Financial Text Mining: Supporting Decision Making Using Web 2.0 Content

Hsin-Min Lu and Hsinchun Chen, *University of Arizona*
Tsai-Jyh Chen, *National Chengchi University*
Mao-Wei Hung and Shu-Hsing Li, *National Taiwan University*

Deep penetration of personal computers, data communication networks, and the Internet has created a massive platform for data collection, dissemination, storage, and retrieval. Everyday, people engage in numerous online activities, including reading the news and product reviews, commenting on developing events, buying and selling stocks, and widening their social networks. This widespread engagement with online worlds has facilitated the creation of large amounts of textual data.

For financial institutions, keeping track of important events and sentiments in a near real-time fashion could accelerate their decision-making process. With the help of text-mining techniques (the "fishing poles"), tangible profits can arise from smart trading strategies and improved portfolio management.

Making sense of the continual stream of textual data requires the development of a surveillance system that can collect, filter, extract, quantify, and analyze relevant information from the Internet. The nature of textual data, however, presents unique challenges. Although researchers and practitioners have achieved a certain level of success in extracting factual information, comprehending the meaning and implication of financial documents and investment strategies from a large collection of Web-based textual data is far more difficult.

Types of Finance-Related Textual Content

Finance-related textual content falls roughly into three categories. The first category includes forums, blogs, and wikis. This category is probably the most representative type of textual-data source on Web 2.0, the latest generation of the WWW. Finance forums such as Yahoo! Finance has existed for a long time and has attracted many postings. A typical IT company forum has hundreds of new messages every day. Users actively share their investment strategies, new product information, and perspectives and opinions. Information about company background, rumors, and news updates are also prevalent in many finance-related blogs and wiki sites.

The second category of finance-related content includes news and research reports. Newspaper articles are often accessible on news Web sites. Moreover, various finance portals provide intraday updates with contents from newswire services. Some portal sites also provide access to research reports generated by analysts.

The third category involves finance-related content generated by firms. Many firms maintain their own Web sites as a communication channel with consumers and investors. Public companies are required to submit their filings to the Edgar system (www.sec.gov/edgar.shtml), which is publicly accessible on the Web.

Capturing Relevant Information

Text mining for financial content has been examined particularly closely within the context of the stock market. One important research question in this context is the representation of relevant information. When the goal is to predict short-term (20-minute) price movements, representing text using proper nouns is more effective than bag-of-words (BOW) or noun phrase representations.[1] Sentiment polarity also draws considerable attention in financial text mining. Recent studies have found that sentiments of online financial forums are positively related to stock return, trading volume, and volatility. Although the effect is statistically significant, the impact is fairly small.[2] Sentiments in news articles have also been found to affect short-term market returns.[3]

Many of the core technologies in financial text mining are related to the development of natural language processing (NLP) tools. For example, the interest in opinion mining has stimulated the development of various tools, and subsequently facilitates studies regarding the relationship between sentiments and stock markets.

One promising research direction in finance text mining is the recognition of *certainty*, which is the state of being free from doubt. Victoria Rubin, Elizabeth Liddy, and Noriko Kando proposed annotating certainty using a four-category scheme: absolute, high, moderate, and low.[4] But, according to their manual-annotation studies, a two-category annotation scheme might be more appropriate, given the subtlety of interpreting certainty embedded in text.

The original purpose of certainty recognition was to distinguish "hard" facts from those with doubts in the context of information retrieval. It is interesting to note that the other side of the coin, uncertainty recognition, is closely related to the problem of recognizing risk-related statements in textual data. Risk-related statement recognition could help process large amounts of textual data and filter out statements related to the assessment of a firm's future value. Risk recognition could potentially improve the speed and quality of the decision-making process.

For example, consider a bank that provides loans to hundreds of institutional clients. This bank might want to monitor documents related to its clients on Web 2.0 and pay special attention to potentially good or bad developments that could affect its clients' ability to repay their debts. If a system can automatically recognize statements that contain information about firms' potential developments, then useful indicators based on the recognized statements can be constructed to help the bank more effectively monitor its clients.

The growing body of Web 2.0 content can facilitate the implementation of this kind of near real-time monitoring system and allow financial institutions to benefit from the continual stream of documents. At the core of this technology is the automatic recognition of risk-related statements in textual data. Unfortunately, current literature doesn't provide sufficient guidance on implementing such systems. Existing studies on certainty recognition focus on developing manual annotation schemes rather than on constructing working systems for monitoring massive amounts of textual data.

Automatic Risk Recognition

Risk is an important concept for decision making. Although various metrics have been developed to quantify risk in numerical data, there are few studies that provide guidelines for recognizing risk-related statements in textual data. In our research, we focus on recognizing risk-related statements related to making decisions on the basis of a firm's future value.

We define a statement as risk related if it imparts information that could affect investors' beliefs about a firm's future value. There are three important concepts in this definition: future timing, firm value, and uncertainty. First, statements that are not related to firm value are not considered. Second, risk is future centric, so the embedded information must have implications for the future. Finally, risk is uncertain, meaning it's possible that more than one outcome or state will occur. A common way to represent uncertainty is by attaching probabilities to possible outcomes. These probabilities are embedded in investors' beliefs regarding a firm's future development.

Figure 1 presents the conceptual model for risk recognition. The three basic concepts just discussed guide investors' judgment in recognizing risk-related statements. A risk-related statement can be further classified according to the direction of impact, which could be positive, negative, or both.

Figure 1. Conceptual model for risk-related statement recognition. Three basic concepts guide investors' judgment in recognizing risk-related statements: future timing, firm value, and uncertainty.

Table 1 presents examples of risk-related statement recognition. The first statement reports decreased market share prediction, which has a direct negative impact on the firm's future value. The second statement reports past company performance, which has no, or at most a weak, connection to its future performance. The third statement reports an upgrade of a credit rating, which can be interpreted as a better financial condition for the future. The first and third statements are risk related, whereas the second is not.

Table 1. Examples of risk-related statement recognition.	
Statement	**Risk impact**
Although many analysts had predicted that the market for implantable cardioverter defibrillators (ICDs) would grow by about 20 percent per year, because of an aging population many now forecast only single-digit annual percentage growth.	Negative
Hitachi returned to profitability in its fiscal third quarter with strong sales of digital products, such as hard-disk drives and liquid-crystal displays, as well as sales of its securities holdings.	None
Deutsche Banc Alex. Brown boosted its credit rating on the Denver telecommunications service provider (Qwest Communications International).	Positive

We formulate risk recognition as three binary classification tasks at the sentence level. The first task is to decide whether a sentence contains risk-related information. The second task is to decide whether the direction of the impact is positive. The last task is to decide whether the direction of the impact is negative. A sentence can have both positive and negative impacts.

We adopted supervised-learning approaches to construct the recognition system. To facilitate the learning process, we had to convert each sentence to a numerical representation. We also included commonly used representations such as unigrams (individual words), bigrams (two-word phrases), trigrams (three-word phrases), part-of-speech (POS) tags, and stylistic features.[5] We used POS tags to capture a sentence's syntactic aspects, which concern expressions of future timing and uncertainty. We also included an epistemic modality lexicon to capture uncertainty. Finally, we included the General Inquirer dictionary (www.wjh.harvard.edu/~inquirer) to capture the underlying meanings of words.

Our study considered two statistical machine-learning approaches: support vector machines (SVM) and elastic-net logistic regressions (ENETs).[6] SVMs consistently deliver good performance across different classification tasks. ENETs conduct feature selection and model learning in one integrated step and achieve better performance than SVMs in some testing data sets. These two models represent current state-of-the-art statistical machine-learning approaches.

In our experiment, we considered the following classical performance measures for text classification: accuracy, recall, precision, and F-measure (the harmonic mean of recall and precision). We evaluated our approach using a random sample of firm-specific news articles from the *Wall Street Journal*, which has a business focus and a fairly high circulation. The research testbed consists of 1,529 sentences from the *Wall Street Journal*. Manual-annotation results indicated that 47 percent of sentences in the testbed were risk related. When the direction of impact was considered, 36 percent of sentences contained a negative impact, and 33 percent contained a positive impact.

Using tenfold cross-validation, we computed classification performance using an SVM and an ENET for the three classification tasks. We tuned the classifiers to optimize either accuracy or F-measure. One interesting observation is that the SVM and ENET achieved similar performance levels. In most cases, the difference was not statistically significant. So, we report the average performance of the two classifiers here.

Table 2 summarizes the performance of risk-related sentence recognition. When the task of recognizing risk impact was considered, SVM and ENET achieved an average accuracy of 70 percent, which was far higher than the accuracy of the majority classifier we used (100 – 47 = 53 percent). (A majority classifier assigns all instances to the majority class, and has often been used as a baseline in text classification research.) The accuracy of recognizing the impact direction was also at the same level (71 percent for positive, and 70 percent for negative). The baseline majority classifier achieved an accuracy of 67 percent (100 – 33) for positive impact and 64 percent (100 – 36) for negative impact. Hence, the improvements from using the machine-learning approaches rather than the majority classifier were smaller for direction identification. The results indicate that recognizing whether a sentence is risk related is easier than identifying the impact direction.

Table 2. Comparison of the average performance of the support vector machine (SVM) and elastic-net logistic regression (ENET) classifiers with that of a majority classifier in recognizing risk impact in risk-related statements.

Task	Classifier	Accuracy (%)	F-Measure (%)	Recall (%)	Precision (%)
Recognizing risk impact (positive or negative)	Average of ENET and SVM	70	70	85	60
	Majority classifier	53	NA	0	NA
Recognizing positive impact direction	Average of ENET and SVM	71	56	72	46
	Majority classifier	67	NA	0	NA
Recognizing negative impact direction	Average of ENET and SVM	70	60	83	48
	Majority classifier	64	NA	0	NA

The F-measure of the three classification tasks was consistent with the classification accuracy. When we optimized the classifiers for F-measure during the training phase, the F-measure for recognizing risk was 70 percent, which was higher than for identifying the impact direction (56 percent for positive, and 60 percent for negative). The recall of risk recognition was 85 percent, and the precision was 60 percent.

These results indicate that, given a set of documents previously not seen by the classifiers, they can recognize 84 percent of all risk-related sentences. For a sentence that the classifiers recognize as risk related, there is a 60 percent chance that the sentence is indeed risk related. Although the recognition results are noisy at the sentence level, we can construct various indicators by aggregating the results from an entire firm or even from an entire industry. The aggregation process can effectively reduce the noise and provide valuable information for decision making.

Our experimental results show great potential for automatic risk recognition in Web 2.0 content. Combined with information-extracting tools such as named-entity recognition and topic classification, our approach can track risks faced by a group of companies and provide timely summaries to highlight important items for further investigation. This approach could potentially accelerate the decision-making process in financial institutions. Possible extensions include the development of automatic trading systems using information provided by various text-mining tools, including the risk recognition tool presented in this essay. Judging from the empirical evidence in past financial text-mining studies, companies could profit from mining textual data if it is implemented correctly.

Those who believe in the efficiency of financial markets might consider developing advanced text-mining technologies to be a waste of time; stock prices adjust quickly to incorporate new information in an efficient market. Nevertheless, in the context of financial decision making, the only way to truly know the value of text-mining techniques is to conduct careful, systematic experiments. The knowledge gained during the experimental process could be a valuable asset for both researchers and practitioners.

Acknowledgments

This work was supported in part by the US National Science Foundation under grant CNS-0709338 and the National Science Council of Taiwan (NSC97-2410-H002-125-MY3).

References

1. R.P. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System," *ACM Trans. Information Systems*, vol. 27, no. 2, 2009, article 12.
2. S.R. Das and M.Y. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science*, vol. 53, no. 9, 2007, pp. 1375–1388.

3. P.C. Tetlock, “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” *J. Finance*, vol. 62, no. 3, 2007, pp. 1139–1168.
4. V.L. Rubin, E.D. Liddy, and N. Kando, “Certainty Identification in Texts: Categorization Model and Manual Tagging Results,” *Computing Attitude and Affect in Text: Theory and Applications*, J.G. Shanahan, Y. Qu, and J. Wiebe, eds., Springer, 2005, pp. 61–76.
5. A. Abbasi, H. Chen, and A. Salem, “Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums,” *ACM Trans. Information Systems*, vol. 26, no. 3, 2008, article 12.
6. H. Zou, and T. Hastie, “Regularization and Variable Selection via the Elastic Net,” *J. Royal Statistical Soc. B*, vol. 67, 2005, pp. 301–320.

Hsin-Min Lu *is pursuing a PhD in the Department of Management Information Systems at the University of Arizona. Contact him at hmlu@email.arizona.edu.*

The biography for ***Hsinchun Chen*** *is on p. XYZ of this issue.*

Tsai-Jyh Chen *is a professor in the Department of Risk Management and Insurance at National Chengchi University in Taipei, Taiwan. Contact her at tjchen@nccu.edu.tw.*

Mao-Wei Hung *is a professor in the Department of International Business at National Taiwan University in Taipei, Taiwan. Contact him at hung@management.ntu.edu.tw.*

Shu-Hsing Li *is a professor in the Department of Accounting at National Taiwan University in Taipei, Taiwan. Contact him at shli@management.ntu.edu.tw.*