Semantic Frames to Predict Stock Price Movement

Boyi Xie, Rebecca J. Passonneau, Leon Wu
Center for Computational Learning Systems
Columbia University
New York, NY USA
(bx2109|becky|leon.wu)@columbia.edu

Germán G. Creamer
Howe School of Technology Management
Stevens Institute of Technology
Hoboken, NJ USA
gcreamer@stevens.edu

Abstract

Semantic frames are a rich linguistic resource. There has been much work on semantic frame parsers, but less that applies them to general NLP problems. We address a task to predict change in stock price from financial news. Semantic frames help to generalize from specific sentences to scenarios, and to detect the (positive or negative) roles of specific companies. We introduce a novel tree representation, and use it to train predictive models with tree kernels using support vector machines. Our experiments test multiple text representations on two binary classification tasks, change of price and polarity. Experiments show that features derived from semantic frame parsing have significantly better performance across years on the polarity task.

1 Introduction

A growing literature evaluates the financial effects of media on the market (Tetlock, 2007; Engelberg and Parsons, 2011). Recent work has applied NLP techniques to various financial media (conventional news, tweets) to detect sentiment in conventional news (Devitt and Ahmad, 2007; Haider and Mehrotra, 2011) or message boards (Chua et al., 2009), or discriminate expert from non-expert investors in financial tweets (Bar-Haim et al., 2011). With the exception of Bar-Haim et al. (2011), these NLP studies have relied on small corpora of hand-labeled data for training or evaluation, and the connection to market events is done indirectly through sentiment detection. We hypothesize that conventional news can be used to predict changes in the stock price of specific companies, and that the semantic features that best represent relevant aspects of the news vary across market sectors. To test this hypothesis, we use price information to label data from six years of financial news. Our experiments test several document representations for two binary classification tasks, change of price and polarity. Our main contribution is a novel tree representation based on semantic frame parses that performs significantly better than enriched bag-of-words vectors.

On Wednesday, April 11th, 2012, Google Inc announced its first quarterly earnings report, a week before the April 20 options contracts expiration in *contrast* to its history of reporting a day before monthly options expirations. The stock price of Google *surged* 3.85% from April 10th's $626.86 to 12th's $651.01. On Friday, April 13th, news reported Oracle Corp would *sue* Google Inc, claiming Google's Android operating system *tramples* its intellectual property rights. Jury selection was set for the next Monday. Google's stock price *tumbled* 4.06% on Friday, and continued to drop in the following week.

Figure 1: Summary of financial news items pertaining to *Google* in April, 2012.

Figure 1 shows a constructed example based on extracts from financial news about *Google* in April, 2012. It illustrates how a series of events reported in the news precedes and potentially predicts a large change in *Google*'s stock price. *Google*'s early announcement of quarterly earnings possibly presages trouble, and its stock price falls soon after reports of a legal action against *Google* by *Oracle*. To produce a coherent story, the original sentences were edited for Figure 1, but they are in the style of actual sentences from our dataset. Accurate detection of events and relations that might have an impact on stock price should benefit from document representation that captures sentiment in lexical items (e.g., *aggressive*) combined with the conceptual relations captured by FrameNet (Ruppenhofer and Rehbein, 2012). A frame is a lexical semantic representa-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 873–883,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

tion of the conceptual roles played by parts of a clause, and relates different lexical items (e.g., *report, announce*) to the same situation type. In the figure, some of the words that evoke frames have been underlined, and role fillers are outlined by boxes or ovals. Sentiment words are in italics.

To the best of our knowledge, this paper is the first to apply semantic frames in this domain. On the polarity task, the semantic frame features encoded as trees perform significantly better across years and sectors than bag-of-words vectors (BOW), and outperform BOW vectors enhanced with semantic frame features, and a supervised topic modeling approach. The results on the price change task show the same trend, but are not statistically significant, possibly due to the volatility of the market in 2007 and the following several years. Yet even modest predictive performance on both tasks could have an impact, as discussed below, if incorporated into financial models such as Rydberg and Shephard (2003). We first discuss the motivation and related work. Section 4 presents vector-based and tree-based features from semantic frame parses, and section 5 describes our dataset. The experimental design and results appear in the following section, followed by discussion and conclusions.

2 Motivation

Financial news is a rich vein for NLP applications to mine. Many news organizations that feature financial news, such as Reuters, the Wall Street Journal and Bloomberg, devote significant resources to the analysis of corporate news.

Much of the data that would support studies of a link between the news media and the market are publicly available. As pointed out by Tetlock et al. (2008), linguistic communication is a potentially important source of information about firms' fundamental values. Because very few stock market investors directly observe firms' production activities, they get most of their information secondhand. Their three main sources are analysts' forecasts, quantifiable publicly disclosed accounting variables, and descriptions of firms' current and future profit-generating activities. If analyst and accounting variables are incomplete or biased measures of firms' fundamental values, linguistic variables may have incremental explanatory power for firms' future earnings and returns.

Consider the following sentences:

Oracle sued Google in August 2010, saying Google's Android mobile operating system infringes its copyrights and patents for the Java programming language. (a)

Oracle has accused Google of violating its intellectual property rights to the Java programming language. (b)

Oracle has blamed Google and alleged that the latter has committed copyright infringement related to Java programming language held by Oracle. (c)

Oracle's Ellison says couldn't sway Google on Java. (d)

Sentences *a*, *b* and *c* are semantically similar, but lexically rather distinct: the shared words are the company names and *Java (programming language)*. Bag-of-Words (BOW) document representation is difficult to surpass for many document classification tasks, but cannot capture the degree of semantic similarity among these sentences. Methods that have proven successful for paraphrase detection (Deerwester et al., 1990; Dolan et al., 2004), as in the main clauses of *b* and *c*, include latent variable models that simultaneously capture the semantics of words and sentences, such as latent semantic analysis (LSA) or latent Dirichlet allocation (LDA). However, our task goes beyond paraphrase detection. The first three sentences all indicate an adversarial relation of *Oracle* to *Google* involving a negative judgement. It would be useful to capture the similarities among all three of these sentences, and to distinguish the role of each company (who is suing and who is being sued). Further, these three sentences potentially have a greater impact on market perception of *Google* in contrast to a sentence like *d*, that refers to the same conflict more indirectly, and whose main clause verb is *say*. We hypothesize that semantic frames can address these issues.

Most of the NLP literature on semantic frames addresses how to build robust semantic frame parsers, with intrinsic evaluation against gold standard parses. There have been few applications of semantic frame parsing for extrinsic tasks. To test for measurable benefits of semantic frame parsing, this paper poses the following questions:

1. Are semantic frames useful for document representation of financial news?
2. What aspects of frames are most useful?
3. What is the relative performance of document representation that relies on frames?

4. What improvements could be made to best exploit semantic frames?

Our work is not aimed at investment profit. Rather, we investigate whether computational linguistic methodologies can improve our understanding of a company's fundamental market value, and whether linguistic information derived from news produces a consistent enough result to benefit more comprehensive financial models.

3 Related Work

NLP has recently been applied to financial text for market analysis, primarily using bag-of-words (BOW) document representation. Luss and d'Aspremont (2008) use text classification to model price movements of financial assets on a per-day basis. They try to predict the direction of return, and abnormal returns, defined as an absolute return greater than a predefined threshold. Kogan et al. (2009) address a text regression problem to predict the financial risk of investment in companies. They analyze 10-K reports to predict stock return volatility. They also predict whether a company will be delisted following its 10-K report. Ruiz et al. (2012) correlate text with financial time series volume and price data. They find that graph centrality measures like page rank and degree are more strongly correlated to both price and traded volume for an aggregation of similar companies, while individual stocks are less correlated. Lavrenko et al. (2000) present an approach to identify news stories that influence the behavior of financial markets, and predict trends in stock prices based on the content of news stories that precede the trends. Luss and d'Aspremont (2008) and Lavrenko et al. (2000) both point out the desire for document feature engineering as future research directions. We explore a rich feature space that relies on frame semantic parsing.

Sentiment analysis figures strongly in NLP work on news. General Inquirer (GI), a content analysis program, is used to quantify pessimism of news in Tetlock (2007) and Tetlock et al. (2008). Other resources for sentiment detection include the Dictionary of Affect in Language (DAL) to score the prior polarity of words, as in Agarwal et al. (2011) on social media data. Our study incorporates DAL scores along with other features.

FrameNet is a rich lexical resource (Fillmore et al., 2003), based on the theory of frame semantics (Fillmore, 1976). There is active research to build more accurate parsers (Das and Smith, 2011; Das and Smith, 2012). Semantic role labeling using FrameNet has been used to identify an opinion with its holder and topic (Kim and Hovy, 2006). For deep representation of sentiment analysis, Ruppenhofer and Rehbein (2012) propose SentiFrameNet.

Category	Features	Value type
Frame	F, FT, FE	$\mathbb{N}$
attributes	wF, wFT, wFE	$\mathbb{R}_{\geq 0}$
BO**W**	UniG, BiG, TriG	$\mathbb{N}$
	wUniG, wBiG, wTriG	$\mathbb{R}_{\geq 0}$
p**D**AL	all-Pls, all-Act, all-Img	$\mathbb{R}_{\sim\mu=0,std=1}$
	VB-Pls, VB-Act, VB-Img	$\mathbb{R}_{\sim\mu=0,std=1}$
	JJ-Pls, JJ-Act, JJ-Img	$\mathbb{R}_{\sim\mu=0,std=1}$
	RB-Pls, RB-Act, RB-Img	$\mathbb{R}_{\sim\mu=0,std=1}$

Table 1: FWD features (**F**rame, bag-of-**W**ords, part-of-speech **D**AL score) and their value types.

Our work addresses classification tasks that have potential relevance to an influential financial model (Rydberg and Shephard, 2003). This model decomposes stock price analysis of financial data into a three-part ADS model - *activity* (a binary process modeling the price move or not), *direction* (another binary process modeling the direction of the moves) and *size* (a number quantifying the size of the moves). Our two binary classification tasks for news, price change and polarity, are analogous to their *activity* and *direction*. In contrast to the ADS model, our approach does not calculate the conditional probability of each factor. At present, our goal is limited to the determination of whether NLP features can uncover information from news that could help predict stock price movement or support analysts' investigations.

4 Methods

We propose two approaches for the use of semantic frames. The first is a rich vector space based on semantic frames, word forms and DAL affect scores. The second is a tree representation that encodes semantic frame features, and depends on tree kernel measures for support vector machine classification. The semantic parses of both methods are derived from SEMAFOR[1] (Das and Smith, 2012; Das and Smith, 2011), which solves the semantic parsing problem by rule-based target identification, log-linear model based frame identification and frame element filling.

[1] http://www.ark.cs.cmu.edu/SEMAFOR.

Frame (F)	Judgment_comm.	Commerce_buy
Target (FT)	accuse sue charge	buy purchase bid
Frame Element (FE)	COMMUNICATOR EVALUEE REASON	BUYER SELLER GOODS

Table 2: Sample frames.

4.1 Semantic Frame based FWD Features

Table 1 lists 24 types of features, including semantic **F**rame attributes, bag-of-**W**ords, and scores for words in the Dictionary of Affect in Language by part of speech (p**D**AL). We refer to these features as **FWD** features throughout the paper. FWD features are used alone and in combinations.

FrameNet defines hundreds of frames, each of which represents a scenario associated with semantic roles, or frame elements, that serve as participants in the scenario the frame signifies. Table 2 shows two frames. The frame *Judgment_communication* (*JC* or *Judgment_comm.* in the rest of the paper) represents a scenario in which a COMMUNICATOR communicates a judgment of an EVALUEE for some REASON. It is evoked by (target) words such as *accuse* or *sue*.

Here we use **F** for the frame name, **FT** for the target words, and **FE** for frame elements. We use both frequency and weighted scores. For example, we define *idf*-adjusted weighted frame features, such as wF for attribute F in document d as $wF_{F,d} = f(F,d) \times log\frac{|D|}{|d \in D: F \in d|}$, where $f(F,d)$ is the frequency of frame F in d, D is the whole document set and $|\cdot|$ is the cardinality operator.

Bag-of-**W**ords features include term frequency and *tfidf* of unigrams, bigrams, and trigrams.

DAL (Dictionary of Affect in Language) is a psycholinguistic resource to measure the emotional meaning of words and texts (Whissel, 1989). It includes 8,742 words that were annotated for three dimensions: Pleasantness (Pls), Activation (Act), and Imagery (Img). Agarwal et al. (2009) introduced part-of-speech specific DAL features for sentiment analysis. We follow their approach by averaging the scores for all words, verb only, adjective only, and adverb only words. Feature values are normalized to mean of zero and standard deviation of one.

4.2 SemTree Feature Space and Kernels

We propose **SemTree** as another feature space to encode **sem**antic information in **tree**s. SemTree can distinguish the roles of each company of interest, or *designated object* (e.g. who is suing and who is being sued).

4.2.1 Construction of Tree Representation

The semantic frame parse of a sentence is a forest of trees, each of which corresponds to a semantic frame. SemTree encodes the original frame structure and its leaf words and phrases, and highlights a designated object at a particular node as follows. For each lexical item (target) that evokes a frame, a backbone is found by extracting the path from the root to the role filler mentioning a designated object; the backbone is then reversed to promote the designated object. If multiple frames have been assigned to the same designated object, their backbones are merged. Lastly, the frame elements and frame targets are inserted at the frame root.

The top of Figure 2 shows the semantic parse for sentence *a* from section 2; we use it to illustrate tree construction for designated object *Oracle*. The parse has two frames (Figure 2-(1)&(2)), one corresponding to the main clause (verb *sue*), and the other for the tenseless adjunct (verb *say*). The reversed paths extracted from each frame root to the designated object *Oracle* become the backbones (Figures 2-(3)&(4)). After merging the two backbones we get the resulting SemTree, as shown in Figure 2-(5). By the same steps, this sentence would also yield a SemTree with *Google* at the root, in the role of EVALUEE.

4.2.2 Kernels and Tree Substructures

The tree kernel (Moschitti, 2006; Collins and Duffy, 2002) is a function of tree similarity, based on common substructures (tree fragments). There are two types of substructures. A subtree (ST) is defined as any node of a tree along with all its descendants. A subset tree (SST) is defined as any node along with its immediate children and, optionally, part or all of the children's descendants. Each tree is represented by a d dimensional vector where the i'th component counts the number of occurrences of the i'th tree fragment.

Define the function $h_i(T)$ as the number of occurrences of the i'th tree fragment in tree T, so that T is now represented as $\mathbf{h}(T) = (h_1(T), h_2(T), ..., h_d(T))$. We define the set of nodes in tree T_1 and T_2 as N_{T_1} and N_{T_2} respectively. We define the indicator function $I_i(n)$ to be 1 if subtree i is seen rooted at node n, and 0 otherwise. It follows that $h_i(T_1) = \sum_{n_1 \in N_{T_1}} I_i(n_1)$

Designated object: Oracle (ORCL)
Sentence: Oracle sued Google in August 2010, saying Google’s Android mobile operating system infringes its copyrights and patents for the Java programming language.
SRL: [Oracle$_{JC.FE.Communicator,Stmt.FE.Speaker}$] [sued$_{JC.Target}$] [Google$_{JC.FE.Evaluee}$] in August 2010, [saying$_{Stmt.Target}$] [Googleś Android mobile operating system infringes its copyrights and patents for the Java programming language$_{Stmt.FE.Message}$].

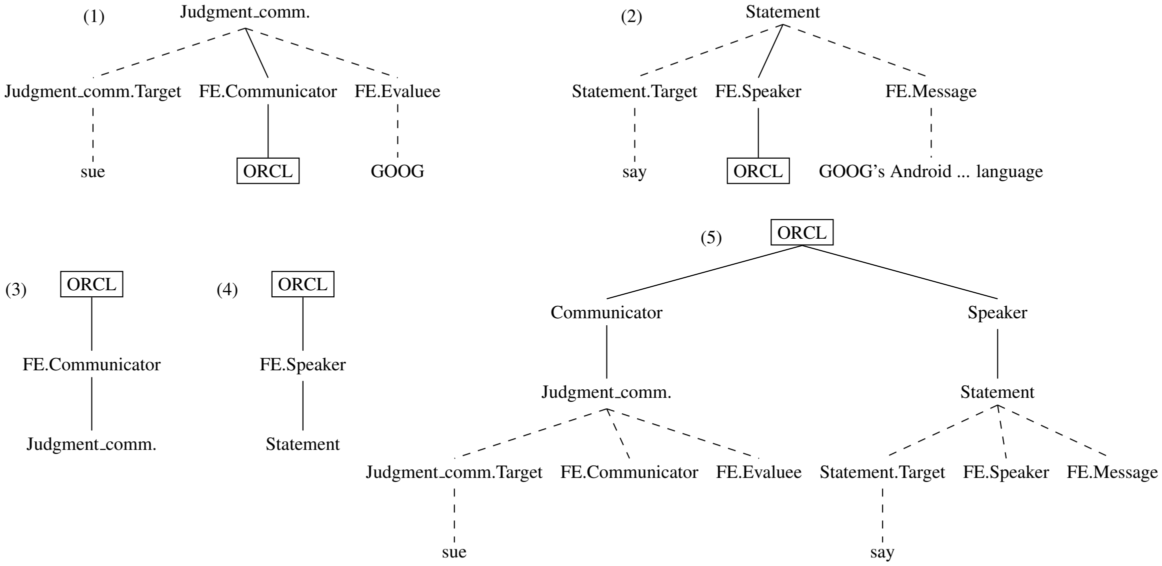

Figure 2: Constructing a tree representation for the designated object *Oracle* in sentence shown.

and $h_i(T_2) = \sum_{n_2 \in N_{T_2}} I_i(n_2)$. Their similarity can be efficiently computed by the inner product,

$$\begin{aligned} K(T_1,T_2) &= \mathbf{h}(T_1) \cdot \mathbf{h}(T_2) \\ &= \textstyle\sum_i h_i(T_1)h_i(T_2) \\ &= \textstyle\sum_i (\sum_{n_1 \in N_{T_1}} I_i(n_1))(\sum_{n_2 \in N_{T_2}} I_i(n_2)) \\ &= \textstyle\sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \sum_i I_i(n_i) I_i(n_2) \\ &= \textstyle\sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1,n_2) \end{aligned}$$

where $\Delta(n_1, n_2)$ is the number of common fragments rooted in the nodes n_1 and n_2. If the productions of these two nodes (themselves and their immediate children) differ, $\Delta(n_1, n_2) = 0$; otherwise iterate their children recursively to evaluate $\Delta(n_1, n_2) = \prod_j^{|children|}(\sigma + \Delta(c^j_{n_1}, c^j_{n_2}))$, where $\sigma = 0$ for ST kernel and $\sigma = 1$ for SST kernel.

The kernel computational complexity is $O(|N_{T_1}| \times |N_{T_2}|)$, where all pairwise comparisons are carried out between T_1 and T_2. However, there are fast algorithms for kernel computation that run in linear time on average, either by dynamic programming (Collins and Duffy, 2002), or pre-sorting production rules before training (Moschitti, 2006). We use the latter.

5 Dataset

We use publicly available financial news from Reuters from January 2007 through August 2012. This time frame includes a severe economic downturn in 2007-2010 followed by a modest recovery in 2011-2012.

An information extraction pipeline is used to pre-process the data. News full text is extracted from HTML. The timestamp of the news is extracted for a later alignment with stock price information, which will be discussed in section 6. The company mentioned is identified by a rule-based matching of a finite list of companies.

There are a total of 10 sectors in the Global Industry Classification Standard (GICS), an industry taxonomy used by the S&P 500.[2] To explore our approach for this domain, we select three sectors for our experiment: Telecommunication Services (TS, the sector with the smallest number of companies), Information Technology (IT), and Consumer Staples (CS), due to our familiarity with the companies in these sectors and an expectation of different characteristics they may exhibit. In the expectation there would be semantic differences associated with these sectors, experiments are performed independently for each sector. There are also differences in the number of companies in the sector, and the amount of news.

We bin news articles by sector. We remove articles that only list stock prices or only show tables of accounting reports. The first preprocessing step is to extract sentences that mention the

[2] Standard & Poor’s 500 is an equity market index that includes 500 U.S. leading companies in leading industries.

	CS (N=40)	IT (N=69)	TS (N=8)
avg # news	5,702±749	13446±1,272	2,177±188
avg # sentences	16,090±2,316	48,929±5,927	6,970±1,383
avg # com./sent.	1.07±0.01	1.06±0.20	1.14±0.03
avg # total	17,131±2,339	51,306±8,637	7,947±1,576

Table 3: Data statistics of mean and standard deviation by year from January 2007 to August 2012, for three sectors, with the number of companies.

relevant companies. Each data instance is a sentence and one of the target companies it mentions. Table 3 summarizes the data statistics. For example, the consumer staples sector has 40 companies. It has an average of 5,702 news articles (16,090 sentences) per year. Each sentence that mentions a consumer staple company mentions 1.07 companies on average. On average, this sector has 17,131 instances per year.

6 Experiments

Our current experiments are carried out for each year, training on one year and testing on the next. The choice to use a coarse time interval with no overlap was an expedience to permit more numerous exploratory experiments, given the computational resources these experiments require. We test the influence of news to predict (1) a change in stock price (*change* task), and (2) the polarity of change (increase vs. decrease; *polarity* task). Experiments evaluate the FWD and SemTree feature spaces compared to two baselines: bag-of-words (BOW) and supervised latent Dirichlet allocation (sLDA) (Blei and McAuliffe, 2007). BOW includes features of unigram, bigram and trigram. sLDA is a statistical model to classify documents based on LDA topic models, using labeled data. It has been applied to and shown good performance in topical text classification, collaborative filtering, and web page popularity prediction problems.

6.1 Labels, Evaluation Metrics, and Settings

We align publicly available daily stock price data from Yahoo Finance with the Reuters news using a method to avoid back-casting. In particular, we use the daily adjusted closing price - the price quoted at the end of a trading day (4PM US Eastern Time), then adjusted by dividends, stock split, and other corporate actions. We create two types of labels for news documents using the price data, to label the existence of a change and the direction of change. Both tasks are treated as binary classification problems. Based on the finding of a one-day delay of the price response to the information embedded in the news by Tetlock et al. (2008), we use $\Delta t = 1$ in our experiment. To constrain the number of parameters, we also use a threshold value (r) of a 2% change, based on the distribution of price changes across our data. In future work, this could be tuned to sector or time.

$$\text{change}=\begin{cases} +1 & \text{if } \frac{|p_{t(0)+\Delta t} - p_{t(-1)}|}{p_{t(-1)}} > r \\ -1 & \text{otherwise} \end{cases}$$

$$\text{polarity}=\begin{cases} +1 & \text{if } p_{t(0)+\Delta t} > p_{t(-1)} \text{ and } change = +1 \\ -1 & \text{if } p_{t(0)+\Delta t} < p_{t(-1)} \text{ and } change = +1 \end{cases}$$

$p_{t(-1)}$ is the adjusted closing price at the end of the last trading day, and $p_{t(0)+\Delta t}$ is the price of the end of the trading day after the Δt day delay. Only the instances with changes are included in the *polarity* task.

There is high variance across years in the proportion of positive labels, and often highly skewed classes in one direction or the other. The average ratios of +/- classes for *change* and *polarity* over the six years' data are 0.73 (std=0.35) and 1.12 (std=0.25), respectively. Because the time frame for our experiments includes an economic crisis followed by a recovery period, we note that the ratio between increase and decrease of price flips between 2007, where it is 1.40, and 2008, where it is 0.71. Accuracy is very sensitive to skew: when a class has low frequency, accuracy can be high using a baseline that makes prediction on the majority class. Given the high data skew, and the large changes from year to year in positive versus negative skew, we use a more robust evaluation metric.

Our evaluation relies on the Matthews correlation coefficient (MCC, also known as the ϕ-coefficient) (Matthews, 1975) to avoid the bias of accuracy due to data skew, and to produce a robust summary score independent of whether the positive class is skewed to the majority or minority. In contrast to f-measure, which is a class-specific weighted average of precision and recall, and whose weighted version depends on a choice of whether the class-specific weights should come from the training or testing data, MCC is a single summary value that incorporates all 4 cells of a 2×2 confusion matrix (TP, FP, TN and FN for True or False Positive or Negative). We have also observed that MCC has a lower relative standard deviation than f-measure.

For a 2×2 contingency table, MCC corresponds to the square root of the average χ^2 statistic $\sqrt{\chi^2/n}$, with values in [-1,1]. It has been sug-

	Change			
test years	BOW	sLDA	FWD	SemTreeFWD
	Consumer Staples			
2008-2010	0.1015	0.0774	0.1079	0.1426
2011-2012	0.1663	0.1203	0.1664	0.1736
5 years	0.1274	0.0945	0.1313	0.1550
	Information Technology			
2008-2010	0.0580	0.0585	0.0701	0.0846
2011-2012	0.0894	0.0681	0.1076	0.1273
5 years	0.0705	0.0623	0.0851	0.1017
	Telecommunication Services			
2008-2010	0.1501	0.1615	0.1497	0.2409
2011-2012	0.2256	0.2084	0.2191	0.4009
5 years	0.1803	0.1803	0.1774	0.3049
	Polarity			
	Consumer Staples			
2008-2010	0.0359	0.0383	0.0956	0.1054
2011-2012	0.0938	0.0270	0.1131	0.1285
5 years	0.0590	0.0338	0.1026	0.1147
p-value		>>0.1000	0.0918	**0.0489**
	Information Technology			
2008-2010	0.0551	0.0332	0.0697	0.0763
2011-2012	0.0591	0.0516	0.0764	0.0857
5 years	0.0567	0.0405	0.0723	0.0801
p-value		0.0626	0.0948	**0.0103**
	Telecommunication Services			
2008-2010	0.0402	0.0464	0.0821	0.0745
2011-2012	0.0366	0.0781	0.0611	0.0809
5 years	0.0388	0.0591	0.0737	0.0770
p-value		>>0.1000	0.0950	**0.0222**

Table 4: Average MCC for the change and polarity tasks by feature representation, for 2008-2010; for 2011-2012; for all 5 years and associated p-values of ANOVAs for comparison to BOW.

gested as one of the best methods to summarize into a single value the confusion matrix of a binary classification task (Jurman and Furlanello, 2010; Baldi et al., 2000). Given the confusion matrix $\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

All sentences with at least one company mention are used for the experiment. We remove stop words and use Stanford CoreNLP for part-of-speech tagging and named entity recognition. Models are constructed using linear kernel support vector machines for both classification tasks. SVM-light with tree kernels[3] (Joachims, 2006; Moschitti, 2006) is used for both the FWD and SemTree feature spaces.

6.2 Results

Table 4 shows the mean MCC values for each task, for each sector. Separate means are shown for the test years of financial crisis (2008-2010) and economic recovery (2011-2012) to highlight the differences in performance that might result from market volatility.

pos. 1	dow, investors, index, retail, data
pos. 2	costs, food, price, prices, named_entity_4
neu. 1	q3, q1, nov, q2, apr
neu. 2	cents, million, share, year, quarter
neg. 1	cut, sales, prices, hurt, disappointing
neg. 2	percent, call, company, fell, named_entity_7

Table 5: Sample sLDA topics for consumer staples for test year 2010 (train on 2009), polarity task.

SemTree combined with FWD (SemTreeFWD) generally gives the best performance in both *change* and *polarity* tasks. SemTree results here are based on the subset tree (SST) kernel, because of its greater precision in computing common frame structures and consistently better performance over the subtree (ST) kernel. SemTree also provides interpretable features for manual analysis as discussed in the next section.

Analysis of Variance (ANOVA) tests were performed on the full 5 years for each sector, to compare each feature representation as a predictor of MCC score with the baseline BOW. The ANOVAs yield the p-values shown in Table 4. There were no significant differences from BOW on the *change* task. For *polarity* detection, SemTreeFWD was significantly better than BOW for each sector (see boldface p-values). No other method was significantly better than BOW, although FWD approaches significance on all sectors, and sLDA approaches significance on IT.

sLDA has promising MCC scores for the telecommunication sector, which has only 8 companies, thus many fewer data instances. Table 5 displays a sample of sLDA topics with good performance on polarity for the consumer staples sector for training year 2009. The positive topics are related to stock index details and retail data. The negative topics contain many words with negative sentiment (e.g., *hurt, disappointing*).

7 Discussion

7.1 Semantic Parse Quality

In general, SEMAFOR parses capture most of the important frames for our purposes. There is, however, significant room for improvement. On a small, randomly selected sample of sentences from all three sectors, two of the authors working independently evaluated the semantic parses, with approximately 80% agreement. Some of the inaccuracies in frame parses result from errors prior to the SEMAFOR parse, such as tokenization or

[3] SVM-light: http://svmlight.joachims.org and Tree Kernels in SVM-light: http://disi.unitn.it/moschitti/Tree-Kernel.htm.

```
+ (Target(jump))
+ (RECIPIENT(Receiving))
+ (VICTIM(Defend))
+ (PERCEIVER_AGENTIVE(Perception_active(Target)
(PERCEIVER_AGENTIVE)(PHENOMENON)))
+ (DONOR(Giving(Target)(THEME)(DONOR)))
+ (Target(beats))
...
- (PHENOMENON(Perception_active(Target)(PERCEIVER
_AGENTIVE)(PHENOMENON)))
- (TRIGGER(Response))
- (Target(cuts))
- (VICTIM(Cause_harm(Target(hurt))(VICTIM)))
```

Figure 3: Best performing SemTree fragments for increase (+) and decrease (-) of price for consumer staples sector across training years.

dependency parsing errors. The average sentence length for the sample was 33.3 words, with an average of 14 frames per sentence, 3 of them with a GICS company as a role filler. Because SemTree encodes only the frames containing a designated object (company), these are the frames we evaluated. On average, about half the frames with a designated object were correct, and two thirds of those frames we judged to be important. Besides errors due to incorrect tokenization or dependency parsing, we observed that about 8% to 10% of frames were incorrectly assigned to due word sense ambiguity.

7.2 Feature Analysis

The experimental results show the SemTree space to be the one representation tested here that is significantly better than BOW, but only for the *polarity* task. Post hoc analysis indicates this may be due to the aptness of semantic frame parsing for polarity. Limitations in our treatment of time point to directions for improvement regarding the *change* task.

Some strengths of our approach are the separate treatment of different sectors, and the benefits of SemTree features. To analyze which were the best performing features within sectors, we extracted the best performing frame fragments for the *polarity* task using a tree kernel feature engineering method presented in Pighin and Moschitti (2009). The algorithm selects the most relevant features in accordance with the weights estimated by SVM, and uses these features to build an explicit representation of the kernel space. Figure 3 shows the best performing SemTree fragments of the *polarity* task for the consumer staples sector.

Recall that we hypothesized differences in semantic frame features across sectors. This shows up as large differences in the strength of features across sectors. More strikingly, the same feature can differ in polarity across sectors. For example, in consumer staples, (EVALUEE(*Judgment_communication*)) has positive polarity, compared with negative polarity in information technology sector. The examples we see indicate that the positive cases pertain to aggressive retail practices that lead to lawsuits with only small fines, but whose larger impact benefits the bottom line. A typical case is the sentence, *The plaintiffs accused Wal-Mart of discriminating against disabled customers by mounting "point-of-sale" terminals in many stores at elevated heights that cannot be reached.* Lawsuits in the IT sector, on the other hand, are often about technology patent disputes, and are more negative, as illustrated by our example sentence in Figure 2.

SemTree features capture the differences between semantic roles for the same frame, and between the same semantic role in different frames. For example, the PERCEIVER_AGENTIVE role of the *Perception_active* frame contributes to prediction of an increase in price, as in *R.J. Reynolds is watching this situation closely and will respond as appropriate*. Conversely, a company that fills the PHENOMENON role of the same frame contributes to prediction of a price decrease, as in *Investors will get a clearer look at how the market values the Philip Morris tobacco businesses when Altria Group Inc. "when-issued" shares begin trading on Tuesday*. When a company fills the VICTIM role in the *Cause_harm* frame, this can predict a decrease in price, as in *Hershey has been hurt by soaring prices for cocoa, energy and other commodities*, whereas filling the VICTIM role in the *Defend* frame is associated with an increase in price, as in *At Berkshire's annual shareholder meeting earlier this month, Warren Buffett defended Wal-Mart, saying the scandal did not change his opinion of the company.*

One weakness of our approach that we discussed above is that there is a strong effect of time that we do not address. The same SemTree feature can be predictive for one time period and not for another. (GOODS(*Commerce_sell*)) is related to a decrease in price for 2008 and 2009 but to an increase in price for 2010-2012. There is clearly an influence of the overall economic context that we do not take into account. For example, the practices of acquiring or selling a business are different in downturning versus recovering markets. An important observation of the MCC values, especially in the case of SemTreeFWD is that MCC increases during the years 2011-2012. We attribute this change to the difficulty of predicting stock price trends when there is the high volatility typical of a financial crisis. The effect of news on volatility, however, can be explored independently. For example, Creamer et al. (2012) detect a strong association.

Another weakness of our approach is that we take sentences out of context, which can lead to prediction errors. For example, the sentence *Longs' real estate assets alone are worth some $2.9 billion, or $71.50 per share, Ackman wrote, meaning that* $\boxed{\textit{CVS}}$ *would essentially be paying for real estate, but gaining Longs' pharmacy benefit management business and retail operations for free* is treated as predicting a positive polarity for *CVS*. This would be accurate if *CVS* was actually going to acquire *Longs*' business. Later in the same news item, however, there is a sentence indicating that the sale will not go through, which predicts negative polarity for *CVS*: *Pershing Square Capital Management said on Thursday it won't support a tender offer from CVS Caremark Corp for rival Longs Drug Stores Corp because the offer price "materially understates the fair value of the company," according to a filing.*

8 Conclusion

We have presented a model for predicting stock price movement from news. We proposed FWD (**F**rames, BO**W**, and part-of-speech specific **D**AL) features and SemTree data representations. Our semantic frame-based model benefits from tree kernel learning using support vector machines. The experimental results for our feature representation perform significantly better than BOW on the *polarity* task, and show promise on the *change* task. It also facilitates human interpretable analysis to understand the relation between a company's market value and its business activities. The signals generated by this algorithm could improve the prediction of a financial time series model, such as ADS (Rydberg and Shephard, 2003).

Our future work will consider the contextual information for sentence selection, and an aggregation of weighted news content based on the decay effect over time for individual companies. We plan to use a moving window for training and testing. We will also explore different labeling methods, such as a threshold for price change tuned by sectors and background economics.

9 Acknowledgements

The authors thank the anonymous reviewers for their insightful comments.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 24–32, Athens, Greece, March. Association for Computational Linguistics.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38. Association for Computational Linguistics.

Pierre Baldi, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412 – 424.

Roy Bar-Haim, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. 2011. Identifying and following expert investors in stock microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1310–1319, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6*.

Christopher Chua, Maria Milosavljevic, and James R. Curran. 2009. A sentiment detection engine for internet stock message boards. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 89–93, Sydney, Australia, December.

Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 263–270, Stroudsburg, PA, USA. Association for Computational Linguistics.

Germán G. Creamer, Yong Ren, and Jeffrey V. Nickerson. 2012. A Longitudinal Analysis of Asset Return, Volatility and Corporate News Network. In *Business Intelligence Congress 3 Proceedings*.

Dipanjan Das and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1435–1444, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dipanjan Das and Noah A. Smith. 2012. Graph-based lexicon expansion with sparsity-inducing penalties. In *HLT-NAACL*, pages 677–687. The Association for Computational Linguistics.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic, June. Association for Computational Linguistics.

William Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. *Proceedings of the 20th International Conference on Computational Linguistics*.

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

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, 16(3):235–250, September.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Syed Aqueel Haider and Rishabh Mehrotra. 2011. Corporate news classification and valence prediction: A supervised approach. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 175–181, Portland, Oregon, June. Association for Computational Linguistics.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

Giuseppe Jurman and Cesare Furlanello. 2010. A unifying view for performance measures in multi-class prediction. *ArXiv e-prints*.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 272–280, Stroudsburg, PA, USA. Association for Computational Linguistics.

Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. 2000. Mining of concurrent text and time series. In *In proceedings of the 6th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, pages 37–44.

Ronny Luss and Alexandre d'Aspremont. 2008. Predicting abnormal returns from news using text classification. *CoRR*, abs/0809.2792.

Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Daniele Pighin and Alessandro Moschitti. 2009. Reverse engineering of tree kernel feature spaces. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore*, pages 111–120.

Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. 2012. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 513–522, New York, NY, USA. ACM.

Josef Ruppenhofer and Ines Rehbein. 2012. Semantic frames as an anchor representation for sentiment analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 104–109, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tina H. Rydberg and Neil Shephard. 2003. Dynamics of Trade-by-Trade Price Movements: Decomposition and Models. *Journal of Financial Econometrics*, 1(1):2–25.

Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*.

Paul C. Tetlock. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*.

Cynthia M. Whissel. 1989. The dictionary of affect in language. *Emotion: Theory, Research, and Experience*, 39(4):113–131.