	0.027	0.075	-0.018	0.023
SENT	-0.001	0.015	-0.013	0.006
PDND	-0.050	-0.007	-0.081	0.020
NIPO	-0.022	0.000	-0.051	0.012
CEFD	0.016	0.022	0.001	0.004
EQTI	-0.005	0.000	-0.039	0.007
D12	-0.229	-0.060	-0.378	0.095
E12	-0.004	0.007	-0.025	0.010
BM	0.002	0.024	-0.083	0.029
Rf	0.012	0.025	-0.004	0.007
DFY	0.016	0.030	-0.001	0.008
LTY	-0.001	0.040	-0.042	0.020
NTIS	-0.055	0.011	-0.097	0.023

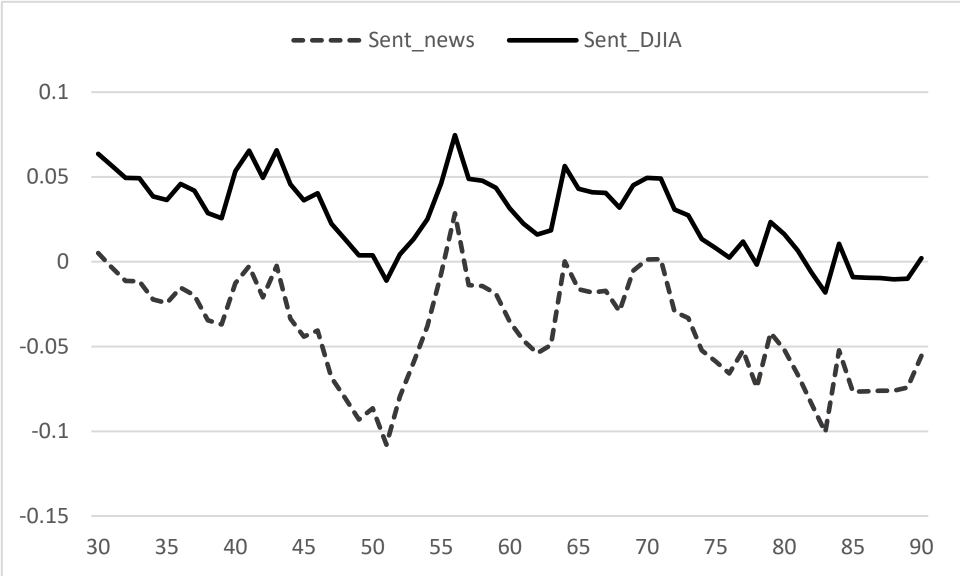
INFL	-0.048	-0.019	-0.076	0.012
LTR	-0.004	0.021	-0.039	0.015
CORPR	-0.014	0.009	-0.037	0.011
SVAR	0.018	0.048	-0.022	0.020

Based on the robustness tests, the unmodified sentiment index (*Sent_news*) only achieves the seemingly decent out-of-sample prediction at $t = 56$. On average, when the choice of time period is not guided by external intervention, it is 4% inferior compared to the historical mean. Unsurprisingly, a majority of the predictors are outcompeted by the historical mean (by up to 23%). Only 6 out of 18 predictors manage to outperform the historical mean and the KG modified index (*Sent_DJIA*) is superior, with an average 2.7% less mean squared error (MSE), which is 50% more than its closest competitor, stock variation (*SVAR*), which has an 1.8% lower MSE. Interestingly, the closed-end fund discount rate (*CEFD*), the underperforming predictor in the in-sample tests, is another challenger to the KG modified sentiment index. Although the *CEFD* has lower average accuracy that the KG modified sentiment index, its R_{OS}^2 is strictly positive throughout all the tests, with a minimum of 0.1% improvement over the historical mean. This stable predictor might be preferred over the KG modified sentiment index by more risk-averse investors. However, the incoherent in-sample and out-of-sample performance puts its robustness in doubt. In conclusion, the KG modified sentiment index is a clear improvement in every aspect compared to the unmodified sentiment index. The KG modification drastically reduces the prediction MSE, and the performance itself become more stable, with a 25% smaller standard deviation.

By directly comparing the out-of-sample R-squared series of *Sent_DJIA* and *Sent_news*, deeper insights into the effects of KG could be excavated, as illustrated in Figure 11. First, the two R_{OS}^2 series are strongly correlated, with a correlation coefficient of 0.95 ($\rho = 0.950$). The result stems largely from the definition of the *Sent_DJIA*, which merely uses an impact factor to modify the *Sent_news*. The strong correlation confirms that the modification does not irrationally alter the modified index from the original one. Second, the influence of the KG modification is strictly positive. For every time window examined, *Sent_DJIA* consistently outperforms *Sent_news*. This suggests that the outperformance brought by KG modification is robust to the time period and data set selections.

Figure 9. Out-of-Sample Robustness Test

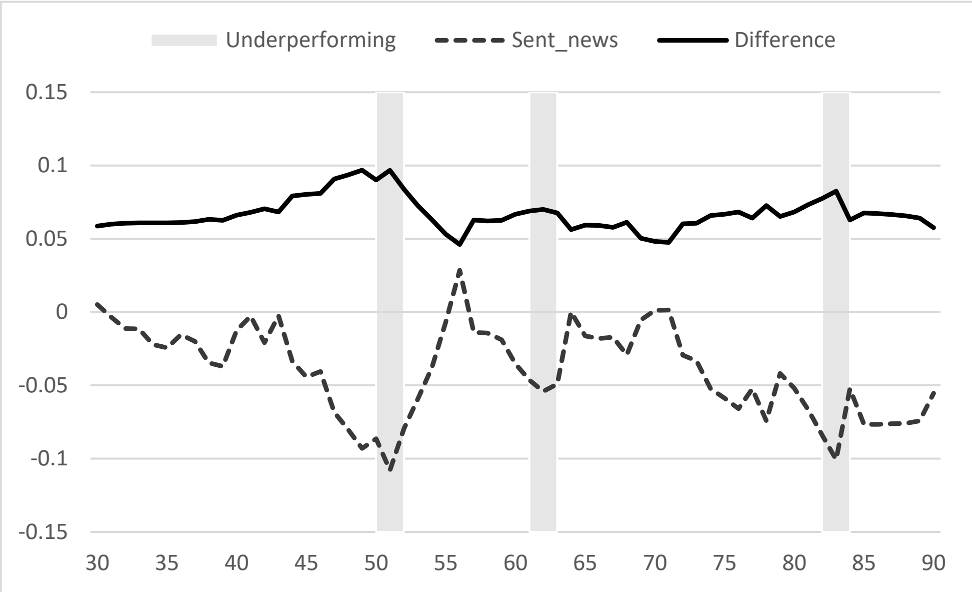
Figure 9 depicts the change of R_{OS}^2 when switching between different starting time periods. The x-axis captures the selected time period t, while the y-axis is the R_{OS}^2 statistics. *Sent_news* depicts the R_{OS}^2 of the unmodified index under different settings of time window, while *Sent_DJIA* depicts the R_{OS}^2 of the KG modified sentiment index in the same settings. The choice of t ranges from 30 to 90, with intervals of 1, which exhausts all reasonable regression setups.



Moreover, an obvious parallel elevation from *Sent_news* to *Sent_DJIA* can be observed. The main reason behind this parallel shift is that the R^2_{OS} captures the cumulative reduction in MSE throughout the testing dataset. For every setup (time window from 30 to 90), data in the last 30 months is included in the testing dataset, which dictates the overall improvements in R^2_{OS}. To characterize the real improvements brought by the KG modification from each time window, I calculate the point-to-point R^2_{OS} difference between these two indexes as follows: $D_t = R^2_{OS,DJIA,t} - R^2_{OS,news,t}$ (Figure 10). Larger D_t indicates better improvements.

Figure 10. $\mathbf{R^2_{OS}}$ Differences Through Different Time Windows

Figure 10 depicts the effective differences in R^2_{OS} statistics between the KG modified sentiment index and the unmodified one during different time periods. The x-axis denotes the time window *t*, while the y-axis denotes the difference in R^2_{OS} statistics D_t. A higher D_t means the KG modification brought higher improvements when time window is *t*. The window size when the unmodified sentiment index is underperforming is highlighted.



Apart from the strictly positive improvement discussed above, another noticeable trait is that the improvement tends to be larger when the unmodified sentiment index underperforms. The R^2_{OS} difference peaks during the 51st, 62nd, and 83rd month, indicating that the KG modification provides stronger improvements during these setups. For the original raw sentiment index, the lowest R^2_{OS} values are observed during these time periods (highlighted in Figure10). To formally examine this, I regressed the difference series D_t on the R^2_{OS} statistics of the unmodified sentiment index $R^2_{OS,news,t}$. The OLS regression results show a strong negative correlation coefficient $\rho = -0.756$ with $R^2 = 0.572$, further proving the effects.

These results suggest that KG modification delivers a significant out-of-sample correction effect on the original raw sentiment index. In out-of-sample prediction tasks, the KG modification impels the predictions from the original raw sentiment index to a better convergence with the real excess return next month. More interestingly, the correction effects become increasingly prominent when the original raw sentiment index underperforms, resulting in a much more stable overall performance throughout different time periods and data sets.

Chapter 7: Conclusion

In this study, I explore how general knowledge can be quantified and used to conduct better financial inferences from media outlets following a similar reasoning process as human newspaper readers. I construct an intuitive index to measure media sentiment on a specific entity. By design, the index not only discriminates news articles based on their relevance to a specific topic, it can also factor in the potential impacts of the articles even if none of the topics are directly mentioned in the article, making it an idea tool to disentangle noisy data.

The most noticeable effect brought by the introduction of external knowledge is that it redirects the estimation of media sentiment to provide a better interpretation of market activity. In both in- and out-of-sample regression tests, KG modification consistently improves the performance of the unmodified sentiment index. The correction effects tend to be more prominent when the unmodified index underperforms, making the performance of the predictor more stable in different model settings and time periods. In addition, the regression results are robust to the choice of parameter and time period. Last but not least, the KG modified sentiment index possesses the highest predictive power among all the benchmark indexes examined in this study. The outstanding performance of the KG modified sentiment index suggests there is great potential in implementing this index in trading strategies.

The unmodified media sentiment index derived from LM shows typical behaviour of sentiment estimations documented by Brown and Cliff (2005). Although it is assertive to conclude that LM can extend its application from financial documents to general documents, it performs reasonably well in this study, showing potentials for applications in business related news articles. On the other hand, the KG modified index reverts to its intrinsic value even quicker than the unmodified index, resulting in a more sensitive estimation of media sentiment.

However, the effect of external knowledge is heavily affected by the quality of the knowledge. A lurking concern shrouding the viability of knowledge graphs is the time-sensitive nature of knowledge, since outdated knowledge could potentially lead to a wrong modification, as shown in Subsection 5.2. Fortunately, knowledge representation technologies have become increasingly popular. There has been an upsurge in structured databases from major tech giants (such as Microsoft Satori) and open source projects (such as DBpedia). The frequently updated external knowledge sources could fuel future research in this topic.

References

Bachmann, R., Elstner, S. and Sims, E.R., 2013. Uncertainty and economic activity: Evidence from business survey data. *American Economic Journal: Macroeconomics*, 5(2), pp.217-49.

Baker, M. and Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4), .1645-1680.

Barberis, N., Shleifer, A. and Wurgler, J., 2005. Comovement. *Journal of financial economics*, *75*(2), pp.283-317.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of Machine Learning Research,* 3(Jan), 993-1022.

Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E., 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: theory and experiment*, 2008 (10), pp.10008.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. and Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems. pp. 2787-2795.

Brown, G.W. and Cliff, M.T., 2005. Investor sentiment and asset valuation. *The Journal of Business*, *78*(2), pp.405-440.

Calomiris, C.W. and Mamaysky, H., 2019. How news and its context drive risk and returns around the world. *Journal of Financial Economics*, 133(2), pp.299-336.

Campbell, J.Y. and Thompson, S.B., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average?. *The Review of Financial Studies*, 21(4), pp.1509-1531.

Clark, T.E. and West, K.D., 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), pp.291-311.

Dangl, T. and Halling, M., 2012. Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, *106*(1), pp.157-181.

Diebold, F.X. and Mariano, R.S., 2002. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1), pp.134-144.

Garcia, D., 2013. Sentiment during recessions. *The Journal of Finance*, 68(3), pp.1267-1300.

Harvey, D., Leybourne, S. and Newbold, P., 1997. Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2), pp.281-291.

Henkel, S.J., Martin, J.S. and Nardari, F., 2011. Time-varying short-horizon predictability. *Journal of Financial Economics*, 99(3), pp.560-580.

Ji, G., He, S., Xu, L., Liu, K. and Zhao, J., 2015, July. Knowledge graph embedding via dynamic mapping matrix. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 687-696.

Daiber, J., Jakob, M., Hokamp, C. and Mendes, P.N., 2013, September. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pp. 121-124.

De Long, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J., 1990. Noise trader risk in financial markets. *Journal of political Economy*, *98*(4), pp.703-738.

Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), pp.383-417.

Griffin, J.M., Hirschey, N.H. and Kelly, P.J., 2011. How important is the financial media in global markets?. *The Review of Financial Studies*, *24*(12), pp.3941-3992.

Kearney, C. and Liu, S., 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, pp.171-185.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S. and Bizer, C., 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2), pp.167-195.

Lin, Y., Liu, Z., Sun, M., Liu, Y. and Zhu, X., 2015, February. Learning entity and relation embeddings for knowledge graph completion. In Twenty-ninth AAAI conference on artificial intelligence.

Liu, B. and McConnell, J.J., 2013. The role of the media in corporate governance: Do the media influence managers' capital allocation decisions?. *Journal of Financial Economics*, 110(1), pp.1-17.

Loughran, T. and McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), pp.35-65.

Loughran, T. and McDonald, B., 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), pp.1187-1230.

Merton, R.C., 1987. A simple model of capital market equilibrium with incomplete information.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111-3119.

Newey, W.K. and West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix, *Economietrica* 55, pp. 703-708.

Ni, Z.X., Wang, D.Z. and Xue, W.J., 2015. Investor sentiment and its nonlinear effect on stock returns—New evidence from the Chinese stock market based on panel quantile regression model. Economic Modelling, 50, pp.266-274.

Pástor, Ľ. and Veronesi, P., 2013. Political uncertainty and risk premia. *Journal of Financial Economics*, 110(3), pp.520-545.

Qiu, L. and Welch, I., 2004. Investor sentiment measures (No. w10794). *National Bureau of Economic Research.*

Rapach, D.E., Strauss, J.K. and Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2), pp.821-862.

Rapach, D.E., Strauss, J.K. and Zhou, G., 2013. International stock return predictability: what is the role of the United States?. *The Journal of Finance*, 68(4), pp.1633-1662.

Russell S, Norvig P. Artificial intelligence: a modern approach[J]. 2002.

Shiller, R.J., 2015. *Irrational exuberance: Revised and expanded third edition*. Princeton university press.

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), pp.1139-1168.

Tetlock, P.C., 2011. All the news that's fit to reprint: Do investors react to stale information?. *The Review of Financial Studies*, *24*(5), pp.1481-1512.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*(3), pp.1437-1467.

Van de Kauter, M., Breesch, D. and Hoste, V., 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications*, *42*(11), pp.4999-5010.

Wang, Z., Zhang, J., Feng, J. and Chen, Z., 2014, June. Knowledge graph embedding by translating on hyperplanes. In Twenty-Eighth AAAI conference on artificial intelligence.

Welch, I. and Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4), 1455-1508.

Zeng-Lin, X.U., Yong-Pan, S., Li-Rong, H.E. and Ya-fang, W.A.N.G., 2016. Review on knowledge graph techniques. *Journal of University of Electronic Science and Technology of China*, 45(4), pp.589-606.

Zhan, Q. and Yin, H., 2018, March. A loan application fraud detection method based on knowledge graph and neural network. In Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence pp. 111-115.

Zhang, L., Wang, S. and Liu, B., 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), p.e1253.

Appendix

Table A.1. Summary of Statistics

This table presents the summary statistics for log excess market return for the DJIA (*ret_DJIA*), risk-free rate (R^f), the original sentiment score from the Business section of NYT (*Sent_news*), the DJIA specific sentiment index (*Sent_DJIA*), investor sentiment index (*Sent_BW*) proposed by Baker and Wurgler (2006), dividend premium (*PDND*, log difference of the weighted average *BM* ratios of dividend paying stocks and nonpaying ones), monthly number of IPOs (*NIPO*), closed-end fund discount rate (*CEFD*, weighted average difference between the share value and net total asset value of closed-end funds), equity shares in new issues (*EQTI*, the value ratio of equity issuance divided by the gross sum of equity and debt issuance) ,dividends (*D12*, 12-month moving sum of dividend payments), earnings (*E12*, 12-month moving sum of earnings), book-to-market ratio (*BM*, BM ratio for DJIA), default yield spread (*DFY*, difference between the yields of AAA and BAA rated corporate bonds), long term corporate bond rate of return (*CORPR*), long-term government bond yield (*LTY*), long term of rate of return(*LTR*), net equity expansion (*NTIS*, 12-month moving sum of new issuance from NYSE listed stocks divided by the end-of-year total capitalization of *NYSE*), inflation rate (*INFL*), and stock variance (*SVAR*, sum of squared daily returns). $\rho(1)$ is the lag 1 autocorrelation coefficient. All statistics are calculated based on the original data series.

	Mean	Stdev	Skew	Kurt	Min	Max	$\rho(1)$
ret_DJIA	0.003	0.042	-0.789	1.532	-0.152	0.091	0.138
Sent_news	-449.706	321.750	-1.156	0.554	-1509.597	-104.574	0.948
Sent_DJIA	0.000	1.000	-2.302	4.616	-3.669	0.714	0.902
Sent_BW	-0.212	0.292	0.110	0.410	-0.900	0.600	0.945
PDND	-4.793	4.661	0.949	1.159	-13.730	10.180	0.886
NIPO	14.725	9.191	0.429	-0.434	0.000	39.000	0.621
CEFD	8.615	4.162	-1.338	2.378	-6.020	18.230	0.920
EQTI	0.135	0.053	0.742	-0.181	0.050	0.270	0.977
D12	31.290	7.406	0.582	-1.015	21.904	45.701	0.999
E12	75.965	25.908	-1.472	1.340	6.860	105.960	0.993
BM	0.319	0.048	-0.248	-0.309	0.216	0.441	0.868
Rf	0.001	0.001	2.380	4.383	0.000	0.004	0.993
DFY	-0.012	0.005	-2.395	6.099	-0.034	-0.006	0.957
LTY	0.034	0.009	0.097	-1.250	0.018	0.052	0.959

NTIS	-0.009	0.020	-0.428	-0.499	-0.058	0.027	0.970
INFL	0.002	0.004	-1.178	4.868	-0.019	0.010	0.572
LTR	0.006	0.035	0.415	2.360	-0.112	0.144	0.010
CORPR	0.006	0.032	0.777	5.004	-0.095	0.156	0.070
SVAR	0.004	0.007	5.228	33.201	0.000	0.058	0.712

Table A.2. Correlation Matrix

This table shows the correlation coefficients between all predictors. "ret" is the log excess return of DJIA. "S_news" is the unmodified sentiment index and "S_DJIA" is the KG modified index, S_BW is the Baker and Wurgler (2006) sentiment index. The pairwise correlation for a given predictor excludes the predictor's correlation with itself.

	ret	S_news	S_DJIA	S_BW	PDND	NIPO	CEFD	EQTI
ret								
S_news	0.31							
S_DJIA	0.37	0.90						
S_BW	-0.11	0.04	0.07					
PDND	-0.17	-0.57	-0.69	-0.40				
NIPO	0.13	0.41	0.49	0.29	-0.50			
CEFD	0.03	0.21	-0.03	-0.74	0.35	-0.21		
EQTI	-0.05	-0.41	-0.31	-0.69	0.46	-0.26	0.40	
D12	0.01	0.57	0.25	0.16	-0.16	0.07	0.36	-0.52
E12	0.10	0.78	0.72	0.49	-0.65	0.52	-0.16	-0.63
BM	-0.19	0.29	0.20	-0.17	-0.03	0.11	0.36	0.22
Rf	-0.08	-0.38	-0.19	0.66	-0.25	0.05	-0.82	-0.48
DFY	0.20	0.63	0.74	0.40	-0.67	0.58	-0.38	-0.42
LTY	-0.09	-0.71	-0.46	0.17	0.10	-0.03	-0.59	0.25
NTIS	0.21	0.46	0.62	-0.16	-0.43	0.49	0.01	0.27
INFL	0.09	0.03	0.08	0.19	-0.23	0.12	-0.41	-0.20
LTR	-0.26	-0.05	0.01	0.01	0.08	-0.06	0.05	0.06
CORPR	0.08	0.05	0.14	-0.13	0.07	-0.08	0.14	0.14
SVAR	-0.43	-0.60	-0.68	-0.15	0.56	-0.40	0.24	0.34

CORPR	LTR	INFL	NTIS	LTY	DFY	Rf	BM	E12	D12
									0.50
								0.28	0.03
							-0.52	-0.04	-0.24
						0.10	-0.07	0.77	0.21
					-0.11	0.61	-0.46	-0.49	-0.69
				-0.02	0.52	-0.31	0.36	0.37	-0.20
			0.04	0.22	0.26	0.22	-0.17	0.04	-0.13
		-0.29	0.04	-0.10	-0.06	0.00	0.14	-0.01	-0.06
	0.74	-0.34	0.10	-0.14	-0.15	-0.10	0.10	-0.10	-0.07
0.05	0.22	-0.46	-0.38	0.18	-0.65	-0.03	0.10	-0.46	-0.16