Identifying Firm-Specific Risk Statements in News Articles

Hsin-Min Lu[1], Nina WanHsin Huang[1], Zhu Zhang[1], and Tsai-Jyh Chen[2]

[1] Management Information Systems Department, The University of Arizona
1130 E. Helen Street, Tucson, Arizona 85721
hmlu@email.arizona.edu, wanhsin.huang@gmail.com,
zhuzhang@u.arizona.edu
[2] Department of Risk Management and Insurance, National Chengchi University
No. 64, Sec.2, ZhiNan Rd., Wenshan District, Taipei City 11605, Taiwan
tjchen@nccu.edu.tw

Abstract. Textual data are an important information source for risk management for business organizations. To effectively identify, extract, and analyze risk-related statements in textual data, these processes need to be automated. We developed an annotation framework for firm-specific risk statements guided by previous economic, managerial, linguistic, and natural language processing research. A manual annotation study using news articles from the Wall Street Journal was conducted to verify the framework. We designed and constructed an automated risk identification system based on the annotation framework. The evaluation using manually annotated risk statements in news articles showed promising results for automated risk identification.

Keywords: Risk management, epistemic modality, evidentiality, machine learning.

1 Introduction

Risk management has long been a topic of interest for researchers and an important issue for business professionals. From the corporate governance prospective, risk management can help managers identify important adverse future events a company may be facing and help establish procedures to measure, report, mitigate, and manage risk. For investors who hold the stocks or bonds of a company, risk management can help them assess potential losses and adjust their portfolios accordingly.

To be able to achieve the expected benefits of risk management, one needs to be able to collect and analyze relevant information from a broad range of data sources. Many of these data sources can be found within a company. For example, records for IT failure, system downtime, or errors caused by production systems [1] are valuable indicators for risks related to production technology. Public data sources such as newspapers and newswires, on the other hand, also play an important role in providing critical information for risk management. For example, various mergers, acquisitions, and business development events, which are important indicators for strategic risk [1, 2], can be found in public data sources.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 42–53, 2009.

One important characteristic of public data sources is that the majority are textual data. News articles typically report a company's past, current, and possible future events considered important for its stakeholders. The reporters of these events, based on their professional judgment, may hint at the potential impacts of these events. Investors need to digest these textual data before meaningful decisions can be made.

The advance of information technology has made the task of information retrieval much easier. For example, by using popular search engines, one can easily locate a set of documents relevant to a company. Only some of these documents, however, are relevant for a user interested in assessing a company's risk factors. To the best of our knowledge, there are few information systems that can help their users further refine and digest risk-related information. Current technology limitations force users to conduct manual analysis on a set of documents that is beyond their capacity.

This information digesting bottleneck may be addressed by systematically studying the characteristics of risk and how the messages are conveyed through textual expressions. Eventually, an information system can be built to help investors better analyze the information and make educated decisions.

Few previous studies have addressed the problem of identifying risk-related statements in textual data or attempted to construct information systems that can assist users in performing relevant tasks. In this study, we proposed a framework for identifying risk-related statements based on previous economic, managerial, linguistic, and natural language processing research. We conducted a preliminary study to verify the framework by manually annotating news articles from the *Wall Street Journal*. A prototype information system that can identify risk-related statements was designed and evaluated.

The rest of the paper is organized as follows. Section 2 summarizes previous studies that are relevant to our goal. Section 3 presents research gaps and questions. We present the annotation framework for risk-related statements in Section 4. Section 5 summarizes the annotation results. In Section 6, we describe the information system that can automatically identify risk-related statements in textual data. The system is evaluated using the human annotated data created in Section 5. We conclude the discussion in Section 7.

2 Background

Rich literature exists for risk management. Most studies have analyzed the problems from an economic or managerial perspective. The other branch of literature comes from current linguistic and natural language processing research. We summarize related studies in this section.

2.1 Definition of Risk

Risk is the possibility of loss or injury [3]. In the context of risk management for business organizations, it can be interpreted as the events and trends that can devastate a company's growth trajectory and shareholder value [1, 4]. Risk is also interpreted simply as "uncertainty" in some microeconomic discussions [5]. These definitions reveal three basic dimensions of risk: timing, uncertainty, and company value.

The first dimension is timing: risk can only exist in events that have not yet happened. The second dimension, uncertainty, is one of the most important characteristics of risk. Uncertainty can be interpreted as the possibility that more than one outcome may occur, which is a typical setting for discussing the decision making process under uncertainty (see, for example, [5]). Finally, risk must have certain impact on company value. The company value can be defined as the company's expected market value.

Note that some studies only consider losses as a contributing factor of risk. Other authors adapt a broader definition and consider the deviation from expectation as risk. For purposes of our study, we adopted the narrower definition and consider an event risky if and only if it may occur in the future and have a negative impact on company value.

2.2 Natural Language Processing Perspective

Despite the lack of direct treatments of risk-related statements in textual data, some studies in natural language processing (NLP) did provide relevant insights. We summarize here the research of subjectivity identification and certainty identification that are related to risk assessment.

Subjectivity Identification. The studies of subjectivity in NLP focus on recognizing expressions that are used to express opinions, emotions, evaluations, and speculations [6, 7]. Articles from newspapers or newswires often contain both subjective and objective expressions. Distinguishing these two kinds of expressions is valuable in applications such as information retrieval or document summarization.

Early studies of subjectivity identification focused on document and sentence level annotation. Annotators were asked to make the judgment of "whether or not the primary intention of a sentence is to objectively present material that is factual to the reporter" [6]. Later studies moved into expression level (words and phrases). Detailed annotation schemes for subjective sentences including the source and target of private state, together with the intensity and polarity of private state used [7]. Note that uncertainty is considered one kind of subjective information. However, this dimension is not emphasized in this line of research.

Certainty Identification. Certainty is the quality or state of being free from doubt, especially on the basis of evidence about the past, present, or future factual or abstract information, expressed by the writer or reported by the writer about others, directly or indirectly involved in the events in the narrative. Certainty is a pragmatic position instead of a grammatical feature.

To capture certainty from textual expressions, Rubin and Liddy [8] proposed a categorization model that characterizes certainty by four dimensions: degree of certainty, perspective of narration, focus, and timing. Four degrees of certainty (absolute, high, moderate, and low) were proposed. They also annotated whether the message was reported from the writer's point of view or from a third party's perspective. Focus was divided into abstract information (opinions, judgments) or factual information (concrete facts). Finally, past timing referred to completed or recent states or events; present timing referred to current, immediate, and incomplete states or events; and future timing referred to predictions, plans, warnings, etc.

Previous studies manually annotated news articles at both the expression and sentence levels [8-10]. Although expression level annotation can provide the richest information, the inter-rater agreement, measured by Cohen's kappa, was low [10]. It indicates that certainty information may be the results of a complicated interaction of various expressions. Although sentence level inter-rater agreement has not been studied yet, it is reasonable to expect better results.

2.3 Linguistic Perspective

Epistemic modality and evidentiality are two semantic categories that are closely related to our study. Epistemic modality is speaker's evaluation toward an event while evidentiality is concerned with the source of information. We discuss them below.

Epistemic Modality. Epistemic modality is "concerned with the speaker's assumptions, or assessment of possibilities, and, in most cases, it indicates the speaker's confidence or lack of confidence in the truth of the proposition expected" [11]. Linguistic devices for epistemic modality include: modal auxiliaries (e.g. could, may, should), epistemic lexical verbs (e.g. appear, seem, suggest), epistemic adverb (e.g. perhaps, possibly, supposedly), epistemic adjectives (e.g. likely, doubtful, apparent), and epistemic nouns (e.g. proposal, suggestion, probability).

Evidentiality. Evidentiality can be narrowly defined as the source of information. A wider definition also include the speaker's attitude (e.g. attested, reported, inferring) toward the source of information. Some authors believe that evidentiality is part of epistemic modality while others believe that these two concepts can be separated . The distinction can be made by noting that "evidentials assert the nature of the evidence for the information in the sentence, while epistemic modals evaluate the speaker's commitment for the statement." [12].

3 Research Gaps and Questions

It is clear from the above discussion that most risk management studies incorporated only numerical accounting and capital market data. Little research has systematically investigated suitable dimensions for risk-related text quantification. We have seen few studies that attempted to develop automatic information systems to help identify and quantify risk-related information.

We notice that, from the previous NLP studies, expression level annotation is the most time consuming and most difficult to conduct. Despite the rich information one may obtain, expression level annotation results are much noisier than the sentence level and document level annotations. Since sentences in a document may or may not contain risk-related information, document level annotation is not a reasonable choice. Sentence level annotation, on the other hand, can provide detailed information with less noise. The results can be aggregated to the document level or, with the help of various statistical methods, drilled down to the expression level.

News articles are one popular data source for obtaining risk-related information. As a result, we chose to focus on firm-specific news articles from the *Wall Street Journal*, one of the most widely read newspapers today.

We set forth a preliminary study that aimed at: 1) developing a suitable framework for quantifying sentence-level risk-related information in news documents, and 2) developing an information system for automatic risk identification and quantification.

4 Risk Quantification Framework

We developed the risk quantification framework based on the characteristics of risk defined in economic and managerial literature as well as various semantic categories mentioned in previous NLP and linguistic studies.

A news sentence can be analyzed using the following five dimensions: timing, epistemic modality, evidentiality, abstract or factual information, and polarity. For the purpose of extracting risk-related information, we focused on the potential impact on three key aspects: future timing, uncertainty, and company value. A risk-related statement can be defined as a sentence that includes future timing, indicates uncertainty, and implies negative impact on company value. We will discuss the framework in detail using sample news articles statements below.

*Kodak, based in Rochester, N.Y., said it **expects** net earnings of $1.15 to $1.30 cents a share for the full year.* (1)

*Mercedes **will** invite potential buyers to hot restaurants and special concerts where it will let them test-drive a C240 or C320.* (2)

Analysts had forecast the company would report earnings of 90 cents a share. (3)

Future Timing. Future timing refers to the expressions that indicate (possible) upcoming events or states. For instance, "expects" in (1) and "will" in (2) indicate future timing. Note that (3) does not have future timing because the forecast is for past company earnings. Binary classification (yes/no) is used for this dimension in our annotation study.

*Although Univision has emerged as the **likely** buyer, Disney and Tribune also have expressed interest.* (4)

*National Semiconductor Corp. cut its **revenue forecast** for the fiscal second quarter, citing inventory and backlog adjustments from customers and distributors.* (5)

Uncertainty. Uncertainty can come from epistemic modality, evidentiality, or both. The expression "likely" in (4) is an example of uncertainty that comes from epistemic modality. In (5), "revenue forecast" indicates the potential uncertainty that is associated with the source of the information.

Various levels of uncertainty can be inferred from the expression. Previous studies had tried to divide the certainty-uncertainty continuum into four or five categories [8-10]. As a preliminary study, we decided to perform a binary classification for this dimension first. A sentence is classified as uncertain or not uncertain.

***Disappointing** ICD sales were offset by a 7% **increase** to $248 million in sales of pacemakers.* (6)

Brown ***boosted its rating*** *on the Denver telecommunications services provider.* (7)

Boeing ***fell*** *2.55 to 58.02.* (8)

Company Value. This dimension captures how the information impacts the market's expectation of the company value. Both abstract and factual information can contribute to this dimension. The polarity in a sentence may also hint at the direction of the impact. It should be pointed out that this dimension cannot be treated as a three-way classification (good, bad, no effect), as a result of our focus on sentence-level information. One sentence may have both good and bad implications. For example, the first half of (6) has a negative implication while the second half has a positive implication. One possibility is to annotate for the net effect on company value. However, the meaning of the original message may be distorted by assuming that only the net effect matters.

We therefore considered company value along two separate dimensions: good news and bad news. A sentence belongs to the good news category if the underlying message has a positive implication for company value and vice versa. A sentence can belong to both categories simultaneously.

Another complication arose from the complex nature of assessing company value. Assessments may differ because of different levels of world knowledge of annotators. It is unrealistic to assume that an annotator (and for that matter, a computer algorithm) can have perfect knowledge about the world. A practical solution is to assume that the annotator has basic business knowledge but does not possess detailed knowledge of a particular event. Under this assumption, (7) is good news because the company rating is a good indicator for the overall financial status of a company; (8) is bad news because stock price reflects current market value of a company.

5 A Manual Annotation Study Using the Wall Street Journal

We annotated news sentences from the Wall Street Journal (WSJ) based on the framework proposed in the previous section. The research test bed is described first, followed by a discussion of the annotation results.

5.1 Research Testbed

We used the following procedure to create our research test bed. First, we drew a random sample of 200 news articles from the WSJ published between 8/4/1999 and 3/2/2007. One annotator manually filtered out non firm-specific news, leaving 103 firm-specific articles. For each firm-specific article, the full text was split into sentences. Each sentence was attached to the original document ID. An Excel file that contained the sentences and document IDs was created. The original order of the sentence in an article was preserved. Only the first 988 sentences from 46 articles were made available to the annotator.

The distribution of article length (measured by # of sentences) is bimodal. A large number of articles contained less than 10 sentences. Another peak was at the group of 30-40 sentences per article. Separating the two groups by a threshold of 20 sentences

per article, we can see that 28 out of 46 articles had length less than or equal to 20. However, this group only contributed to 235 of the total 988 sentences.

The annotator read the sentences in the Excel file in sequence. The experience was much like reading the original article except that the sentences had been split and listed row by row. Each sentence was annotated with four 0s or 1s indicating whether it belonged to the four dimensions proposed: future timing, uncertainty, good news, and bad news.

5.2 Annotation Results

On average, 14.3% of sentences were marked with future timing, 17% uncertainty, 12.9% good news, and 24.6% bad news. The proportion of risk-related sentences (those with future timing, uncertainty, and bad news simultaneously) was only 4.7%. The number of bad news sentences was about twice as much as those with good news.

Separating the collection into short and long articles (at the cutoff point of 20 sentences per article), we can see clearly that these two types of news articles had different distributions in terms of the four dimensions under consideration. Short articles contained more future timing and uncertainty than the longer ones. On the other hand, bad news sentences were more prevalent in long articles. Interestingly, the proportion of risk-related sentences was about the same in these two types of articles.

The differences may reflect the nature of these two types of articles. Long articles, in many cases, contained comments and analysis of current and past events. These articles might be, in part, stimulated by recent bad news about a company. Short articles, in many cases, were quick notices of recent and future events.

6 An Automatic Risk Identification System

Based on the risk quantification framework and the manual annotation results, we designed a risk identification system to extract risk-related information at the sentence level using the four dimensions proposed. Each dimension was separately modeled.

For each input news articles, the system first splits the full text into individual sentences. A feature extraction routine then converts each sentence into the bag-of-words representation. Four binary classifiers are used to identify information related to future timing, uncertainty, good news, and bad news. The scores from the 4 classifiers then can be used to compute the risk scores of a sentence.

It is clear from the design that the performance of the system depends heavily on the four classification modules. As a preliminary study, we decided to evaluate the system performance based on the four classifiers individually.

6.1 Features

We adopted the baseline bag-of-words representation, together with the part of speech (POS) tags. The Porter stemmer was used to find the lemma of individual words [13]. Four combinations of features are considered:

1. Bag-of-words (no stemming)
2. Bag-of-words (stemmed)

3. Bag-of-words (no stemming) with POS tags
4. Bag-of-words (stemmed) with POS tags

When POS tags were considered, the same word with different POS tags was treated as a different feature.

6.2 Machine Learning Approaches

We considered two popular machine learning approaches in this study. The first approach is the maximum entropy (Maxent) model [14]. This model is mathematically equivalent to the discrete choice model. To handle the large number of features, Gaussian priors with mean 0 and variance 1 is imposed on the parameters. The second model considered is the support vector machine (SVM) classifier with a linear kernel [15].

6.3 Baseline Models

We considered two baseline models in this study. The first baseline model is the majority classifier. The majority classifier classifies each instance to the majority class. For example, 14.3% of sentences in the test bed have future timing. The majority classifier assigns all sentences to the class of no future timing since the majority of sentences do not have this tag.

The second baseline classifier is the agnostic classifier. The agnostic classifier assigns a random score between 0 and 1 to each instance. Given a threshold, all instances above the threshold are positively classified while instances below the threshold are negatively classified. The ROC curve of an agnostic classifier is a straight line connecting the origin to (1, 1) [16]. Note that the majority classifier is a special case of the agnostic classifier. Depending on whether the negative or positive tagging is the majority class, an agnostic classifier with cutoff equals 0 or 1 is equivalent to the majority classifier.

6.4 Performance Measures

We considered accuracy, recall, precision, and F-measure in this study. Let TP, FP, TN, and FN denote true positive, false positive, true negative, false negative. These measures can be computed as follows:

Accuracy = (TP+TN) / (TP+FP+TN+FN)
Recall = TP / (TP+FN)
Precision = TP / (TP+FP)
F-measure = 2 * Precision * Recall / (Precision + Recall)

Accuracy is the probability that a case is correctly assigned. Recall is the probability that a positive case is correctly assigned. Precision is the probability that an assigned positive case is correct. F-measure is the harmonic mean of precision and recall. Most of these measures had been used in prior classification studies [17, 18].

6.5 Experimental Results

We used 10-fold cross validation to test the performance of the 4 classifiers. Different performance can be achieved by choosing different threshold values. To make the subsequent discussion easier, we present the figures that maximized the F-measure for each classification task. Tables 1 and 2 summarize the performance of the 4 classification tasks. The bag-of-words features with no stemming were denoted as "B" in the parentheses of the first column; "BS" denotes bag-of-words with stemming. We reported figures that maximized the F-measure. Adding POS tags had only minor impacts on the performance. The outcomes using POS tags, as a result, were omitted to save space.

The majority classifiers have accuracy rates of 85.7%, 83%, 75.4%, and 87.1% for future timing, uncertainty, bad news, and good news, respectively. However, since the majority classes are all negative (the absent of a characteristic), the recalls were all zero and the precisions could not be calculated (since there were no positively assigned cases). Compared to the performance of the Maxent and SVM models, the accuracy was similar but the recalls were much lower.

The agnostic classifier, which can be considered as a generalization of the majority classifier, achieved much lower performance compared to the Maxent and SVM models. The accuracy across four classification tasks was at the range of 21.6% to 33.7%. The F-measures were all less than 39.5%. It is interesting to observe that the recall rates were much higher than those for precision. One possible reason is that the agnostic classifier cannot distinguish cases with different tags. As a result, the only way to boost the F-measure is to have positive assignments to most instances.

The SVM model outperformed the Maxent model in all feature-task combinations. On average, the F-measures of the SVM model were about 10% higher than those of the Maxent model. In most cases, the performance gaps came from higher recall of the SVM model.

In most classification tasks (future timing, uncertainty, and good news), stemming improved the performance. However, the best performance of bad news classification came from bag-of-words without stemming. Error analysis shows that stemming mapped words with different semantics to the same lemma (e.g. "willfully" to "will") caused false positives in subsequent classification.

Table 1. Performance Summary for Future Timing and Uncertainty

	Future Timing				Uncertainty			
	Acc.	Recall	Prec.	F	Acc.	Recall	Prec.	F
Majority	85.7	0	NA	NA	83.0	0	NA	NA
Agnostic	22.4	92.9	14.8	25.5	21.6	97.0	17.5	29.6
Maxent (B)	91.7	57.4	78.6	66.4	82.0	42.3	46.7	44.4
SVM (B)	93.1	75.2	76.3	75.7	83.4	72.0	50.8	59.6
Maxent (BS)	91.8	55.3	81.3	65.8	82.3	42.9	47.7	45.1
SVM (BS)	**94.0**	**75.9**	**81.1**	**78.4**	**87.9**	**56.0**	**67.1**	**61.0**

Table 2. Performance Summary for Bad News and Good News

	Bad News				Good News			
	Acc.	Recall	Prec.	F	Acc.	Recall	Prec.	F
Majority	75.4	0	NA	NA	87.1	0	NA	NA
Agnostic	24.6	100.0	24.6	39.5	33.7	80.3	13.9	23.7
Maxent (B)	73.3	44.4	45.6	45.0	80.0	47.2	31.4	37.7
SVM (B)	**73.4**	**62.6**	**46.9**	**53.6**	88.1	40.9	54.7	46.8
Maxent (BS)	62.3	68.3	36.0	47.2	83.9	40.9	38.2	39.5
SVM (BS)	72.1	64.2	45.2	53.1	**88.4**	**45.7**	**55.8**	**50.2**

To gain a better understanding of the classification task, we computed the conditional probability associated with each lemma from the Maxent model. The model is trained using all training data with words stemmed. The conditional probability is the probability that the Maxent model will have positive assignment given a particular lemma. For example, given the lemma "outlook," we computed the conditional probability that a sentence belonged to the future timing category given that the sentence contained only the word "outlook." A similar procedure was repeated for the remaining three dimensions. The conditional probability is a good proxy for the importance of each lemma.

We sorted the conditional probability of each lemma in descending order. Top 10 lemmas from each classifier were analyzed. We observe from the results that lemmas such as "will," "expect," and "estim" were good indicators for future timing. Lemmas such as "expect," "if," and "may" hinted at uncertainty. Bad news sentences may contain lemma such as "fall," "problem," or "risk" while good news sentences may contain lemma such as "strong," "unit," or "rose."

Note that some lemmas were not semantically related to the underlying classes. For example "cable-tv" in future timing, uncertainty, and bad news did not have a clear connection to these three dimensions. One possible reason is that our training dataset was small and the Maxent model over-generalized from this particular dataset.

7 Concluding Remarks

We developed an annotation framework for risk identification based on the previous literature. The framework models risks along four dimensions: future timing, uncertainty, good news, and bad news. We applied the framework on firm-specific news articles from the *Wall Street Journal*. The annotation results showed that bad news is the most commonly annotated dimension across the four dimensions considered. We designed an automatic risk identification system based on the annotation framework and trained the underlying classifiers using the manual annotation results.

Using the bag-of-words representation, we achieved F-measures between 50.2% and 78.4% for the four classification tasks under consideration. Important features of these four classifiers showed consistent semantics as indicated by the definitions of these four dimensions. The results are promising for the development of a full-fledged system.

We are currently recruiting and training more human annotators to conduct manual annotation based on the proposed framework. The validity of the framework can be further confirmed by analyzing the annotation results from multiple sources. We also plan to continue the research by developing an information system that can automatically identify risk-related statements from various business-related news sources.

Acknowledgments. This work was supported in part by Taiwan National Science Council through Grant #NSC97-2410-H002-002 and U.S. National Science Foundation through Grant #IIS-0428241 ("A National Center of Excellence for Infectious Disease Informatics"). The authors wish to thank Dr. Hsinchun Chen and Dr. Shu-Hsing Li for their constructive comments.

References

1. Slywotzky, A.J., Drzik, J.: Countering the Biggest Risk of All. Harvard Business Review, 1–11 (2005)
2. Chapelle, A., Crama, Y., Hubner, G., Peters, J.-P.: Practical methods for measuring and managing operational risk in the financial sector: A clinical study. Journal of Banking & Finance 32, 1049–1061 (2008)
3. Merriam-Webster: Risk: Definition from the Merrian-Webster Online Dictionary (2008)
4. COSO: Enterprise Risk Management - Intergrated Framework. COSO(Committee of Sponsoring Organizations of the Treadway Commission) (2004)
5. Mas-Colell, A., Whinston, M., Green, J.R.: Microeconomic Theory, Oxford (1995)
6. Bruce, R.F., Wiebe, J.M.: Recognizing subjectivity: a case study of manual tagging. Natural Language Engineering 1, 1–16 (1999)
7. Wiebe, J., Wilson, T., Cardie, C.: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation 39, 165–210 (2005)
8. Rubin, V.L., Liddy, E.D.: Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In: Shanahan, J.G., Qu, Y., Wiebe, J. (eds.) Computing Attitude and Affect in Text: Theory and Applications. Springer, Heidelberg (2005)
9. Rubin, V.L.: Certainty Categorization Model. In: AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications, Stanford, CA (2004)
10. Rubin, V.L.: Starting with Certainty or Starting with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In: Proceedings of NAACL HLT 2007, pp. 141–144 (2007)
11. Coates, J.: Epistemic Modality and Spoken Discourse. Transactions of the Philological Society 85, 110–131 (1987)
12. de Haan, F.: Evidentiality and Epistemic Modality: Setting Boundaries. Southwest Journal of Linguistic 18, 83–101 (1999)
13. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14, 130–137 (1980)
14. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Comput. Linguist. 22, 39–71 (1996)
15. Joachims, T.: Making large-Scale SVM Learning Practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT-Press, Cambridge (1999)

16. Cortes, C., Mohri, M.: Confidence Intervals for the Area under the ROC Curve. In: Advances in Neural Information Processing Systems (NIPS 2004), vol. 17. MIT Press, Vancouver (2005)
17. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. 26, 1–29 (2008)
18. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam (2005)