www.joim.com

Journal Of Investment Management, Vol. 15, No. 1, (2017), pp. 39–49

A MACHINE LEARNING APPROACH TO RESEARCH CURATION FOR INVESTMENT PROCESS

Sonya Cates[a], *Stephen Lawrence*[b,*], *Carla Penedo*[c] *and Viktoriia Samatova*[b,†]

Many investment professionals consider academic research instrumental in improving the quality of the investment process. However, it is hard to extract investment insights from the vast and rapidly expanding research corpus, which requires a large amount of time and human effort in order to absorb. We offer a novel solution to this problem by introducing a machine learning approach to research curation. By comparing the performance and accuracy of humans and machines, we show that a machine learning approach approximates the quality of human curation but offers the strategic benefits of scalability, efficiency and lower cost.

What is the role of machine learning in the curation of academic research? Social Science Research Network (SSRN) reports that it received 67,318 papers over a 12-month period.[1] This volume presents a challenge for investment professionals who want to use insights from academic research to generate profitable investment strategies. Hiring a small army of experts who can read academic research and select the best ideas is expensive and time consuming.

Machine learning techniques have been promised as a scalable alternative to lower-level cognitive tasks for which a human is overqualified and inefficient as a resource.[2] Certainly, formulaic tasks can easily be proxied by an algorithm, but how successful are classification algorithms at proxying for something as abstract as the editorial curation of human research? Is there an optimal mix of human and machine curation that outperforms either an entirely human or entirely algorithmic curation? What are the risks of relying heavily on a machine-based classification algorithm for selecting areas of research focus? And by automating the selection and classification

[a]Roger Williams University, USA. Tel: +1 401-254-3482, E-mail: scates@rwu.edu.
[b]State Street Corporation, Portugal.
[*]Tel.: +1 617-662-5623; E-mail: sclawrence@statestreet.com.
[†]Tel.: +1 647-775-5426; E-mail: vsamatova@statestreet.com.
[c]Novabase, USA. Tel: +351 21 383 6300, E-mail: carla.penedo@novabase.pt.

steps, can we free up researchers to focus on more complex aspects of a typical research workflow?

In this paper we offer an alternative solution for this problem by applying machine learning techniques to academic research curation. We seek to benchmark the performance of machine-based classification algorithms against the inherent variability of human classification of academic papers for the purpose of idea generation in investment management.

The idea is simple: in order to select the best insights from academic research, at the very least one needs to (1) determine if a given paper is relevant for investment professionals; and (2) organize papers by predefined categories to help an investor understand what kind of strategy each paper is referring to and what asset classes are involved. Machine learning algorithms are able to successfully address both of these questions and streamline the curation process.

We have collected a unique dataset based on a human editorial review process for evaluating academic research conducted over a period spanning February 2010–September 2016. This process of human curation seeks to identify relevance for investment management and classify research based on pre-set list of asset, topic and region "tags". This review process relies on several rounds of evaluation by finance industry professionals with graduate degrees in economics or finance who have vast experience reading academic research as well as constructing and testing investment strategies. The process requires two researchers look at the same paper to mitigate behavioral selection bias. By comparing relevance ratings by two independent researchers, we find agreement 83% of time—suggesting an upper limit on the consistency of qualitative assessment by human editors. Our model for relevance determination is in agreement with a human researcher 78% of the time, which is a close match to human performance. Moreover, tagging model accuracy is 87% on average.[3] Given the time and cost saving benefits of a machine learning approach to both tasks, this is an attractive alternative to the human research curation model.

Our approach to determining relevance and topic is based on the textual content of the research papers. In order to apply machine learning methods to text, each document must be represented as a set of features. We calculate the features for each document by measuring word frequencies. This method of representing text is known as a bag-of-words approach. Resulting measures of word frequency for papers classified by a human reviewer are then used to create a model for the classification of unlabeled documents. A number of different models may be used in this step, including the logistic regression and support vector machine approaches that we employ and describe in more detail in Sections 2 and 3.

The general problem that we are addressing in this work is one of automated text classification, i.e., the problem of assigning a given document to one or more predefined categories. Text classification is a well-studied area in the field of machine learning, and many approaches have been developed and applied to a variety of diverse problems. We review the work most closely related to ours and refer interested readers to overviews of the field by Dalal *et al.* (2011) and Sebastini (2002), which provide a more comprehensive review of existing algorithms and techniques.

Several machine learning approaches have been applied to the problem of classifying research papers. Taheriyan (2011) seeks to classify computer science research papers by mapping the relationships among authorship, citations, and references and then using those relationships to infer subject classifications for unclassified papers based on their similarity to a set of human-classified papers. However, the approach does not consider the textual content of the unclassified papers. Such an approach may prove to be a useful complementary approach to our content-focused method of classification.

Papadatos *et al.* (2014) address the problem of research curation in the field of chemistry. Their work aims to distinguish between scientific publications that are relevant for a particular field of chemistry (small molecule drug discovery) and those that are not relevant for this area. Specifically, they focus on prioritizing documents so that extraction of chemical structure information from relevant publications by a human may be conducted in a more efficient manner. This is similar to our goal in that we do not intend, or expect, an automated approach to replace human involvement in the research curation process. Rather, we seek to deploy high-level human abilities more effectively by automating lower-level cognitive tasks.

Automated approaches to analyzing and classifying documents have also been applied to the field of finance. Prior applications to textual data in finance have focused on news stories and corporate financial reports rather than financial research, the focus of this work. Applications to financial news stories include work by Tetlock *et al.* (2008). They predict firm-level earnings and stock returns based on the fraction of negative words in news stories. Work by Manela and Moreira (2015) constructs a measure of "news implied volatility" from *Wall Street Journal* articles, which is related to the VIX.

Machine learning applications to analyzing corporate financial reports include work by Vilain *et al.* (2007) who apply multiple methods to the problem of processing and classifying information contained in the tables of such reports published via the web. We encounter a related problem in that many of the documents we seek to categorize contain information in tables. Currently we exclude this information; however, in the future we hope to incorporate the information contained in tables and figures. Recent work by Gao (2016) also addresses the problem of extracting information from corporate financial filings. This work applies a method similar to ours (a regression approach applied to a bag-of-words model) to SEC financial filings in order to measure the informativeness of the filings with regard to macroeconomic market conditions.

This paper is organized as follows. In Section 1 we introduce the data used for model training and testing. Section 2 reviews the methodology used to predict relevance, while Section 3 describes our categorization algorithms.

In Section 4, we test the ability of machines to determine relevance by comparing the agreement between a machine learning model based on a bag-of-words approach and the human research curation process to the degree of agreement between two independent researchers. We find that the model is in agreement with a human 78% of the time (relative to 83% of human/human agreement). This suggests that our model is almost as good at determining relevance of a given paper as the fully human curation process but offers significant advantages in efficiency, speed and scalability.

In Section 5 we test the ability to automate the classification of research based on asset, topic and region tags, by training a model on a set of academic and financial industry research reports and testing the classification accuracy on a sample of 700 documents. The resulting average F-score success rate for different tags is 87%. In Section 6 we conclude. We find that machine learning algorithms are sufficiently accurate to be a cost-effective alternative to a basic human curation process. However, we suggest that these cost efficiencies work best if editorial efforts instead focus their energy on summarization of content and deeper qualitative assessments of uniqueness or author competence—skills which require significant domain knowledge, and for which we have not yet developed a sufficient machine-based alternative.

1 Data

The primary data for this analysis is a set of 3,735 human-classified academic papers on topics in economics and finance. These papers were presented and/or published in prominent finance conferences and journals from January 2010 to September 2016. The scope of finance research includes micro- and macrofinance, theoretical and empirical models. As part of an ongoing editorial exercise conducted over a six-year period, a team of researchers and investment professionals at State Street sought to identify, categorize and summarize a subset of papers from these sources, noted for their relevance to investment management.

Paper classification was performed by a team of quantitative finance researchers with advanced degrees in economics and finance, prior experience in reading academic papers and expertise in constructing and testing investment strategies. All reviewers receive extensive training on the norms that are sought during the review process. A core editorial team promotes consistency and accuracy across the team.

The structure of the data is as follows: for each paper, we have the paper contents, which are extracted from a PDF, and associated metadata from the review process. The human review process has two rounds—each paper is reviewed by two different people in order to reduce human selection bias. In most cases, the second reviewer is more experienced than the first one. Both reviewers read the paper, however, the second reviewer can see the rating of the first reviewer when making her rating choice. The decision-making process for each paper involves reading the paper and making a decision about its relevance for the investment process. If the paper is of interest to institutional investors because it helps them learn something about asset pricing and, more so, develop investment strategies, the paper is categorized as "selected," otherwise, the paper is "rejected." Other factors affecting reviewer decisions are innovation of research question, good data sample, robustness of methodology and analysis, clear hypothesis and its proof, and paper's authors and their affiliation.

Moreover, selected papers are classified based on their relevance to the investment community by topics (e.g., asset allocation, investor behavior, liquidity, macroeconomics, manager selection, market dynamics, policy, pricing factor, risk management). Topic relevance is determined by the primary use of this paper by investment professionals and the model focus.

In aggregate each record has text contents of the paper itself, information about the paper, separate rating feedback from both stages of the review process (selected/rejected) and, for selected papers, a list of relevant topics associated with the paper.

To create our model for relevance determination, we use a training set consisting of the final ratings and topic classification for exactly 3,000 academic papers over the entire period of 2010–2016. Our primary test set consists of 735 similar papers. For the training set, the total number of selected papers is 1,683 and the total number of rejected papers is 2,052. For the test set, the number of selected papers is 339 and the number of rejected papers is 396. The natural ratio of accepted papers to rejected papers is around 10%. As a result we limit the number of rejected papers used in the sample to keep the training select/reject ratio approximately equal to maximize the efficacy of the training set. We use all available selected papers from the sample period for our analysis. Rejected papers in the training set include all rejected papers from all sources over the period 2015–2016 plus rejected papers from select conferences in 2013–2015.[4]

We also test our model using a second test dataset consisting of papers from a single conference in 2013 for which we solicited ratings and topics through two separate and independent editorial processes. These independent review processes will serve as a baseline for understanding the consistency of human judgment and classification. For this test set, the number of selected papers is 55 and the number of rejected papers is 387. This ratio of selected to rejected papers more closely reflects the ratio of selected and rejected papers in practice.

For our classification model, we used a combination of 2,939 academic and financial industry reports. The reason for using two different sets of data is by construction: while we need to determine relevance of academic papers, financial industry reports are always relevant. However, topic classification is important for both types of documents. The test set for classification contains 700 documents.

We estimate that the combined number of hours spent reviewing research for this sample exceeds 12,000 person hours over the past six years.[5]

2 Methodology for relevance determination

After the human labeling described in Section 1, each document is converted into a text only format and the main text is extracted. This step removes the bibliography and most tables and figures. The next step is to convert each document into a numerically represented set of features. This feature vector, along with the human supplied select/reject label, forms the training set for our model. With this training set we construct a logistic regression model to classify new documents as selected or rejected. We use Scikit-learn and the NLTK toolkit for Python to implement the process described below.

We use a bag-of-words approach to construct feature vectors. In its simplest form a bag-of-words approach uses raw word frequencies, meaning that each document would be represented as a list of numbers (the features), with each number corresponding to the number of times a given word appeared in the document. For example, taking the first two sentences of this paragraph as two documents would result in the word counts and feature vectors shown in Table 1. We apply two main improvements to the basic approach, a 2-gram model and term frequency–inverse document frequency, and describe each of these methods below.

Rather than simply count the frequency of individual words, known as a unigram or 1-gram model, our approach uses a 2-gram model. This means that we also count the frequency of continuous sequences of two words as well as individual words. Using the example sentences above, this means we would now include counts of word

Table 1 Word counts and feature vectors.

	Sentence 1 feature vector	Sentence 2 feature vector
Appeared	0	1
Approach	1	1
As	0	1
Bag-of-words	1	1
Be	0	1
Construct	1	0
Corresponding	0	1
Document	0	2
Each	0	2
…	…	…

pairs such as "feature vectors" and "word frequencies" in our feature vectors. This approach is known generally as an n-gram model. Increasing the value of n can yield more descriptive features; for example, counts of "feature vectors" or "word frequencies" are likely to be a better indicator of the content of this document than counts of the individual words. However, as n increases the required computation time increases significantly and must be balanced with improvements in performance, thus an n-gram model is only practical for very small values of n.

We also consider that words or groups of words that appear frequently in both classes of documents (select and reject) are unlikely to contain useful information for determining whether to select or reject a new document. Thus rather than creating a feature vector for each document based on term frequencies alone we down-weight frequently occurring terms using the term frequency–inverse document frequency as follows:

Let $f_{t,d}$ be the frequency of a term t in a document d, described previously.

The inverse document frequency is defined as

$$idf(t,d) = \log \frac{n}{1+n_t} \tag{1}$$

where n is the total number of documents and n_t is the number of documents that contain the term t. The term frequency–inverse document frequency is defined as the product of term frequency and inverse document frequency:

$$tf \cdot idf(t,d) = f_{t,d} \times \log \frac{n}{1+n_t} \tag{2}$$

Once we have created a feature vector for each document in our training set, we then train a logistic regression model with the human-supplied labels to perform classification of new documents. A logistic regression model determines a new document's classification by calculating

$$p_l = \frac{1}{1+c^{-z_i}} \tag{3}$$

where

$$z_i = w_i^T x = w_0 + w_1 x_1^i + \cdots + w_n x_n^i \tag{4}$$

In Equation (4) x_n^i represents the n-th feature of the i-th document and w_n represents the weight given to that feature. We may then interpret p_l as the probability of a particular document belong to the 'select' class and label documents with a value greater than 0.5 as selected according to our model.

Training such a model involves determining the weights for each feature (word count) using standard gradient descent methods to minimize a cost function, which takes into account a penalty for misclassified training examples as well as a regularization term that penalizes extreme weights in order to reduce overfitting. Our model uses an L2

Table 2 Most heavily weighted features.

Term	Weight
CDS	3.85
Hedge fund	3.46
VIX	3.34
IPO	3.25
Insider	3.19
Sentiment	3.07
Futures	3.01
Institutional ownership	2.93
Illiquidity	2.83
Mutual fund	2.83
Alpha	2.78
Currency	2.71
Predictability	2.67
Risk premium	2.47
Excess returns	2.40

norm for this regularization. Murphy (2012) provides a more thorough treatment of the mathematical details of the method, which we apply with the logistic regression methods of the Scikit-learn toolkit. Some of the most heavily weighted terms resulting from this process (which become document features with the term frequency–inverse document frequency method described above) are shown in Table 2.

3 Methodology for classification labeling

After identifying relevant documents, we further refine the research curation process by classifying documents based on their contents according to their asset class, regional focus and primary topic of research focus. Our goal is to make a large number of industry documents and academic papers searchable to enable researchers to discover relevant insights and generate new investment ideas.

We apply a hybrid approach that combines two main techniques: standard machine learning classification algorithms and similarity metrics that combine several natural language processing (NLP) techniques. The classification algorithms are most effective (and used) when there are enough training documents for a given tag. In this case, we use a state-of-the-art classification algorithm: Support Vector Machine (SVM). The similarity metrics are useful for making classifications when there are few training documents. In this situation we combine several NLP metrics, including Term Frequency, word2vec and Latent Semantic Indexing, to make the classification more robust when there are few training examples (or observations).

SVM classifiers are a machine learning technique that represents observations as points in hyperspace, also called vectors, with dimensionality equal to the number of features used to describe the observations (documents in our case). The SVMs determine a hyperplane that separates the observations of different classes with a vector such that the separation vector is equidistant to the observations that are closest to the frontier, creating an optimal partition of the observations. The power of SVM is its ability to transform a complex classification boundary into more tractable combinations of features that describe the key distinguishing features of a document class, providing a faster way to classify incoming content. In our case, the true relationship between features and document classification can be complex. The SVM technique streamlines the process of identifying key features that determine the approximate classification of the document. Murphy (2012) provides a more complete explanation of the mathematical basis for this method.

We also apply several NLP metrics. Term Frequency, a simple metric, essentially counts term repetition in a document. More complex metrics, including word2vec and Latent Semantic Indexing, take into account the word context and assesses the semantic similarity of words. The NLP metrics are used as features for the SVM algorithm to analyze deeper aspects of the documents and uncover patterns that are more complex.

Classification algorithms learn by generalizing from a given set of examples, in our case a collection of documents in which the occurrences of a set of entities of interest (predefined list of "tags") are manually classified. Each document can have multi-tags classification. The collection represents the training dataset used to predict the occurrences of those entities in a previously unseen set of documents and hence, it is crucial to have a representative dataset that is sufficiently large and of high quality.

We used a baseline of 400 documents tagged by human researchers to create a trained corpus based on a tag tree with three different category

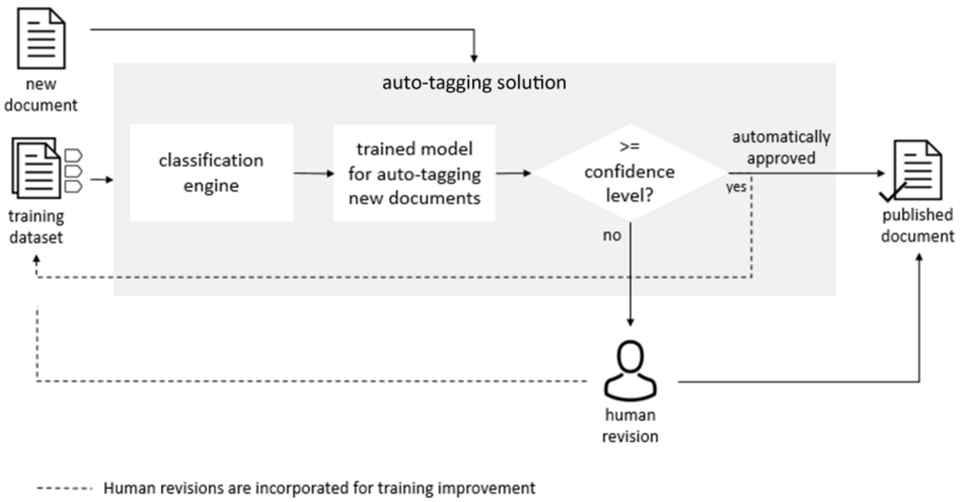

Figure 1 Classification model training process.

tags (asset class – 8, region – 122 and topic – 122). While the creation of such a training dataset was significantly demanding and time consuming, it was completely crucial to guarantee the quality of the corpus (i.e., high confidence in the hand-tagged documents and concept standardization among the elements in the tag tree) and hence, the success of the classification process.

In practice, our auto-tagging method is embedded in a hybrid human/machine workflow which leverages human curation by using a trained classification model to predict tags in new documents and incorporates human feedback into a continual improvement cycle. The overall approach is shown in Figure 1. The classification engine is based on the SVM algorithm (for tags with enough training documents) complemented by similarity metrics (for tags with fewer training documents) that can semantically relate terms, described above. When there is a new document to analyze, the classification engine is used for the automatic classification. If the system is confident with the result (above a predefined threshold), the new document is automatically approved and made available to the end users. Otherwise, skilled researchers review the generated tags and the human tagging is then integrated into the solution and used to update the training model for the continuous and automatic improvement of the system.

4 Performance of relevance classifier

From our set of 3,735 human-classified papers described in Section 1, we reserve 735 to form a test set and train the model described above on the remaining 3,000 documents. Our model correctly classifies 567 of the 735 documents in our test set resulting in a classification accuracy of 77.1%.

Precision and recall are common measures of classifier accuracy. Precision is equal to the number of true positives divided by the sum of the number of true positives and the number of false positives. Recall is equal to the number of true positives divided by the sum of the number of false negatives and the number of true positives. The F-score combines these two measures as:

$$F = 2\left(\frac{precision \times recall}{precision + recall}\right) \tag{5}$$

Our model achieves a precision and recall of 75.7% and 74.3% respectively, yielding an F-score of 75.0%. The confusion matrix containing a breakdown of the model's performance is shown in Table 3.

Table 3 Confusion matrix.

	Model reject	Model select
Human reject	315 (true negatives)	81 (false positives)
Human select	87 (false negatives)	252 (true positives)

Table 4 Relevance model accuracy.

Classifier pair	Agreement
Human1/Human2	83.1%
Human1/Model	78.1%
Human2/Model	77.4%

As discussed in Section 2, humans sometimes disagree when assigning labels, thus to accurately assess our model's potential to automate the process of human curation we also compare the rate of agreement of two humans on the same set of documents with the rate of agreement of our model with a human labeler. For this comparison we use a smaller test set of papers from a single conference that have been independently labeled by two humans. We find that the human labelers agreed 83.1% of the time. Our model's classification of the documents agreed with the humans' 77.8% of the time, an accuracy similar to that reported on the larger test set. These results are presented in Table 4.

5 Performance of classification algorithm

We evaluate the performance of our classification solution in a test set of 700 documents, using an F-score, as defined above. This measure provides a more accurate gauge of performance for datasets that have few examples of some classes. Our solution achieves an average F-score of 87% for the most representative tags (tags presented in more than 50 trained documents), which will improve in time with more documents to be used in training. Table 5 presents the results detailed by category tags for the most representative tags.

Table 5 Classification model accuracy.

Category tags	F-score
Asset class	50–88%
Region	70–100%
Topic	60–100%

The lower F-scores are associated with generic tags that simultaneously classify sell-side research and academic papers, because we are currently using a single training model for both types of documents. These two types of documents can be quite different in the vocabulary used and in the incidence of certain tags, as they are written by different authors and targeting different audiences, which poses a challenge for an automated approach. However, given the inconsistency present in human tagging and the cost of maintaining a significant editorial review process, the ability of the auto-tagging approach to tag documents quickly and consistently provides a cost-effective baseline which can be readily enhanced with additional, limited human supervision.

6 Conclusion and future work discussion

We tested the performance of machine learning models trained to determine the relevance of academic papers and classified them based on pre-specified asset class, topic and region categories. We found that model performance is similar to human research curation but offers advantages of efficiency, scalability, cost and speed. Specifically, we find that when assessing relevance, humans agree with one another 83% of the time while the model is in agreement with human assessments 78% of time. For document classification, our model's average success rate is 87%.

The machine learning research curation process outlined in this paper is a promising alternative to the naïve use of human labor. We expect that a systematic combination of increasingly intelligent algorithms and strategic human oversight will further streamline research curation and should be the topic of future research.

The relevance model can be improved by leveraging other elements of the document typically used during cursory human review. The outlined approach currently extracts only the main portion of a document's text but classifiers could readily make use of information contained in figures and bibliographies, as well as author and publication information using an ensemble learning approach.

Based on our initial findings, differentiating documents by type and source prior to training should dramatically improve the accuracy of classification algorithms for topics that span academic and non-academic research. The primary benefit of the classification algorithm results in accelerated searches for relevant content, so approaches that increase the relevant dimensions of search and granularity of topics are likely to further refine the search process. Such refinements benefit from additional focused training sets and an iterative learning process.

Anecdotal evidence points to the viability and practicality of this approach. State Street has implemented the classification algorithm as part of human–machine curation model for streamlining research organization. The relevance scoring is currently being evaluated as an ancillary tool to assist the human-driven paper selection process to help researchers focus on papers with the highest chance of selection. The combination of machine learning and human expertise boasts the greatest promise for streamlining the research curation process. By pre-screening or scoring research prior to human involvement, researchers can focus their attention on aspects of research that require more significant training, such as assessing the uniqueness or the quality of research. Validating and refining machine classifications is faster and more productive than starting with human classification. Ultimately, the outcome of a research curation process is typically a decision to further explore an idea through implementation or additional research with the interests of a particular audience in mind—a decidedly human decision process which is not easily replicated algorithmically. So, while our approach is changing the way we deploy researchers, we see plenty of future opportunities for editor and researcher alike.

Notes

[1] September 25, 2015–September 24, 2016. "Search eLibrary". Social Science Research Network, Accessed September 25, 2016.

[2] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). "How to Grow a Mind: Statistics, Structure, and Abstraction," *Science* **331** (6022), 1282.

[3] In this case, human accuracy is assumed to be 100% since we only take the final tags and do not differentiate between tagging performance of first and second researcher.

[4] The breadth of content covered varies significantly by source. While the selected papers are homogenous in their content, rejected conference papers are significantly different from rejected journal papers. We bolstered the use of conferences in the training set to more closely resemble the types of content experienced in the test set —which focuses on a single conference.

[5] Estimate based on six years of executing a quarterly cycle requires an average of five researchers to work at least five hours over a twenty-day period.

Disclaimer

The material presented is for informational purposes only. The views expressed in this material are the views of the authors and are subject to change based on market and other conditions and factors, moreover, they do not necessarily represent the official views of State Street Global Exchange^{SM} and/or State Street Corporation and its affiliates.

References

Aggarwal, C. C. and Zhai, C. (2012). *Mining Text Data*. Springer Science & Business Media.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

Dalal, M. K. and Zaveri, M. A. (2011). "Automatic Text Classification: A Technical Review," *International Journal of Computer Applications* **28**(2), 37–40.

Gao, L. (2016). "Applications of Machine Learning and Computational Linguistics in Financial Economics," PhD Thesis, Tepper School of Business, Carnegie Mellon University.

Joachims, T. and Sebastiani, F. (2001). "Guest Editors' Introduction to the Special Issue on Automated Text Categorization," *Journal of Intelligent Information Systems*, Forthcoming.

Knight, K. (1999). "Mining Online Text," *Communications of the ACM* **42**(11), 58–61.

Manela, A. and Moreira, A. (2015). "News Implied Volatility and Disaster Concerns," Available at SSRN 2382197.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Papadatos, G., van Westen, G. J., Croset, S., Santos, R., Trubian, S., and Overington, J. P. (2014). "A Document Classifier for Medicinal Chemistry Publications Trained on the ChEMBL Corpus," *Journal of Cheminformatics* **6**(1), 1.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., and Cournapeau, D. (2011). "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830.

Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization," *ACM Computing Surveys (CSUR)* **34**(1), 1–47.

Taheriyan, M. (2011). "Subject Classification of Research Papers Based on Interrelationships Analysis," *Proceedings of the 2011 Workshop on Knowledge Discovery, Modeling and Simulation* (pp. 39–44). ACM.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). "How to Grow a Mind: Statistics, Structure, and Abstraction," *Science* **331**(6022), 1279–1285.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *The Journal of Finance* **63**(3), 1437–1467.

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). Butterworths.

Vilain, M., Gibson, J., and Quimby, R. (2007). "Table Classification: An Application of Machine Learning to Web-hosted Financial Texts," *Proceedings of the 2007 International Conference on Recent Advances in Natural Language Processing* (pp. 619–623).

Wu, H., Luk, R. Wong, K., and Kwok, K. (2008). "Interpreting TF-IDF Term Weights as Making Relevance Decisions," *ACM Transactions on Information Systems* **26**(3).

Keywords: Machine learning; investment process; research curation; relevance determination; topic classification; automated categorization