Determining Lead-Lag Structure Between Sentiment Index and Stock Price Returns

Alex Momotov[1], *Xianghua* Xie[2,*]

[1-2] Department of Computer Science, Swansea University, UK, http://csvision.swan.ac.uk

Abstract. Sentiment mining has a long history of empirical research, with a number of applications in analysing financial markets, social media trends, and company-specific news. To this end NLP approaches to sentiment extraction can be categorised into lexicon-based and classification algorithms. The former provides a mapping between lexical items and their semantic orientation, and was shown to be less sensitive to the issues of concept drift and domain-specificity. On the contrary, classification approaches often learn corpus-specific attributes that are unique to the context of a given discipline, but result in a greater accuracy within the same context. This research contrasts and compares the e of the art techniques of the two approaches within the domain of news sentiment analysis, as well as, investigates a novel document encoding representation of the `TF-IDF momentum matrix'. The presented lexicon-based methodology is centred around Loughran & McDonald financial sentiment word lists and reaches 86.4% explained stock momentum variance, whereas the classification approach follows a thematic analysis pipeline implementing Latent Dirichlet Allocation and achieves that of 94.8%. As an additional element of model evaluation, the research implements Thermal Optimal Path method which relies on a dynamic programming approach for performance optimisation. The technique originates in statistical physics and serves to uncover a lead-lag relationship between sentiment signal and stock price momentum time series data.

1 Introduction

Sentiment mining algorithms have long been of interest to researchers, traders, and the wider NLP community. As such, research literature distinguishes two fundamental strands of methodology for analysing document tone - dictionary-based, and the machine learning classification-focused approaches [2, 19, 22]. The former group of algorithms is also known under the generic term of lexicon-based approaches, and operates by providing a database-like mapping between individual words and their semantic score. Sentiment dictionaries have become one of the widely applied techniques, with applications including opinion extraction, news sentiment analysis, document tone classification, and stock price prediction. On the contrary, classifier-centred group of methods includes algorithms such as Naive Bayes, SVM, logistic regression, and neural networks, where the model is selected in conjunction

* Corresponding author: x.xie@swansea.ac.uk

with a document-term representation matrix. The methodology of document-encoding representations, in turn, has expanded from bag-of-words to TF-IDF matrices, and more recently to thematic modelling techniques of Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA).

1.1 Lexicon-Based Sentiment Analysis

The advantages of classifier-based methods are well known and widely documented throughout the computational linguistics literature, however, the traditional lexicon techniques are still relevant and yield more generalisable results when the sentiment mining task is unsupervised. The literature brings two explanations for this phenomena - the issues of domain-specificity and the lack of context. Domain specificity refers to the behaviour when classifier-based techniques perform well within a given document collection but fail to generalise their performance onto contexts with new vocabulary meanings. This happens because classifiers learn domain-specific attributes which are likely to be of different or opposite sentiment polarity when the context is changed. For example, [22] report that Support Vector Machines and the machine learning-based approaches in general tend to outperform lexicon-based techniques within a single domain, however their accuracy has been noted to drop dramatically when applied to texts different in context to the original corpus [22]. Secondly, `lack of context' refers to the loss of semantic information contained in word order during the construction of TF-IDF matrix. In practice, n-gram co-location detection rarely captures negating, amplifying and downtoning token combinations and in scenarios such as `quite well', `not well', `extremely well', the token `well' will carry the same semantic polarity and magnitude [2, 22].

Sentiment lexicon approaches are not subject to the `domain specificity' and `lack of context' drawbacks because they operate under the assumption of *prior polarity*. That is, each word within a text has a sentiment orientation (SO) and magnitude that are independent of the context and that these can be represented as scalar values [22]. Computational linguistics literature discusses several ways to realise the idea of the dictionary-based model in practice, where the distinction between generic and proprietary implementations can be outlined as follows. General purpose lexicons, such as Diction and VADER are created with no particular application context in mind, where each token's SO is computed based on the word's meaning as per its most frequent usage in the generic domain. On the opposite, domain-specific dictionaries are built using word senses as per their most probable usage in a specific discipline and, thereby, known to result in a substantial performance improvement [15]. The phenomena is explained by a large presence of polysemes (words with multiple meanings) in English language. In fact, it can be argued that as a document corpus becomes more discipline-specific, the sentiment score derived from non-proprietary dictionaries becomes increasingly unreliable [15, 16, 17].

This brings us to discuss our rationale behind the choice of Loughran & McDonald financial dictionary as part of our lexicon-based sentiment mining model. The task of mining financial news sentiment constitutes the scenario where the document corpus is highly discipline specific. This calls for a need of a finance-oriented dictionary, where, to the best of our knowledge, the Loughran & McDonald (LM) word lists introduced in [15] are the state of the art proprietary sentiment lexicon in the field. The resource has been compiled using 10Ks reports from US companies, where the authors manually labelled each item according to its most likely sentiment value in the context of business news and corporate developments. Based on this, we build the central dictionary-based sentiment analysis algorithm of this study around negative LM word lists.

1.2 LDA-Based Sentiment Analysis

From the early 2000s the field of computational linguistics has moved from document modelling techniques such as BOW and TF-IDF matrices to topic and thematic modelling approaches. Latent Semantic Analysis (LSA) and Latent Semantic Indexing (LSI) were early developments within topic modelling which applied singular value decomposition on sentence-level term matrix and have been extensively experimented with for text summarisation [3, 11]. Following LSA the research area focused on Bayesian and probabilistic topic models which gave rise to Latent Dirichlet Allocation (LDA) which is considered the current state of the art approach in text modelling [3]. LDA is a generative probabilistic algorithm which aims to model a corpus of texts, or otherwise discrete data, and was first introduced to the literature in [7]. The model was designed to overcome the shortcomings of the TF-IDF scheme and can, similarly, be used for summarisation, classification, similarity and distance reporting [6, 7]. LDA operates under a series of assumptions about how each document was initially generated. The generative process has two main stages where discrete data corpus is viewed as a randomly selected mixture of latent topics, where each topic is, in itself, a probabilistic distribution over a fixed thematic vocabulary [7].

The types of feature-vectors constructable from LDA can represent a weighted allocation of terms where the weight reflects importance of the word to each existing topic, or a probabilistic distribution of k topics across a single document, or a multidimensional matrix expressing both [5, 8]. For example, [8] trains Latent Dirichlet Allocation on 7645 press releases from 499 German market companies to extract 40 topics across the corpus, then proceeds to assign each document its predominant theme based on individual thematic probability distributions θ_d. As a result, [8] have drawn a corpus-level statistical summary of thematic allocations for eight company sectors, as well as, a ranking of LDA topics by their covariance with non-zero abnormal returns. While their exploration did not go beyond examining the impact of topics within company announcements on the stock market's median abnormal returns - we find the central avenue of their research very applicable to our LDA-based sentiment mining model, and therefore, choose to structure our `bottom-up' part of methodology around theirs.

LDA-constructed feature vector can also represent a higher level statistic on the number of times each topic occurs in the collection of articles during a certain time interval. For instance, in their study [5] use Latent Dirichlet Allocation to construct this type of document-vector representation to classify sentiment of news article collection from Thomson Reuters archive and predict directional changes in price and volatility for a range of US stocks. The authors construct their feature space by combining 100 topics extracted from article bodies and 20 topics from titles into a single 120 dimensional vector. The feature vector thus represents the number of times each of the 120 themes appeared in the NER-extracted news corpus during a single one hour period before the event window. The principal result of Atkins et al. (2018) methodology is that while the Naive Bayes classifier trained on news topic distributions did not show any predictive power for the asset close price (accuracy 49%) it was more effective at forecasting secondary time-series attributes - namely volatility at 56% accuracy [5].

1.3 Thermal Optimal Path

Determining the direction of causality between two time-series variables is a problem with long history [20]. One way of detecting such a relationship is by finding a global maxima of the lagged cross-correlation function, where the information of the first variable $X(t)$ is related to the future value of the second variable $Y(t + k)$ for $k \neq 0$ [22]. The method, however, cannot

give us confidence to conclude causality between the two time-series, since both of them may be a result of a common phenomena external to the model [20]. A more pragmatic method, known as Granger causality, is an extension of lagged correlation and works by testing for predictability between X and Y [20, 22]. That is, if the knowledge of X at time t is able to improve predictability of Y after a time period $t + k$ for $k > 0$, then it can be stated that "X Granger causes Y" [20][p. 578], [4, 9]. Although useful in practice, the Granger method still does not offer explanation of the real causal behaviour, as it lacks information about the forces operating and influencing the model from outside - what [23][p. 288] call addressing fundamental `epistemological question'. Based on this premise, the researchers introduce a novel model termed `optimal thermal causal path', also known as, `thermal optimal path' (TOP) which belongs to the same category as Granger causality method, yet was designed to provide more functional utility in uncovering lag relationships between two time-series data [20, 22, 23].

The underlying assumption behind the TOP method is allowing for the possibility that the lead-lag structure between two variables may change dynamically, as opposed to being a constant temporal interval k, and that it can be expressed as a function of time [20]. In other words, the method primarily addresses the case where the dependence between $X(t)$ and $Y(t + k)$ is expressed with $k = f(t)$, such that the solution to discovering the causality relationship becomes identical to finding the function $f()$. Assuming the timelines of the two variables can be denoted with discrete units of time t_1 and t_2, the first step towards implementing thermal optimal path method is to calculate the distance matrix between $X(t_1)$ and $Y(t_2)$ [20]. The distance matrix is denoted as $E_{X,Y}$ and contains distance values between the realisations of X and Y at every possible pairwise combination of t_1 and t_2 [20].

Generalising this idea, the TOP method aims to find a set of values wandering alongside the main diagonal of the distance matrix, such that their global summation is minimised, in turn, defining the shape of the lead-lag structure between the two time series [20]. If we define the main diagonal of the matrix as t, where $t_1 = t_2$ i.e. $k = 0$, then intuitively, one could search for locally minimal values of all lines perpendicular to t as a solution to the global minimisation mapping. However, this idea proves to be problematic, in that, such a mapping would result in large jumps between consecutive values of t, as well as, a potential for one-to-many relationship, breaking the definition of a function [20]. Therefore, as a preliminary calculation to the thermal optimal path the research defines *energy landscape* - a derivative from the distance matrix where each entry reflects minimum energy path taken from the matrix origin $(t_1, t_2 = 0)$ to the coordinate pair under the constraint of continuity in t [20]. Consequently, *energy* is calculated as a cumulative sum of all distance entries within $E_{X,Y}$ encountered along the minimum path. In this case the global minimum path is also termed `optimal path at zero temperature' and travels from the origin of the energy landscape to $t_1, t_2 = N - 1$ such that the summation of its values is minimised and where the deviation from the main diagonal t defines the lead-lag relationship between X and Y [20][p. 579].

Although achieving the goal of determining the causal structure, zero temperature path has one major drawback. That is, the absolute minimum direction within the energy landscape is sensitive to random structural realisations and noise present between X and Y to an extent that small changes in the time-series values lead to large fluctuations of the thermal path [20]. This issue of sensitivity to noise is well documented in statistical physics research, such as [13] and constitutes the central motivation behind the development of TOP method [20]. Building on the concept of global optimisation problem within the energy landscape, [20] put forward the `finite temperature' model and allow deviations from the absolute minimum path to a degree determined by the `temperature' parameter. That is, the new optimal path now represents probability of a certain path configuration that is different from the absolute minimum, and is expressed with Boltzmann weight as $\exp[-\Delta E/T]$, where the temperature t determines the allowed deviation from the minimum energy [20][p. 582]. As

the temperature parameter approaches zero, the optimal path configuration matches closer to that of zero temperature path. Likewise, as *t* increases, the finite path reflects a greater sample of probabilities around the minimum and results in a smoother structure that is less sensitive to noise and random data realisations [22, 23].

2 Methodology

As an initial stage, we build a web scraping algorithm and gather a corpus of financial news articles spanning across an 8-year period between 2010 and 2018 from US-based Thomson Reuters news archive. The stage involves an element of named entity recognition in order to identify news stories relating to the selected stock assets and groups the gathered document samples into stock specific corpora. We proceed to acquire historical daily closing stock price for each identified asset and pre-process the time-series data using differencing, simple moving average, and standartisation. This is followed by implementing a lexicon-based approach centred around the percentage of negative words and smoothed across a different set of lengths in order to investigate long, medium, and short-run relationships between sentiment and stock price. Similarly, the LDA-based model, then evaluates such relationship using a thematic distribution momentum vector and topic ranking using least mean squares regression. Finally, we rely on dynamic programming to implement Thermal Optimal Path method which, in turn, helps to describe the lead-lag relationship between sentiment and stock price momentum signals.

2.1 News Data Collection

Following the footsteps of [5] we have chosen Thomson Reuters US news archive as our primary source of news data. We focus on the US strand of the archive and how news present therein relate to composite US index of Dow Jones Industrial Average (DJIA), as well as, individual firm stock indices of Apple, Amazon, Google and Microsoft. We incorporate a news scraping pipeline using *requests* and *BeautifulSoup* and gather a total of 8.5 years news interval between the dates of 1-1-2010 and 20-6-2018. Altogether, we collect 6,862,707 news articles, out of which 3,908,333 were empty due to the `Page Not Found' error and another 216,968 contained empty main body. These were discarded, resulting in a total of 2,737,406 news publications that are valid and used for all subsequent phases in this work. Our rationale for collecting a large dataset, is to enable a later calculation of a long-term moving average statistic on the sentiment index, as well as, allow for a narrower NER search, as justified by the need for highly relevant company-specific news.

2.2 Financial Data

We use daily asset closing price (as traded in NASDAQ stock exchange market) as the primary financial time-series data. We collect the historical prices for Amazon, Microsoft, Apple, and Google from Yahoo Finance resource and gather the closing values for the of Dow Jones composite index from Thomson Reuters Datastream outlet. The original close price data is not stationary and hence provides no information about short term trends and changes. Following the traditional preprocessing methodology of the financial and NLP fields we, first, transform the price time-series using the percentage differencing, calculated as follows:

$$\frac{p(t) - p(t-1)}{p(t)} \tag{1}$$

Where, *p()* is the price of a given stock at a discrete unit of time *t* representing a single day. Next, in order to represent a long-term trend within the time-series, we calculate an *n*-day simple moving average (SMA), according to the equation below:

$$SMA = \frac{p(t)+p(t-1)+\cdots p(t-(n-1))}{n} \quad (2)$$

Where, *n* is the number of days across which the returns average is taken, and is followed by discarding the first *n* days from the data series - as these values were used for the SMA calculation. Applied together, differencing and SMA transformations comprise a `momentum' statistic which represents the speed of change in close price data. Following this, we carry out financial dataset standardisation by subtracting mean from each momentum value, then dividing it by standard deviation.

2.3 Lexicon-Based Sentiment Analysis

In our survey of literature we identified Loughran & McDonald (LM) financial dictionary as the state of the art domain-specific lexicon in the field and, hence, use the LM negative word list as a primary semantic lexicon for our investigation. The dictionary provides word inflections as an alternative to the traditional stemming procedure, and so our text preprocessing only involves tokenisation.

An initial exploratory data analysis has lead us to develop a document filtering algorithm, similar to the Luhn's cut-off principle discussed in [13, 18] as a strategy to dealing with LM percentage outliers. That is, we remove document samples with greater than a certain LM word percentage, as well as, a fixed LM word count, where these threshold values are determined experimentally. This step has shown to improve the percentage of explained variance by approximately 5% for the DJIA dataset and around 2.5% for other firm-specific regression results and is a technique we adopt for all the subsequent tests.

To derive sentiment index, we employ a sequence of calculations which is much similar to that of the financial data transformation, and outputs a stationary daily sentiment time-series. As a first step we compute an average percentage of Loughran & McDonald keywords across all news articles on a given day, and vectorise the resulting data, such that each timeline instance is now associated with a single LM percentage score. This is followed by simple moving average, and standardisation by mean and standard deviation. These steps enable correlation and regression between stock momentum and sentiment index (which is now also sentiment momentum) as the both time-series are mean-centred with one standard deviation at 1 and -1.

As a second stage of the lexicon-based methodology, we implement ranking of individual Loughran & McDonald words by their strength of regression with stock index momentum. The bag of words representation no longer sufficiently reflects importance of token frequencies for this purpose, and thus, the first step is generating a TF-IDF document encoding matrix. Analogous to the BOW construction, we develop a custom algorithm which builds and returns a Numpy memory-map of TF-IDF matrix with the advantage of constant memory complexity, according to the following:

$$TFIDF_{i,j} = TF_{i,j} \times log\left(\frac{n}{DF_i}\right) \quad (3)$$

Where, $TF_{i,j}$ denotes frequency of term *i* in document *j* and is already provided within the bag of words, *n* is the total number of documents in the dataset, and DF_i is a corpus-wide number of texts that contain the term *i* [1]. As an additional step, we normalise TF-IDF rows by their length, such that the resulting feature vectors are in their unit form with vector magnitudes equal to 1. Following TF-IDF construction, we apply our sequence of transformations to

generate sentiment index to each matrix column individually. That is, we average all TF-IDF scores across all feature vectors with a common date of publication, calculate an *n*-day simple moving average, discard an *m* number of first timeline instances and standardise the matrix entries by mean and standard deviation. The resultant matrix can be interpreted as a momentum of unit vector TF-IDF LM keyword percentage scores, or in other words, how the mean of importance and usage frequency of individual Loughran & McDonald negative words changes over the course of the timeline.

The goal of this methodological phase is to see which words are most informative to the overall sentiment score derivation in our main lexicon-based model. Therefore, following the same rationale, we choose the least mean squares linear regression as a primary ranking tool for this purpose. We proceed to pair each TF-IDF momentum matrix column with financial time-series vectors using common date values and fit the linear regression model for each vector pair. As a result, we obtain a set of significance statistics (R^2, adjusted R^2, *F*-statistic, *t*-statistic, *p*-value) for the momentum TF-IDF score of each negative LM word. This enables us to filter word-regressors with significant relationships (p-value < 0.05) and rank them by the percentage of explained variance, where we report such ranking tables in the results section.

2.4 LDA-Based Sentiment Analysis

We have discussed a rationale behind why financial news data requires an alternative treatment to conventional LDA-based sentiment mining and that utilising price data as an element of supervision for LDA has been the focus of recent NLP research [3, 5, 8]. To the best of our knowledge, this goal has no de-facto methodology, and constitutes the current arena of empirical investigation within the field. A primary research avenue in this direction, however, can be outlined as iterating over extracted latent topics, while running an external statistical model on a single theme at a time, with the goal of learning most informative topics, as well as, most significant thematic keywords present therein. The external model evaluates each theme's distribution against a dependent variable of interest, such as stock returns signal, and enables the researcher to rank themes by their discriminative power.

Starting with the original asset-specific subsets of news data we pre-process documents using word-level text segmentation, removing non-alphanumeric characters, and removing stopwords with NLTK - the cleaning steps from the earlier `top-down' methodology. This time, however, we incorporate two additional phases of stemming and removing words shorter than 3 characters, where our stemming algorithm is NLTK's *SnowballStemmer* - an improved version of the popular Porter stemmer [5]. This is justified by no longer having available a dictionary with word inflections, leading to a need of minimising the resultant lexicon length manually. Further to this, we follow [5] and set Luhn's cut-off intervals as 15 documents for the lower cut-off and 50% for the upper cut-off which shorten the lexicon length further. That is, the lower cut-off is a raw document count where terms appearing in fewer that this amount of texts are discarded, while the upper cut-off is a document percentage.

Using *Gensim* library we generate BOW and TF-IDF matrices, then proceed to train a multi-core and memory efficient implementation of LDA, called `Online Latent Dirichlet Allocation'. Earlier we discussed how there exist several ways of constructing a feature vector from the LDA-extracted thematic distribution. For example, the vector may represent a keyword distribution of a single topic of interest for each document; an overall percentage composition of *N* topics on a single day; or a one-hot representation of which topics appeared in the news during a certain time interval preceding a price fluctuation [5, 8]. For the purposes of comparability to the stock price momentum signal, we use a similar sequence of calculations as for transforming the close price time-series data. Stating with the thematic

distribution vector which encodes the probabilistic topical composition for each document, we calculate column mean of all data points with the same publication date. The resultant vector reflects an average presence of each theme on a given date of news publication, while still remaining unit length. Next, we calculate an *n*-day simple moving average, where *n* matches that of the financial signal's SMA and is also a theme-individual computation. Lastly, we discard the first *m* timeline dates, standardise each theme by its mean and standard deviation, and synchronise the data with the stock price time-series. The final matrix can be interpreted as containing the momentum signal for each individual topic, or in other words, how a smoothed presence of each latent topic in news changes over the course of the publication timeline.

Similarly to how we raked individual financial dictionary keywords by their explained variance we now implement such ranking statistic for LDA topics. To this end, we run least mean squares linear regression on each column of the `thematic momentum matrix' against the stock price momentum data using Python's scientific library *statsmodels.api.sm.OLS()*. We filter out signals that are not significant ($p > 0.05$), sort themes by their percentage of explained variance (R^2), and report the ranking tables in the results section. As a second element of evaluation, we use our thermal optimal path methodology to find zero and finite temperature paths between each of the topic momentums and the financial time-series. Based on their lead-lag structure, this step helps us understand which themes are more likely to precede stock price movement and we report such structures for a few topics with the highest R^2 value. Finally, we fit multiple linear regression using all theme momentum series. We report on significance coefficients of each fit, as well as, if the percentage of explained variance is higher than that derived from the lexicon-based approach.

2.5 Thermal Optimal Path

We have touched upon the thermal optimal path method as a tool for discovering lead-lag relationship between two time-series variables in our survey of literature, and will now discuss the implementation. The two central assumptions behind the TOP method are that the lag of one variable behind another changes dynamically, and that a measure of this lag can be expressed as a function of time. In other words, where lagged cross-correlation finds *k* such that *X(t)* and *Y(t + k)* are maximally close, the thermal path method attempts to solve the problem by finding a function *f()*, such that *k = f(t)*. Assuming t_1 and t_2 are discrete time variables, the first step towards implementing thermal optimal path method is to calculate the distance matrix between $X(t_1)$ and $Y(t_2)$ [20]. The distance matrix is denoted as $E_{X,Y}$ and contains difference values between the realisations of *X* and *Y* at every possible pairwise combination of t_1 and t_2 [20]. Each entry within $E_{X,Y}$ is computed according to the following distance measure:

$$\varepsilon(t_1, t_2) = |X(t_1) - Y(t_2)|^q \tag{4}$$

Where, *q* controls how much weight is placed on discrepancies and following the TOP convention described in [20] we set $q = 2$. The main diagonal *t* of the distance matrix is defined as $t_1 = t_2$, where the energy $\varepsilon(t_1, t_2)$ between *X* and *Y* is zero when the two time-series are equal $X(t) = Y(t)$. Then, *x* is defined as an axis perpendicular to the main diagonal, and at any point quantifies a deviation from *t* [20]. It is, then, useful to construct a rotated coordinate scheme, as per the transfer matrix method in [20] with the following equation pair:

$$\begin{cases} x = t_2 - t_1 \\ t = t_2 + t_1 \end{cases} \tag{5}$$

By this definition, when $x \neq 0$ there exists a deviation from $t_1 = t_2$ implying one time-series variable leads, and the other lags [23]. The TOP method, then tries to find a path starting at the origin $t_1 = 0$, $t_2 = 0$ and wandering alongside the main diagonal, such that to minimise the summation of distances $\varepsilon(t_1, t_2)$ within the path's realisation structure. A prerequisite to this, however, is the computation of an *energy landscape* matrix. Building on the earlier definition of the distance matrix entries $\varepsilon(t_1, t_2)$, the corresponding values of the energy landscape are denoted $E(t_1, t_2)$ and calculated as a cumulative distance (defined in (4)) of the optimal thermal path starting at the origin $(t_1 = 0, t_2 = 0)$ and leading to (t_1, t_2) [20]. The cumulative distance is, then, determined using the following relation between the two matrices, as described in [20][p. 581]:

$$E(t_1, t_2) = \varepsilon(t_1, t_2) + Min[E(t_1 - 1, t_2), E(t_1, t_2 - 1), E(t_1 - 1, t_2 - 1)] \quad (6)$$

Following this, the computation of a global minimum energy path is carried out starting from the energy landscape origin and considering a minimum of three subsequent entries $Min[E(t_1 - 1, t_2), E(t_1, t_2 - 1), E(t_1 - 1, t_2 - 1)]$. In this case the path is also termed `optimal path at zero temperature' and defines the lead-lag relationship between X and Y [20].

Although plausible in theory, the optimal path at zero temperature is highly sensitive to noise and to random structural realisations between X and Y [22]. Based on this, the original TOP research proposes a new method that now allows `thermal fluctuations' around the absolute minimum energy path, where the probability of observing such path decreases with its energy [20]. The probability of a new path configuration with energy ΔE above the energy of a zero temperature path is proportionate to the Boltzmann weight $exp[-\Delta E/T]$. Where, the temperature t determines the allowed deviation from the minimum energy and is the only parameter in the method [20]. Following the optimal choice of temperature parameter as described in [10, 12] we set $T = 2$ for all stock-specific datasets in our work.

The new optimal path obtained from the above method is now termed *optimal path at finite temperature*, where we describe its implementation as follows. First, the principal equation defining realisation of path deviation x at position t is given by [20]:

$$\langle x(t) \rangle = \sum_x x G(x, t) / G(t) \quad (7)$$

Where $\langle x(t) \rangle$ denotes expectation of the path position at time t, $G(x,t)$ is the partition function, and $G(t)$ is a total partition function. The partition and total partition functions are, in turn, defined by the following equations from [20][p. 590]:

$$G(x, t + 1) = [G(x - 1, t) + G(x + 1, t) + G(x, t - 1)] e^{-\varepsilon(x,t)/T} \quad (8)$$

$$G(t) = \sum_x G(x, t) \quad (9)$$

We have described the mathematical relations involved in TOP pipeline computation and would like to touch upon its implementation in practice. In total, the methodology involves derivation of four matrices - distance matrix, energy landscape, partition matrix, and the total partition matrix, where the matrices coordinates are derived using neighbouring values. This leads to the need of computing matrix coordinates repeatedly and calls for the use of a dynamic programming approach. For instance, instead of calculating $E(t_1 - 1, t_2)$, $E(t_1, t_2 - 1)$, and $E(t_1 - 1, t_2 - 1)$ within the energy landscape anew for each $E(t_1, t_2)$, a more efficient algorithm could store the intermediate matrix results and check whether the neighbouring matrix values are already available before proceeding with the calculation.

3 Results and Discussion

3.1 Lexicon-Based Model

Our first set of results concerns least mean squares linear regression statistic for each stock-specific news sentiment index and their respective stock price momentum signals. In order to examine long, medium and short-term relationships between the two, we use varying n - day momentum indices, where n determines the smoothing parameter in the SMA calculation. Table 1 provides a summary of LMS regression coefficients including percentage of explained variance (R^2), adjusted R^2, F-statistic, t-statistic, and a measure of significance determined by the p-value. As a first observation, all regression results are significant with the p-value below 0.05 and the t-statistic greater than 3. Secondly, all datasets resulted in R^2 values which increase with the value of n in the momentum calculation. That is, as we take a longer-term smoothing interval, the sentiment signal is able to explain a proportionately larger percentage of variance within the close price momentum data.

Table 1. Linear regression between stock and sentiment indices at varying momentum intervals.

	Momentum	180	240	300	365
DJIA	R^2	0.680	0.771	0.817	0.864
	Adj. R^2	0.679	0.771	0.817	0.864
	F-statistic	3203.	4891.	6224.	8452.
	t-statistic	56.592	69.933	78.892	91.933
	p-value	0.000	0.000	0.000	0.000
Amazon	R^2	0.323	0.453	0.629	0.730
	Adj. R^2	0.322	0.452	0.629	0.730
	F-statistic	546.2	906.1	1776.	2680.
	t-statistic	23.371	30.102	42.139	51.765
	p-value	0.000	0.000	0.000	0.000
Apple	R^2	0.174	0.328	0.414	0.509
	Adj. R^2	0.174	0.328	0.414	0.508
	F-statistic	267.7	590.9	812.3	1122.
	t-statistic	16.360	24.309	28.501	33.502
	p-value	0.000	0.000	0.000	0.000
Google	R^2	0.195	0.291	0.431	0.438
	Adj. R^2	0.195	0.291	0.430	0.438
	F-statistic	316.4	511.7	896.7	876.5
	t-statistic	17.787	22.620	29.946	29.607
	p-value	0.000	0.000	0.000	0.000
Microsoft	R^2	0.239	0.296	0.426	0.547
	Adj. R^2	0.239	0.296	0.426	0.547
	F-statistic	399.4	509.0	854.4	1312.
	t-statistic	19.984	22.562	29.229	36.218
	p-value	0.000	0.000	0.000	0.000

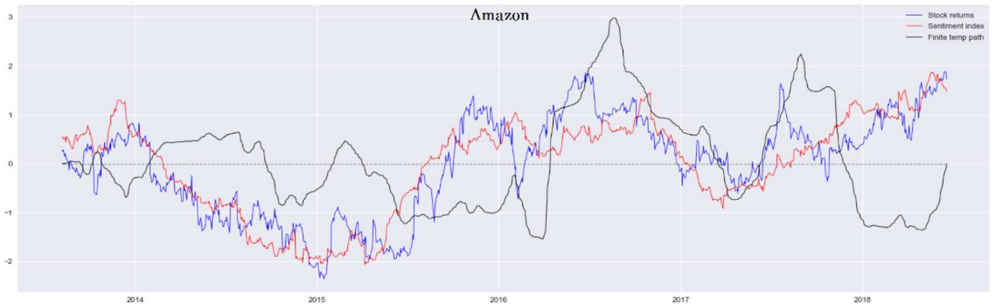

Figure 1. Thermal optimal path structure between sentiment and close price momentum indices.

The next set of results describes dynamic lead-lag relationships uncovered using the thermal optimal path method. Figure 1 shows the finite temperature path structure for the Amazon dataset with temperature $T = 2$ (black curve), where sentiment and close price momentum indices are indicated in red and blue, respectively. The x-axis travels alongside the length of the main diagonal within the distance matrix $E_{X,Y}$ and is equal to t as defined in the relation (5). Y-axis measures the deviation x from the main diagonal and quantifies the structural lead-lag relationship between sentiment and financial time-series data. Subsequently, any structural realisations of optimal thermal path at any given t that are above the horizontal line $x = 0$ indicate that the sentiment index leads stock price momentum. This, in turn, can be interpreted in favour of robustness of the sentiment mining model, since situations where the first variable X leads the second Y are a precursor to X causing Y, thereby increasing the potential for predictability. On the contrary, where the black curve travels below $x = 0$ indicates that the financial index leads, and sentiment signal lags behind. Therefore, an overall evaluation of the sentiment mining model is equivalent to analysing the proportion of the path realisation that is positioned above zero.

Table 2. LM negative dictionary word rankings by R^2 using LMS regression on individual TF-IDF momentum scores for Dow Jones news dataset.

Dow Jones Industrial Average				
Word	R^2	t-val	p-val	grad.
weakness	0.676	52.5	0.0	0.756
difficulty	0.601	44.7	0.0	0.715
cutbacks	0.594	44.1	0.0	0.733
disappointing	0.572	42.1	0.0	0.757
distorting	0.566	41.6	0.0	0.753
pervasive	0.552	40.4	0.0	1.952
deficient	0.551	40.3	0.0	0.728
misled	0.551	40.3	0.0	0.805
bans	0.529	38.6	0.0	0.681
easing	0.521	37.9	0.0	0.693
suspected	0.515	37.5	0.0	0.702
deterioration	0.512	37.3	0.0	0.675
challenges	0.502	36.6	0.0	0.709
bailout	0.502	36.5	0.0	0.851
violated	0.487	35.5	0.0	0.702
delay	0.487	35.4	0.0	0.675
deteriorated	0.484	35.3	0.0	1.704

We now turn to discuss the results of Loughran & McDonald financial dictionary word rankings by their R^2 values using individual linear regressions with TF-IDF momentum matrix. Table 2 summarises these results for the Dow Jones dataset, where the right-most column represents gradient of the best fit line as per the least mean squares regression and is positive for all keywords. The *p*-value was rounded to three decimal places, thus the absence of significant figures indicates the values are below 0.0005 and implies strongly significant regression tests in all cases. Similarly, the *t*-statistic is highly greater than 3 for all tests. The most striking observation, however is that the percentage of explained variance (R^2) for the highest ranking word-regressors is extremely high. This finding can be attributed to the fact that the sentiment index of the primary lexicon-based model was calculated using LM word percentage, which had no information about the weights of the dictionary words, or their significance within the corpus. Furthermore, we conclude that our primary sentiment mining model has a substantial potential for improvement if enhanced with the inverse-frequency normalisation that is present within the TF-IDF momentum matrix.

3.2 LDA-Based Model

The first group of results from the thematic modelling approach describes LDA topic rankings from individual topic momentum regressions against their asset's close price momentum time series. The ranking statistics are drawn for Dow Jones in table 3, where the summary displays extracted themes with significant regression relationships only ($p < 0.05$) and sorted by their R^2 scores. *Topic ID* represents theme number allocated by the LDA model, and *grad*. refers to the gradient of the best fit LMS regression line. The right-most column compiles three highest probability stemmed keywords in each topic.

Table 3. LDA topic ranking table for the Dow Jones news dataset.

Dow Jones Industrial Average					
Topic ID	R^2	*t*-val	*p*-val	grad.	Topic keywords
21	0.517	37.8	0.0	0.673	walgreen, unitedhealth, insur
*5	0.509	-37.1	0.0	-0.786	court, lawsuit, patent
4	0.486	35.5	0.0	0.67	text, coverag, eikon
7	0.465	34.0	0.0	0.768	india, indian, rupe
0	0.425	31.4	0.0	0.77	wto, attack, browser
2	0.365	-27.7	0.0	-0.723	appl, microsoft, verizon
29	0.34	-26.2	0.0	-0.544	allergan, gupta, rajaratnam
20	0.31	24.5	0.0	0.514	pfizer, drug, patient
19	0.239	-20.5	0.0	-0.544	goldman, bank, sach
10	0.229	19.9	0.0	0.664	peltz, trian, conf
17	0.228	-19.8	0.0	-0.451	index, dow, nasdaq
16	0.227	19.8	0.0	0.455	oct, china, hong
18	0.22	19.4	0.0	0.462	fitch, rate, ecuador
25	0.216	19.1	0.0	0.439	caterpillar, mine, equip
6	0.199	18.2	0.0	0.42	johnson, recal, xarelto

Since themes have been filtered for significance, their gradient coefficient can be interpreted as an element of sentiment polarity. That is, borrowing the idea from the lexicon-based approaches, where each word has a semantic orientation and magnitude, we can now infer each theme's semantic orientation using the *grad* column. For example, the second best topic in the DJIA table with topic ID 5 has a negative relationship (gradient = -0.786) with Dow Jones stock returns. This brings us to discuss the interpretation of the right-most column

displaying each theme's highest probability keywords. Even though the tokens have been stemmed during the data cleaning phase, their original meaning can be inferred, to an extent, with some words being more recognisable than others. Similarly, the most probable words relate to a clearly distinguishable themes in some cases, but appear quite randomly selected in others. For example, topic #5 contains stems `court', `lawsuit', `patent', `case', `district' which clearly indicate a topic about litigious company events. This compliments the interpretation of the gradient coefficient and suggests a negative relationship between lawsuit related news disclosures and company's stock returns. Observing the respective R^2 column, topic #5 also explains 50.9% of variance present within the momentum data. Treating the lexicon-based model as a baseline, this is a relatively large percentage.

Table 4. Multiple regression with a varying *n*-day momentum interval. F-statistic reflects the quality of the overall fit.

	Momentum	120	180	240	300	365
DJIA	R^2	0.798	0.883	0.902	0.913	0.946
	Adj. R^2	0.794	0.881	0.900	0.911	0.945
	F-statistic	203.5	373.8	437.4	478.6	758.3
	P. (F-statistic)	0.000	0.000	0.000	0.000	0.000
Amazon	R^2	0.756	0.840	0.853	0.874	0.917
	Adj. R^2	0.756	0.836	0.849	0.870	0.915
	F-statistic	122.4	198.6	210.1	238.8	361.5
	P. (F-statistic)	0.000	0.000	0.000	0.000	0.000
Apple	R^2	0.843	0.935	0.944	0.946	0.948
	Adj. R^2	0.839	0.933	0.942	0.945	0.947
	F-statistic	232.9	593.0	660.2	656.8	644.5
	P. (F-statistic)	0.000	0.000	0.000	0.000	0.000
Google	R^2	0.638	0.793	0.776	0.874	0.869
	Adj. R^2	0.630	0.788	0.770	0.871	0.865
	F-statistic	79.36	164.3	141.7	271.6	244.9
	P. (F-statistic)	0.000	0.000	0.000	0.000	0.000
Microsoft	R^2	0.820	0.879	0.900	0.929	0.922
	Adj. R^2	0.816	0.876	0.898	0.927	0.920
	F-statistic	199.0	304.3	359.3	492.5	422.0
	P. (F-statistic)	0.000	0.000	0.000	0.000	0.000

Our last series of results concern multiple linear regression fit using all 30 topics from the thematic momentum matrix. For the purpose of comparability to the lexicon-based model performance we report the regression tests for a varying length smoothing interval set by the *n*-day parameter including 120, 180, 240, 300, and 365 days in Table 4. Here, the *p*-value significance test is replaced by the probability of *F*-statistic, where the *F*-statistic itself reflects the quality of the overall fit.

As we can see, all regression tests are significant, where the probability of observing the given *F*-values randomly is below 0.0005. Similarly to our earlier observation regarding the lexicon-based model performance, the percentage of explained variance increases with the momentum interval. Following this, the regression tests with the longest momentum interval of 365 result in R^2 scores which are higher than the respective values from the lexicon-based model when using most effective LM word lists for each stock. In addition, the percentage of explained variance is greater than that of the dictionary approach for every respective pair of stock and *n*-day momentum interval. Consequently, when interpreting linear regression

coefficients as a primary reference for model evaluation, the LDA-based sentiment mining model outperforms the dictionary-based pipeline in all tested scenarios.

4 Conclusion

This research described an unsupervised methodology of mining sentiment from financial news releases using a combination of lexicon-based, LDA-driven, and external model evaluation techniques. In this section we will discuss ways to extend the developed models and the future research avenues. One research direction that we briefly touched upon was concerned with enhancing existing semantic dictionaries with the knowledge of coefficients from individual word regressions while attempting to maintain the position of generalisability to other corpora. This stems from our finding that a single keyword-regressor was able to explain up to 83.5% of stock momentum variance, outperforming the method of basing sentiment extraction on percentage of negative words in news by 32.6%. Simply assigning each dictionary item an LSM gradient-derived weight will violate the principle of an *uninformed* semantic lexicon and result in sentiment labels that are specific to the given stock. Therefore, as an attempt to extend discipline-specific sentiment dictionaries, research might explore methods to learn such weights as to reflect term importance in a maximally generic financial context. The second future research avenue concerns alternative ways of using close price data as an element of supervision for LDA. In our methodology we have discussed how due to the unsupervised nature of news sentiment mining the researcher needs to consider alternative supervision and evaluation strategies. We mentioned that one of such strategies involves uncovering a contrast between thematic distribution matrices, and would like to elaborate on this idea further. First, instead of using a strict *positive* and *negative* document class dichotomy - news texts released in a given event window preceding stock price fluctuation might be treated as merely having an increased likelihood of belonging to their respective class. Then, this could be followed by training two separate LDA models on the opposing class corpora, and computing the difference between their thematic model distributions. We hypothesised, that after discarding an intersection of word probabilities from all theme pairs, the resulting difference matrix may contain class discriminative information. In this light, future research might explore methods of generating a feature vector from the aforementioned matrix that would preserve attributes informative of document's sentiment.

References

[1] C. C. Aggarwal, Data mining: the textbook. Springer, 2015.
[2] S. Ahire, “A survey of sentiment lexicons,” 2014.
[3] M. Allahyari et al., “Text summarization techniques: A brief survey,” arXiv preprint arXiv:1707.02268, 2017.
[4] R. Ashley, C. W. Granger, and R. Schmalensee, “Advertising and aggregate consumption: An analysis of causality,” Econometrica: Journal of the Econometric Society, pp. 1149–1167, 1980.
[5] A. Atkins, M. Niranjan, and E. Gerding, “Financial news predicts stock market volatility better than close price,” The Journal of Finance and Data Science, 2018.
[6] D. M. Blei, “Probabilistic topic models,” Communications of the ACM, vol. 55, no. 4, pp. 77–84, 2012.
[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[8] S. Feuerriegel, A. Ratku, and D. Neumann, “Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation,” in System Sciences (HICSS), 2016 49th Hawaii International Conference on, 2016, pp. 1072–1081.
[9] J. Geweke, “Inference and causality in economic time series models,” Handbook of econometrics, vol. 2, pp. 1101–1144, 1984.
[10] C.-C. Gong, S.-D. Ji, L.-L. Su, S.-P. Li, and F. Ren, “The lead–lag relationship between stock index and stock index futures: A thermal optimal path method,” Physica A: Statistical Mechanics and its Applications, vol. 444, pp. 63–72, 2016.
[11] T. L. Griffiths and M. Steyvers, “A probabilistic approach to semantic representation,” in Proceedings of the Annual Meeting of the Cognitive Science Society, 2002, vol. 24, no. 24.
[12] K. Guo, Y. Sun, and X. Qian, “Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market,” Physica A: Statistical Mechanics and its Applications, vol. 469, pp. 390–396, 2017.
[13] T. Halpin-Healy and Y.-C. Zhang, “Kinetic roughening phenomena, stochastic growth, directed polymers and all that. Aspects of multidisciplinary statistical mechanics,” Physics reports, vol. 254, no. 4–6, pp. 215–414, 1995.
[14] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.
[15] T. Loughran and B. McDonald, “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks,” The Journal of Finance, vol. 66, no. 1, pp. 35–65, 2011.
[16] T. Loughran and B. McDonald, “The use of word lists in textual analysis,” Journal of Behavioral Finance, vol. 16, no. 1, pp. 1–11, 2015.
[17] T. Loughran and B. McDonald, “Textual analysis in accounting and finance: A survey,” Journal of Accounting Research, vol. 54, no. 4, pp. 1187–1230, 2016.
[18] H. P. Luhn, “The automatic creation of literature abstracts,” IBM Journal of research and development, vol. 2, no. 2, pp. 159–165, 1958.
[19] S. Sohangir, N. Petty, and D. Wang, “Financial Sentiment Lexicon Analysis,” in Semantic Computing (ICSC), 2018 IEEE 12th International Conference on, 2018, pp. 286–289.
[20] D. Sornette and W.-X. Zhou, “Non-parametric determination of real-time lag structure between two time series: the ‘optimal thermal causal path’method,” Quantitative Finance, vol. 5, no. 6, pp. 577–591, 2005.
[21] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” Computational linguistics, vol. 37, no. 2, pp. 267–307, 2011.
[22] W.-X. Zhou and D. Sornette, “Non-parametric determination of real-time lag structure between two time series: The ‘optimal thermal causal path’ method with applications to economic data,” Journal of Macroeconomics, vol. 28, no. 1, pp. 195–224, 2006.
[23] W.-X. Zhou and D. Sornette, “Lead-lag cross-sectional structure and detection of correlated–anticorrelated regime shifts: Application to the volatilities of inflation and economic growth rates,” Physica A: Statistical Mechanics and its Applications, vol. 380, pp. 287–296, 2007.