Predicting Market-Volatility from Federal Reserve Board Meeting Minutes NLP for Finance

Reza Bosagh Zadeh, Andreas Zollmann

1 Introduction

Predicting markets has always had a certain appeal to researchers, and news can have huge impacts on markets. Announcements about corporate profits (or lack thereof), a change in management, a new promising product, talk of a takeover, even the weather. All these events can cause a company's share price to move wildly up or down. To automate trading strategies that can leverage this news information, the mining and sifting of news takes place in computers dedicated to algorithmic trading, removing human emotions from the decision making process and speeding up reactions. The meetings of the federal reserve should also have a huge impact on the market since fiscal policy has a direct impact on the economy.

The use of tools from natural language processing for finance has been studied in numerous publications. A survey is available in (Mittermayer and Knolmayer, 2006a). In this work, we perform experiments in predicting volatilities of several financial indicators based on the publicly available minutes of the Federal Reserve Board meetings.

2 Past work

Much work on news-based market forecasting has occurred in recent years. Some research focused on confirming or refuting different variants of the *efficient market hypothesis* (Fama, 1965), stating that excess returns per unit risk cannot be consistently generated using public information, with differing conclusions. (Gidófalvi and Elkan, 2003) develop a Naive-Bayes prediction model for short-term stock market returns (classes 'up', 'down', 'normal') based on the contents of financial news articles treated as bags-of-words. They find the model that has the highest predictive accuracy over the 20 minutes trading window *before* the publication time of the respective article, suggesting that market corrections for news are primarily due to insiders. Support-vector machine (SVM) and regression text classification models have also successfully been applied to predict intra-day stock market returns, e.g. by (Pui Cheong Fung et al., 2003; Schumaker and Chen, 2008) based on Reuters news data, and by (Mittermayer and Knolmayer, 2006b) based on press releases. On the other hand, (Luss and d'Aspremont, 2008) found their model, an SVM using both past equity returns and press release bags-of-words, unable to predict stock return differences.

The text-based features employed for news-based stock market prediction have been fairly simple in past work (usually words or word tuples selected according to frequency, TF-IDF score, or information gain). In (Schumaker and Chen, 2008), the effect of using proper nouns, noun phrases, and/or named entities as features was investigated, the proper noun representation being the most successful. In a paper about forecasting public opinion from news, (Lerman et al., 2008) go much further in the modeling of the news documents as feature vectors. In addition to the bag-of-word features, *news focus features* are introduced, representing the change in occurrence frequency of a word in the current day's news coverage compared to the average news coverage of the past N days. A further novelty is the introduction of *dependency features*, essentially fragments of the dependency tree of a sentence in the concerned news article. Examples of these features are "won→Kerry", "Kerry←spokesperson→campaign", and "agenda→'s→Bush".

Another market variable apart from price that researchers have attempted to forecast based on relevant text data is volatility (e.g. (Seo et al., 2002)). Text data attempted as right-hand-side variable has been news, press releases, and SEC filings, amongst others. For example, in (Kogan et al., 2009), volatility of stock returns is predicted based on annual reports of the respective companies.

We are not aware of any attempts to use Federal Reserve Board meeting minutes to predict financial indicators. However, as (Boukus and Rosenberg, 2006) show, market participants do extract complex signals from these minutes. Using Latent Semantic Analysis, the authors found correlations of e.g. Treasury yields with specific themes of the meeting minutes.

3 General Framework

We developed a framework that allows predicting realized volatility in stock market prices from Federal Reserve Board meeting minutes. The first step was to harvest all the meeting minutes from the web, split the documents into individual units corresponding to sentences, and tokenize the sentences. Meeting minutes from 1993 onwards are in HTML format and could easily be harvested. Minutes before that only are available as PDF documents, but we were able to retrieve the text with reasonable accuracy using a PDF-to-HTML converter.

We obtained 398 meeting minutes, ranging from 1967 to March 2008.[1] This corpus consists of 28,817 sentences and 1,031,662 words. We noticed that the vocabulary of the meeting minutes is extremely limited. We compared its language with that of a corpus of Wall Street Journal news by counting the number of unique words from each genre in subcorpora of 470,000 words each. While the meeting minutes subcorpus only has a vocabulary of 7300 words, the news corpus vocabulary is of size 41,289—nearly six times as large.

Due to the time-dependent nature of our prediction task, we progressively split the data, using the first 80% (roughly years 1967-2000) for training, and reserving the last 20% (2001-2008) for testing.

As these minutes refer to monetary policy and economic outlook affecting all U.S. financial markets as a whole rather than just individual companies, it makes sense to predict volatility as an index over many market variables. For most of our experiments, we chose the volatility of the daily returns of the S&P 500 index incurred over the next n days following the meeting. As mentioned in Section 5, we investigated several other indices in preliminary experiments.

A natural question arising is what measure of volatility is most suitable, and what time span n we should use to compute it. We used the sample variance of log-returns as volatility measure:

$$\text{vol} = 1/(n-1) \sum_{i=1}^{n} [\ln \text{return}(t+i) - 1/n \sum_{j=1}^{n} \ln \text{return}(t+j)]^2 \ ,$$

where t is the time of the meeting and

$$\text{return}(t) = \text{price}(t)/\,\text{price}(t-1) - 1 \ .$$

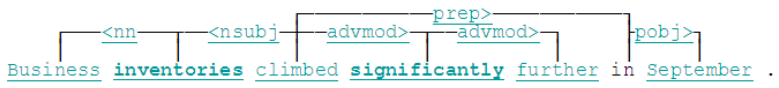
We experimented with time spans n from just two days up to one year.

3.1 Text mining

Deciding upon which features to use is a challenge as in most machine learning and natural language processing tasks. Conventional approaches are bag-of-words. The set of words used as features may be hand-selected or automatically selected.

Following (Kogan et al., 2009), we extracted unigram as well as bigram bag-of-word features from the meeting minutes. As a preprocessing step (apart from sentence splitting and tokenization), we stemmed the words using the Porter2 stemming algorithm and removed stop words.

[1] Data after March 2008 was discarded since we were aiming to predict up to one year into the future, which required index prices for up to one year after the last meeting date for evaluation.



- Using seedword 'inventories', extract fragments:

Inventories→climbed, business→inventories→climbed, inventories→climbed←further

Figure 1: A sentence and its extracted dependency features for the seed word 'inventories'.

We experimented with simple word frequencies, as well as TFIDF scores, and LOG1P (defined as $\ln(1+f)$, where f is the word/bigram frequency).

Many bag-of-words features have already been tried, so we propose some features that are less common in the literature. The only paper for text-based market prediction we are aware of that uses more advanced features than bag-of-words is Lerman et al. (Lerman et al., 2008) who used dependency parses to extract relations amongst entities. To apply this model to our task, we asked a financial expert to select from a list of all words from the meeting minutes, sorted by decreasing frequency, the words deemed most relevant to changes in volatility. We then used these words as 'seeds' in the dependency tree fragment model: all subtrees of the form 'child→parent', 'child→parent←child', and 'child→parent→grandparent' that had at least one seed word in them were extracted. We used the Stanford dependency parser to parse the (pre-tokenized) meeting minute sentences. Figure 1 shows a (simplified) sentence from one of the meetings with extracted dependency fragments for the seed word 'inventories'.

4 Evaluation

For training and evaluation, we may use two different methods: classification and regression. When classification is used, we only predict whether volatility will rise or fall, and when regression is used, we will predict the volatility itself.

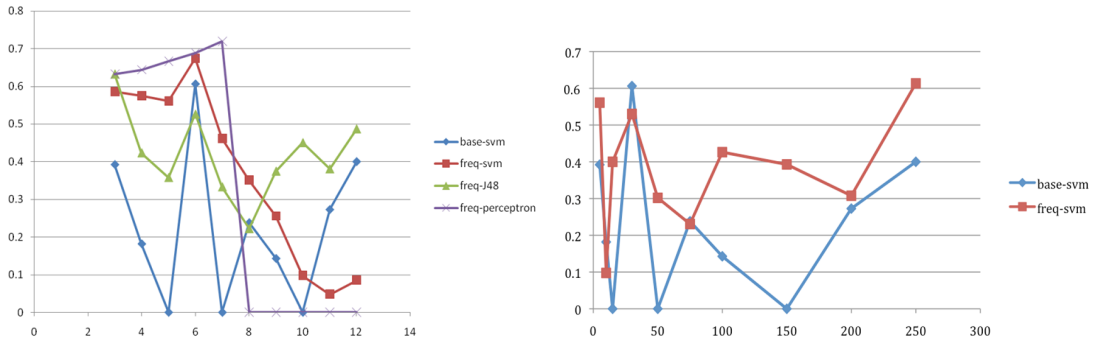
Our baseline will be the volatility of the past n days. So for example, say we are predicting the volatility of n days from today, we will use the crude prediction of the previous n days as a baseline. Then, using our federal reserve board minutes and the SVMs, we try to predict the volatility n from now. There are two modes of evaluation: accuracy—for the case of classification—and mean squared error—for the regression scenario. For accuracy, we claim success if the SVM (or other classifier) can predict the direction of the volatility (whether it goes up or down). Our experiments will vary n to see how the system performs for shorter versus longer delays.

If we fail to beat the baseline, then we will create a hybrid of our system and the baseline in order to beat the baseline. This is valid because it shows the text-based system adds information where the crude baseline cannot.

5 Results

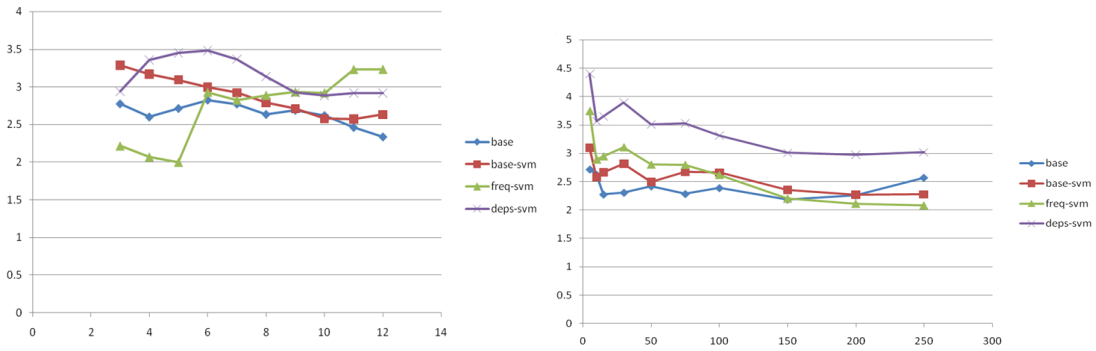
To see the effect of looking ahead far versus near terms, we used two time scales. One linear and another exponential. In all the graphs below, the horizontal axis is number of days (n) looked ahead. So for example if today is the meeting day and $n = 100$, we would be trying to predict market volatility from today till 100 business days from today - similarly the information available for baselines is computed by taking 100 business days before today. For regression, the vertical axis is Root Mean Squared Error (RMSE), so lower is better. For classification, the vertical axis corresponds to accuracy, except for Figure 2, from an early experiment in which we were evaluating F1 measure; in both cases higher is better.

5.1 S&P500 Regression and Classification for TF-IDF bag-of-word models



(a) Short periods (b) Long periods

Figure 2: Classification for TF-IDF bag-of-word ('freq') models. The baseline is only trained on the class labels. F1-measure vs Lookahead (days)



(a) Short periods (b) Long periods

Figure 3: Regression for TF-IDF bag-of-word ('freq') and dependency features ('deps'), when including previous volatility as a feature. The baseline 'base' simply predicts previous volatility, whereas 'base-svm' is an SVM trained on previous volatilities. RMSE vs Lookahead (days)

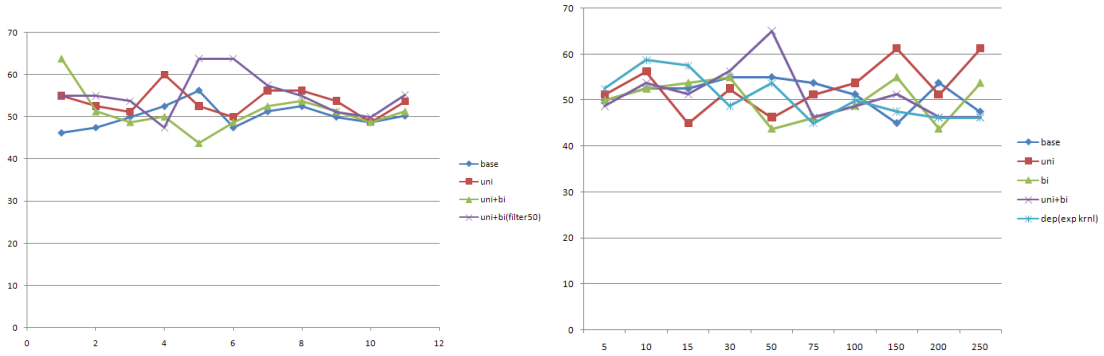
For most experiments there are two graphs, one which looks into the future with linearly increasing number of days (short periods) and another which does the same with exponential increase (long periods). We hypothesize that short term changes will be predictable with higher accuracy since the impact of the meetings is expected to dampen with time, as the newly revealed information has time to spread across markets.

As can be seen from figures 2 and 3, unfortunately neither Classification nor Regression provide consistently positive results in this task. For classification in the first 3-7 days, the SVM classifier can consistently beat the baseline, but fails thereafter.

5.2 S&P500 LOG1P frequency Regression and Classification

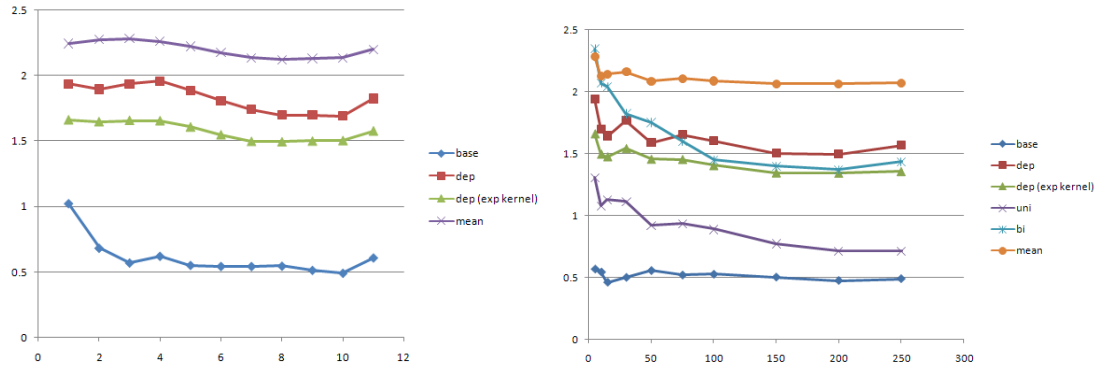
Since we already remove stop-words, there is no clear benefit from dampening features by IDF(inverse document frequency). There are more encouraging results with LOG1P features, shown in figure 4. Again, short-term results are better than long-term. In these experiments, the regression models did not include previous volatility as a feature, which gives a better insight in how well the text only can predict volatility. A simple baseline 'mean' is to always predict the mean volatility encountered during training. Baseline 'base' is again the (much stronger) model predicting just previous volatility.

Since it is unclear how much of an impact the meetings of the Federal Reserve Board really



(a) Short periods (b) Long periods

Figure 4: Classification for models with LOG1P bag-of-word ('uni') and additional bag-of-bigram features ('uni+bi'). Filter50 removed all features occurring less than 50 times. Baseline always predicts the majority class. Accuracy vs Lookahead (days)



(a) Short periods (b) Long periods

Figure 5: Regression without use of previous volatility (except for model 'base', which uses just that). RMSE vs Lookahead (days)

have on the constituent stocks of the S&P 500, we investigate further with other indices. We try predicting the volatility of 13-week Treasury Bills and 10-year Treasury Notes.

5.3 Results for other indices

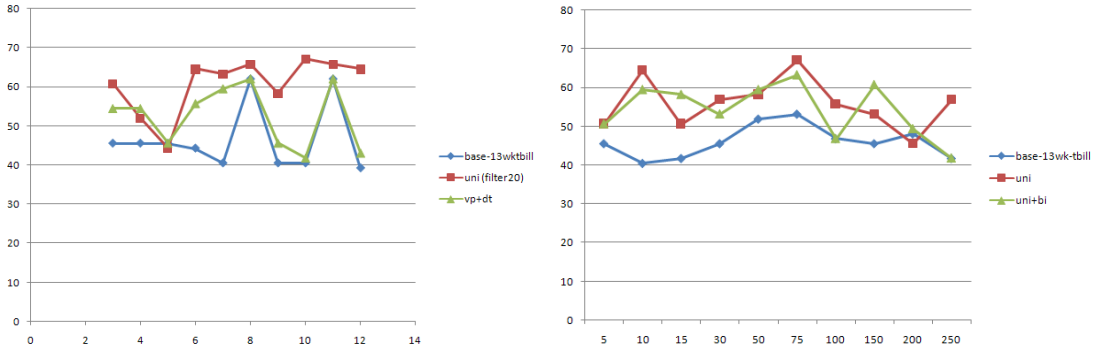
Classification for 13 week treasury bills shows promise for both long and short terms (figure 6), in that we consistently beat our baseline for both short term and long term prediction. It should be noted that the baseline here had no information other than the words of meetings.

A potential explanation for the better predictability of 13 week treasury bills could be that their prices are more directly tied to the actions of the Federal Reserve Board. Encouraged by this result, we tried predicting yet a different index, namely the 10 year treasury note index.

Perhaps surprisingly, the 10 year treasury notes volatility is harder to predict (figure 7) than the 13 week treasury bill volatility. A reasonable explanation for this is that the longer timespan of these bills (10 years versus 13 weeks) makes them less affected by the short-term discussions of the Federal Reserve.

6 Interesting Models

It is interesting to look inside our trained models and see what the classifiers are learning. For this purpose, we provide the learned classification trees and some prominent terms from the hyperplane learned by the support vector machines in Figure 8. The classification accuracy of each learned function is given.



(a) Short periods (b) Long periods

Figure 6: Classification for 13 week Treasury Bills. Accuracy vs Lookahead (days)

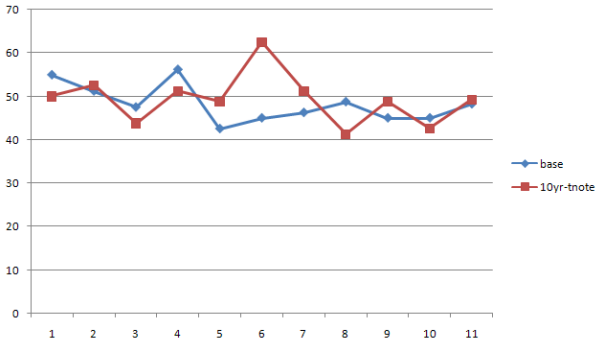


Figure 7: Classification for 10 year treasury notes. Accuracy vs Lookahead (days)

Surprisingly, such simple trees can predict with 62% accuracy the direction in which the volatility is expected move. A leaf node with UP label means we expect the volatility to rise. The features here are LOG1P.

7 Conclusions and future work

We have shown that the Federal Reserve Board meeting minutes have predictive power on future volatility of several indices: In a classification setting, a support-vector-machine model is able to outperform a random-guessing as well as a majority-class guessing baseline when predicting short-term volatility of the S&P500, when predicting short-term as well as long-term volatility of 13-week Treasury Bills. The outperformance is strongest for short-term volatility of 13-week T-Bills. No consistent improvements could be achieved on 10-year Treasury Notes. These preliminary results suggest that FRB meetings might have more influence on the bond market than on the stock market, and that it affects markets most strongly in the days after the meeting.

As part of this work, we harvested the meeting minutes since their beginning in 1967 and have made them available online, in the original format, in plain text, and in stemmed form, at:

`http://rezab.ca/useful/fomc_minutes.html`

We hope this will be a useful contribution to the research community.

A possible extension of our work is the prediction of implied instead of realized volatility, i.e., the future volatility in a stock price or index expected at a given time by market participants, which can be inferred indirectly from the price of an option on the stock or index, its strike, and the price of the stock or index itself. Note that this is a much more ambitious aim than predicting realized volatility since change in realized volatility can be consistent with the belief of market participants, whereas a change in expected future volatility necessarily reflects a change

```
w_stimulu < 6.2E-4
|  w_capit < 0.00253
|  |  w_statist < 0.00146: DOWN(80.0/45.0)
|  |  w_statist >= 0.00146
|  |  |  w_specifi < 0.00173: UP(120.0/79.0)
|  |  |  w_specifi >= 0.00173: DOWN(17.0/3.0)
|  w_capit >= 0.00253: UP(19.0/1.0)
w_stimulu >= 6.2E-4: DOWN(28.0/4.0)
```

(a) 52% Accurate

```
w_execut < 0.00149: DOWN(107.0/53.0)
w_execut >= 0.00149
|  w_dissent < 0.00307
|  |  w_month < 0.00198: DOWN(36.0/24.0)
|  |  w_month >= 0.00198: UP(105.0/56.0)
|  w_dissent >= 0.00307: DOWN(15.0/1.0)
```

(b) 62% Accurate

Figure 8: Classification Trees for 13 week Treasury Bills.

```
negative:
      -0.5677 * (normalized) w_action
      -0.554  * (normalized) w_manufactur
      -0.4998 * (normalized) w_slowli
      -0.4965 * (normalized) w_craven
      -0.4947 * (normalized) w_recent
      -0.4771 * (normalized) w_outcom
      -0.3755 * (normalized) w_crude
      -0.3732 * (normalized) w_institut
      -0.3718 * (normalized) w_affect
      -0.3694 * (normalized) w_climb
      -0.3551 * (normalized) w_canadian
      -0.3533 * (normalized) w_cumul
positive:
       0.3664 * (normalized) w_surg
       0.3679 * (normalized) w_polici
       0.3735 * (normalized) w_warehous
       0.375  * (normalized) w_resum
       0.3864 * (normalized) w_job
       0.4039 * (normalized) w_implement
       0.4054 * (normalized) w_outlook
       0.4059 * (normalized) w_struckmey
       0.4068 * (normalized) w_cutback
       0.6298 * (normalized) w_downward
       0.6536 * (normalized) w_curtail
```

Figure 9: Prominent terms in the learned SVM hyperplane

in the participants' beliefs and thus being able to predict implied volatility ourselves would allow us to devise trading strategies enabling us to profit from our predictions. This would mean that volatility markets are not efficient. We are considering using the *Chicago Board Options Exchange Volatility Index* (VIX), also popularly known as the 'fear index', as measure of implied volatility. VIX is a weighted sum of short-term out-of-the-money S&P 500 options. There are also futures and options with VIX as the underlying, making the devising of a trading strategy trivial.

References

Boukus, Ellyn and Joshua V. Rosenberg. 2006. The information content of FOMC minutes. http://ssrn.com/abstract=922312.

Fama, Eugene F. 1965. The behavior of stock-market prices. *The Journal of Business*, 38(1):34–105.

Gidófalvi, G. and C. Elkan. 2003. Using news articles to predict stock price movements. Technical report, Department of Computer Science and Engineering, University of California.

Kogan, Shimon, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting Risk from Financial Reports with Regression. In *Proceedings of the North American Association for Computational Linguistics Human Language Technologies Conference*, Boulder, CO.

Lerman, Kevin, Ari Gilder, Mark Dredze, and Fernando Pereira. 2008. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 473–480, Manchester, UK, August. Coling 2008 Organizing Committee.

Luss, Ronny and Alexandre d'Aspremont. 2008. Predicting abnormal returns from news using text classification. http://arxiv.org/abs/0809.2792v2.

Mittermayer, M.A. and G. Knolmayer. 2006a. Text Mining Systems for Market Response to News: A Survey. Technical report, Working paper.

Mittermayer, Marc-Andre and Gerhard F. Knolmayer. 2006b. Newscats: A news categorization and trading system. *IEEE International Conference on Data Mining*, pages 1002–1007.

Pui Cheong Fung, G., J. Xu Yu, and Wai Lam. 2003. Stock prediction: Integrating text mining approach using real-time news. *IEEE International Conference on Computational Intelligence for Financial Engineering*, pages 395–402, March.

Schumaker, Robert P. and Hsinchun Chen. 2008. Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies. *J. Am. Soc. Inf. Sci. Technol.*, 59(2):247–255.

Seo, Young-woo, Joseph A. Giampapa, and Katia P. Sycara. 2002. Text classification for intelligent agent portfolio management. In *In International Joint Conference on Autonomous Agents and Multi-Agent Systems*.