Textual Outlier Detection and Anomalies in Financial Reporting

Leslie Barrett
Bloomberg, LP
New York, USA
lbarrett4@bloomberg.net

Sidney Fletcher
Bloomberg LP
New York, USA
sfletcher28@bloomberg.net

Robert Kingan
Bloomberg, LP
New York, USA
rkingan@bloomberg.net

Mrinal Kumar
Bloomberg, LP
New York, USA
mkumar145@bloomberg.net

Alexandra Ortan
Bloomberg, LP
New York, USA
aortan@bloomberg.net

Siddharth Parikh
Bloomberg, LP
New York, USA
sparikh63@bloomberg.net

Anu Pradhan
Bloomberg, LP
New York, USA
apradhan11@bloomberg.net

Ryon Smey
Bloomberg, LP
New York, USA
rsmey@bloomberg.net

ABSTRACT

Outlier detection in text poses unique challenges due to the high-dimensional nature of language data and the difficulty in isolating a single definition of "textual outlier." Textual outliers can be viewed from the perspective of topic, stylistic variation or literal/metaphoric phraseology. This paper applies recent approaches to outlier detection on high-dimensional datasets to textual data focusing on topical outliers. We use the best performing approaches on selected standard corpora to find outliers in a corpus of Risk Factors from U.S. corporate annual reports. Our results show promise in detecting unexpected off-topic language present in this dataset.

Keywords

Text, topical outlier, annual reports

1. INTRODUCTION

According to Hawkins [13], an outlier is an observation which deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Outliers, also known as "anomalies" or "novelties," can appear in a variety of contexts, such as health monitoring systems, sensor networks and even biological ecosystem models.

Compared with more popular applications such as fraud or intrusion detection, outlier detection in text has received relatively little attention in the literature. Text poses unique challenges in part because of the multiple ways that "anomalous" text can be defined and, in part, because of the high-dimensional nature of textual feature representation.

Outlier detection models work according to a few basic mechanisms and may be supervised or unsupervised. The most basic approach, extreme value analysis, is used for one-dimensional data and finds outliers through deviations from the mean.

Statistical models assume a particular data distribution and compute the outliers as points that have the lowest fit value to that distribution. Barnett and Lewis [1] discuss tests for various types of distributions and provide a comprehensive overview.

Linear models project the data into a lower-dimensional sub-space, then estimate the distance of each point to the subspace. Examples of this type of approach are PCA-based methods.

Proximity-based models are among the most popular outlier detection approaches. In these models, points that are separated from the rest of the data by a certain distance are considered outliers. These methods have the following subclasses:

- Cluster-based methods such as k-means (MacQueen [20]), where an object is an outlier if it is associated with no clusters, with a small cluster, or there is a large distance between the point and the nearest cluster
- Density-based methods such as LOF (Breuning et al. [2]) and DBSCAN (Ester et al., [7]), where an object is an outlier if its density is relatively much lower than that of its neighbors
- Distance-based methods, such as K-NN (Cover and Hart [5]), where an object is an outlier if its neighborhood does not have enough other points
- Depth-based methods (e.g., ISODEPTH) where data objects are organized in layers, where shallow layers are more likely to contain outliers (Ruts and Rousseuw [24])

One-class classification approaches train traditional binary or multi-class classifiers on a single class and the outliers are data points that will fail to be scored for that class. Such methods include one-class SVMs (OCSVMs) and One Class Random Forests (OCRFs).

Finally, special-purpose methods used for high-dimensional datasets, such as angle-based outlier detection (Kriegel et al. [17]), are proposed to escape the "curse of dimensionality" inherent in

statistical and distance-based methods. Such methods analyze the variance of the angles between data points rather than the distance.

We focus here on recent methods that have been shown to perform well either on high-dimensional data in general or on textual data in particular.

Our analysis is the first to analyze textual outliers within a single data set rather than using intrusion tests from different texts. The Risk Factors data set has certain unique properties that make it suitable for this type of analysis and also challenging due to its multi-topical nature. We show, despite these challenges that such an analysis can be successful in identifying "unusual" Risk Factors.

Our work is divided into two basic experiments. In the first experiment, we review recent high-performing models from the literature and test these models on three publicly available datasets. Following the work of Kannan et al. [15], we consider textual outliers to be texts from a given topic that have been inserted into a much larger set of texts from a separate topic. This choice is motivated by the goal of the second experiment, which is to identify the Risk Factor sections of Company Annual Reports that use "unusual" language, in particular discussing topics counter to the expectations of the reader. The first experiment is also motivated by a desire to gain a deeper understanding of the performance variance across different types of text, especially for models that have not previously been used on textual data. To achieve this goal, we run the models that performed best in our first experiment on the three public datasets to identify anomalous Risk Factor sections.

2. PREVIOUS WORK

Although outlier detection in text has not received as much attention as other contexts, there are some notable exceptions. We discuss other relevant work here, including recent approaches that have performed well on high-dimensional datasets.

2.1 Author and Genre

Guthrie [12] and Guthrie et al. [11] consider texts that are unusual because of author, genre, style or emotional tone, using hand-coded stylistic features to identify textual outliers. Guthrie [12] states "…abnormalities in text can be viewed as a type of outlier detection because these anomalies will differ significantly from the writing style in the majority of the data." The goal of identifying outliers along these parameters includes automated plagiarism detection, detection of deceptive writing and improvement in the homogeneity of large corpora.

Guthrie [12] calculates, for each segment of a document, the distance to its complement in the text (the union of the remaining segments) where the vector space is a list of stylistic features rather than words. The inlier dataset used was a collection of fictional works, varying the word-length of the intrusion sets between 100 and 1,000 words.

2.2 Metaphorical Language

Feldman and Peng [8], in an analysis of figurative language, view idiom recognition as a type of outlier detection. Idioms are figurative expressions where the underlying meaning cannot be derived from the literal meaning of the phrase. Typical examples in English include "kick the bucket," "have a cow," and "let the cat out of the bag." Feldman and Peng [8] observe that idioms have three main properties that make detection more likely using methods for finding outliers. First, literal phrases tend to have many neighbors, whereas figurative phrases have only a few. Second, words making up idiomatic phrases have low semantic relatedness. Third, idiomatic phrases are not strongly related to the preceding and following text. Based on these three observations, they reduce the problem of idiom recognition to that of detecting semantic outliers.

2.3 Topic

Using the semantics of document-level contexts, Kannan et al. [15] investigated textual outliers as purely topical deviants (in single-topic texts) within various corpora including news and Wikipedia. They note that non-negative matrix factorization (NMF), as a low-rank approximation technique, is particularly well suited to high-dimensional datasets and thus a natural candidate for text-based outlier detection. They generate two separate matrices: a low-rank matrix, created by a generative process and a separate matrix not generated by this process representing the outlier entries. Applying L1 and L2 norms to columns in the second matrix, Kannan et al. [15] minimize the sum of the outlier scores, using block coordinate descent. This approach, Text Outliers using Nonnegative Matrix Factorization (TONMF), is then applied to standard text-based datasets in various ratios of outliers to inliers.

Kannan et al. [15] note that the performance of TONMF and other competing methods used in their study varies considerably with the corpus and choice of inlier-outlier pair.

2.4 Other Approaches

Several recent approaches show success on a variety of high-dimensional datasets. One-class classification has gained some recent attention in the literature for outlier detection. One-class Support Vector Machines (Manevitz and Yousef [21]) have proven successful on the Reuters test set with various sparse feature representations. Similarly, One Class Random Forests (Goix [10], Désir et al. [6]) creates an approach to decision-tree splitting based on having only one class present in the data. The OCRF method selects an artificial outlier distribution at each node to be split and adapts based on the inlier probability at that node. The method outperforms LOF, iForest and OCSVM on 12 benchmark datasets from the UCI machine learning data repository.

Angle-based outlier detection (Kriegel et al. [17]) addresses the problem of high-dimensionality by comparing angles between pairs of distance vectors rather than using distance directly. If the range of observed angles for a point is large, the point will be surrounded by other points in all directions indicating that the point is positioned inside a cluster. This situation is characteristic of inliers, whereas angles with less variation tend to characterize outliers. This method outperformed Local Outlier Factor on a synthetic dataset.

3. DATA AND METHODS

In this section we discuss both the datasets used for the initial experiment, in which we select a model to be used for finding outliers in financial data, and the Risk Factors dataset used for the second experiment. We discuss the process through which we arrived at the choice of an outlier detection model.

Our approach follows a topical definition of outlier similar to Kannan et al. [15], but focuses on a specific set of documents in the finance industry. The document set comprises particular sections of corporate annual reports known as Risk Factors, which tend to be fairly formulaic in their construction. As such, when these sections do not follow prescribed patterns, they tend to signal unusual and

important corporate events. The character of the unusual sections has no discernable pattern in terms of stylistic or syntactic differences, but rather tends to raise a topic that is not normally discussed in that location. In particular, the purpose of the "Risk Factor" sections is to convey the types of risks that investors in the company would incur at the time the report is issued under normal circumstances. When risks fall outside of what would be considered normal circumstances, these sections will reflect that with unusual language. In what follows, we will describe our approach to identifying sections of unusual language in these documents using the best performing outlier model from a set of recent approaches

3.1 Data for the Initial Experiment

In the initial experiment, we compare selected recent approaches that have either performed well on textual data or have performed well on other data types, but are well-suited to high-dimensional data. Because we consider outliers to be topical text variants, we follow the approach of Kannan et al. [15] wherein we selected outliers and inliers from different texts in 20Newsgroups, Reuters and Wikipeople respectively. We follow the authors' strategy for the inlier-outlier selection for all three corpora. All models were run on these datasets. Stemming was not applied except on Reuters, which is stemmed by default.

20Newsgroups: The 20Newsgroups dataset[1] is a publicly available collection of approximately 20,000 newsgroup documents organized into 20 topical subgroups. Some of the newsgroups are very closely related to each other (e.g., IBM / Mac hardware), while others are highly unrelated (e.g., for sale / Christian religion). For this dataset, we cleaned the data before using it with any of the selected models, removing non-ASCII characters. We used "comp.sys.ibm.pc.hardware" and "comp.sys.mac.hardware" for the inliers and 50 samples from "soc.religion.christian" as outliers. The samples were selected randomly.

Reuters-21578 Text Categorization Collection Dataset: The Reuters dataset[2] is a publicly available dataset of stories that appeared on Reuters' newswire in 1987. It contains 21,578 documents that have been indexed and assigned categories by members of the Reuters Ltd. staff. Here we used the "acq" and "earn" text as inliers and 100 samples from the "interest" texts as outliers. The samples were selected randomly.

WikiPeople: The WikiPeople dataset is a publicly available dataset[3] constructed from Wikipedia articles about people. The dataset consists of 1,285 documents with 18,833 words and contains only sections where headings appeared more than ten times in the entire corpus. For this dataset, we used all "life" sections (9,246 samples) for inliers and 100 samples from the "death" section for outliers. The samples were selected randomly.

3.2 Risk Factor Data for U.S. Companies

The U.S. Securities and Exchange Commission (SEC) requires the Form 10-K, or annual report, to be filed by every publicly-traded U.S. company at the end of its fiscal year. It includes a summary of the company's results for that year, risk factors, management discussions, analysis of financial conditions, and details of all the annual financial statements. The Risk Factors section, mandated by the SEC, is found in Part 1 of the annual report and may have sub-sections as needed to describe the different types of risks faced by potential investors. These sub-sections may vary by industry, although the topics relating to high risks tend to be fairly industry invariant. Per Item 503(c) of Regulation S-K – a regulation the SEC began vigorously enforcing in the wake of the 2009 financial crisis – a company "[should] not present risks that could apply to any issuer or any offering." Figure 1 shows some examples of Risk factors, including typical samples and outliers as determined by analysts.

The SEC also states that the Risk Factor data "includes information about the most significant risks that apply to the company or to its securities. Companies generally list the risk factors in order of their importance. In practice, this section focuses on the risks themselves, not how the company addresses those risks." [4]

DATA	STATUS
Our Manager has limited experience operating a public company or complying with regulatory requirements, including the Sarbanes-Oxley Act, which may hinder its ability to achieve our objectives.	Outlier
Our success depends on our ability to attract and retain key personnel. Our success depends to a large extent upon our ability to attract and retain key executives, managers and skilled personnel.	Inlier
There Is No Assurance That Our Products Will Have Market Acceptance	Inlier
We rely on third-party license agreements; impairment of those agreements may cause pr oduction or shipment delays that could harm our business.	Inlier
We have licensing agreements with other entities for patents, software and technology used in our manufacturing operations and products.	Inlier
We may require additional capital to finance our growth or to fund acquisitions or investments in complementary businesses, technologies or product lines.	Inlier
The opportunity to earn incentive compensation may lead our Manager to place undue emphasis on the maximization of dividends at the expense of other criteria, such as preservation of capital	Outlier

Figure 1: Risk Factors

The Risk Factors text tends to vary very little from year-to-year, unless an important event has taken place, such as an upper management departure, accounting issue or significant change in the company's financial outlook. These sorts of events would be expected to pair with unusual language found in the text. Thus, we define an "outlier" for this type of data to be a risk factor that describes an unusual amount of risk or a particular event that could have imminent negative consequences on the financial health of the reporting company. While typical texts are characterized by conditionals and speculative statements, texts signaling unusual risk or serious problems tend to be characterized by very specific language.

The existing finance literature has used a variety of techniques to analyze the informational content of risk factors. Campbell et al. [3] use the total word count and hand-selected keywords from risk topics to show significant associations between risk factor content and measures of market risk. Kravet and Muslu [16] use the number of sentences to show a similar effect on market risk, as well as on the risk perception of investors. Nelson and Pritchard [23] use trigram overlap to show that high-risk firms disclosed more material compared to low-risk firms in the voluntary disclosure regime (pre-2005). Hope et al. [14] leverage a NER system to show that more specific risk factors elicit a stronger market/analyst

[1] https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

[2] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[3] http://www.cs.gmu.edu/~sblasiak/wikipeople.tar.gz

[4] https://www.sec.gov/fast-answers/answersreada10khtm.html

response. Also, Gaulin [9] uses LDA-based topic detection to model firm-specific topics within risk factors.

However, no current analysis exists on simply identifying anomalous language in Risk Factor data, nor are we aware of any treatment of unusual Risk Factors as textual outliers.

For this experiment, we selected 150,000 risk factor sections which are tagged by an automated parser[5]. This set included 11 Industry classes and ranged from the years 2016 to 2017.

4. RESULTS AND DISCUSSION

In this section, we discuss the findings of the first experiment and review the qualities of the models that were chosen as "winners" for the purposes of being used in the second experiment on financial text.

4.1 Description of Candidate Models

Our candidate models are derived from seven base models.

KNN: The K-nearest-neighbors model is the prototypical distance-based method which might be expected to perform well on text-based distances and was used on metaphorical outlier detection by Feldman and Peng [8].

Local Outlier Factor (LOF): LOF is a density-based method (Breunig et al. [2]) used in a variety of outlier detection experiments on high-dimensional data (Goix et al. [10] inter alia).

Angle-based Outlier Detection (ABOD): The ABOD model (Kriegel at al. [17]) calculates the variance in angles between the vectors of the samples, overcoming the dimensionality problems of pure distance-based approaches like LOF.

One-Class Support Vector Machines (OCSVM): In this approach (Scholkopf et al. [25]) the model maps input data into a high-dimensional feature space (via a kernel), then iteratively finds the maximal margin hyperplane that separates the training data from the origin. The origin is treated as the only member of the second class. This was previously used on the Reuters corpus by Manevitz and Yousef [21].

Text Outliers using Nonnegative Matrix Factorization (TONMF): This model (Kannan et al. [15]) uses a variation of low-rank matrix factorization, as is discussed in detail in section 2.3. It has shown good results on standard text data.

Isolation Forest (iForest): This method (Liu et al. [19]) works by identifying the trees with smaller average decision path lengths that are characteristic of outlier samples.

Because the approach is not distance or density-based, it is a strong candidate for analyzing high-dimensional datasets.

One-Class Random Forests (OCRF): This model (Désir et al. [6]) is based on a random forest algorithm and an outlier generation process that uses classifier ensemble randomization principles. The method has shown strong performance on high-dimensional datasets.

4.2 Results from Benchmark Datasets

We ran 10 models in total in a supervised experiment over the three benchmark datasets. We use two metrics to report quality. First, we show area under the ROC curve (AUC) drawn from the threshold range over the outlier scores, which is standard for outlier analysis.

[5] Access to this proprietary system provided by Bloomberg LP

Second, we measure Matthews Correlation Coefficient (MCC) based on Matthews (1975), as this metric is known for stability over differing class sizes. MCC is defined as follows:

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

We note that the models differed significantly, both in their performance on text in general and on each dataset individually. In some cases, this was unexpected given previous results in the literature. For example, despite the results in Liu et al. [19] showing iForest strongly outperforming LOF on high dimensional datasets, as well as large datasets, it did not do so on any of the datasets in our benchmark sample. On the other hand, models that had performed well on text in particular also did well in our samples.

	20NewsGs	Reuters	WikiP	**Avg**
k-NN	0.775	0.527	0.524	0.609
OCSVM + Count + NMF	0.671	0.728	0.607	0.669
TONMF + Count	0.709	0.762	0.416	0.634
TONMF	0.552	0.692	0.618	0.621
LOF	0.577	0.509	**0.694**	0.593
LOF + PCA	0.681	**0.778**	0.612	**0.690**
ABOD + PCA	**0.844**	0.333	N/A	0.556
iForest	0.355	0.610	0.301	0.422
Best Model	ABOD + PCA	LOF + PCA	LOF	LOF + PCA

Table 1: AUC Results

	20NewsGs	Reuters	WikiP	**Avg**
k-NN	**0.549**	0.054	0.050	**0.368**
OCSVM + Count + NMF	0.136	0.131	-0.022	0.082
TONMF + Count	0.159	0.044	0.009	0.212
TONMF	-0.005	0.054	0.011	0.060
LOF	0.101	0.198	**0.070**	0.123
LOF + PCA	0.226	**0.281**	0.056	0.188
ABOD + PCA	0.214	0.003	N/A	0.089
iForest	0.117	0.084	-0.011	0.063
Best Model	k-NN	LOF + PCA	LOF	k-NN

Table 2: MCC Results

We found that certain models, such as Local Outlier Factor, Angle Based Outlier Detection and TONMF were particularly sensitive to certain representations and we updated these models accordingly using the best-performing representation from Binary, Count-based, TFIDF or PCA. TONMF responded to simpler sparse representations, whereas other methods responded to a lower dimensional representation (PCA-based). Where it is not specified, we used a tf-idf representation. Table 1 below shows AUC results and Table 2 shows MCC results. We did not process the texts to remove stop words and did not perform stemming. Best scores are highlighted.

One-Class Random Forest examples are omitted from the table of outcomes, as the results were below our threshold cutoff of .4 for averaged AUC and .05 for averaged MCC. ABOD is not included in the table of outcomes as the performance of ABOD on large sparse representations did not meet our minimum threshold necessary to run it on the Risk Factor corpus. ABOD + PCA results for WikiPeople are not available, as the program did not complete.

In general the results on WikiPeople were below those for the other datasets for most of the models tested. While we don't see an obvious cause, we note that the "Death" sections used as outliers tend to be very topically diverse and that this dataset tends to have a high ratio of common-to-proper nouns. This could contribute to very low term-densities, increasing dimensionality.

4.3 Results on Risk Factors Data

We extracted the Risk Factors sections from 150,000 10-K annual reports for the filing years 2016-2017.

As an unsupervised experiment, we ran the three algorithms that performed best on the public datasets based on mean averaged rank of AUC and MCC scores over the three datasets. These were, in order of rank, LOF+PCA, TONMF and KNN respectively.

For the Risk Factor data, we did not preprocess the text to exclude stop words nor did we use a stemmer. The feature set used in this experiment consisted simply of unigram-based bag-of-words.

We recorded the top 50 outlier scores for each of the three models and combined them with 50 inliers selected randomly. This data was then randomized and sampled without replacement by unique ID, so that the human reviewer would not see the same sample twice. The randomized set without the scores was sent to a domain expert for review and contained 285 samples. The original 300 samples were reduced due to certain samples being erroneously parsed from a previous section in the document and not containing a complete Risk Factor section. The results of the review are in Figure 2. True positives are samples that the model scored above the threshold that were selected by the reviewer. False positives are samples that scored above the threshold that the reviewer called an inlier. False negatives are samples that scored below the threshold that the reviewer picked as an outlier.

The TONMF model was the best performing with an F1 score of .62 and an MCC score of .47. For LOF_PCA and KNN respectively F1 was .21 and .39. MCC for LOF_PCA was -0.24 and -0.10 for KNN. While TONMF does have higher recall than the other models, it is notable for its high precision.

Fisher's Exact Test was run on the confusion matrix for each model. The p-values were 0.0002, 0.012 and 0.096 for TONMF, LOF+PCA and KNN respectively.

The average Krippendorff's alpha (Krippendorff [18]) across models was -0.01, indicating that the models in general do not make human-like selections despite at least one model having a reasonable portion of true positives. We suspect that additional features, either lexical features identifying the complex semantics of different risks, or non-lexical features reflecting the company's financial health may improve this behavior.

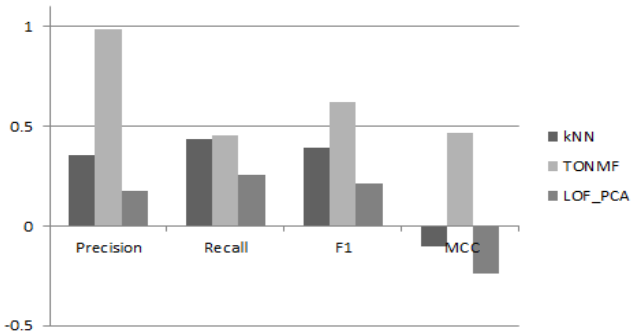

Figure 2: Results on Risk Factors

5. CONCLUSIONS AND FUTURE WORK

In this paper we have described two experiments. First, we discussed the performance of eight outlier detection methods[6] on three benchmark textual datasets. For each model we selected different feature representations and reported numbers for the ones that did best on each. Second, we showed the results of running the three top-performing candidates from the initial experiment on a hand-scored sample of outliers and inliers from a dataset of Risk Factors found in SEC 10-K documents. Results indicate strong potential for using this technique to automatically identify unusual topics in this dataset.

We note that Risk Factor outliers tend to be very topically diverse, which makes them a challenging application for outlier detection. However, the topics related to unusual risk are finite, being restricted to those risks impacting investment. Therefore, despite the difficulty, we believe this data is a good candidate for this approach and for further research into topical outlier detection in text.

Of the three candidate models run on the Risk Factor dataset, our results show that the TONMF model performs best. This model has performed well on high-dimensional data in general and seems particularly suited to textual outlier sets where the outliers themselves are topically diverse, as is the case with Risk Factors.

Future work includes extending the present analysis to other models, such as autoencoders. These models are similar to TONMF in that they use a low-rank approximation, and have shown promising results for outlier detection in other areas (Zhou and Paffenroth [26]).

[6] Two of the ten did not complete or were below our performance threshold so these numbers did not appear in the table

Finally, we have noticed that outliers are characterized by higher densities of particular syntactic classes, such as pronominals and modals. An experiment with the benchmark datasets showed a small improvement on POS-tagged data on the Newsgroups Corpus for the TONMF model. However, we would expect this improvement to be enhanced in the Risk Factors due to the importance of pragmatic and semantic factors, including the frequent use of conditionals and specificity that tend to distinguish the typical risk discussions from unusual risk language. We would use a POS-tagged dataset in future iterations.

6. ACKNOWLEDGMENTS

We would like to acknowledge Alex Kreonidis on the Bloomberg Law Data team for providing subject matter expertise in scoring our samples

7. REFERENCES

[1] Vic Barnett and Toby Lewis. 1994. Outliers in Statistical Data. John Wiley, 1994.

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander. 2001. LOF: Identifying Density-based Local Outliers. In Proceedings of SIGMOD Conference, pages 93–104, 2000

[3] John L. Campbell, Hsiu-chin Chen, Dan S. Dhaliwal, Hsin-Min Lu and Logan B. Steele. 2013. The Information Content of Mandatory Risk Factor Disclosures in Corporate Filings. Review of Account-ing Studies 19(1)

[4] V. Chandola, A. Banerjee and V. Kumar. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41(3):15:1–15:58.

[5] T. Cover and P. Hart. 1967. Nearest Neighbor Pattern Classification. 2006. IEEE Transactions on Information Theory. Volume 13(1): 21-27

[6] C. Désir, S. Bernard, C. Petitjean, and L. Heutte. 2013. One class random forests. Pattern Recognition.

[7] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xu, Xiaowei. 1996. A density-based algorithm for dis-covering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.

[8] Anna Feldman and Jing Peng. 2013. Automatic detection of idiomatic clauses. In Computational Linguistics and Intelligent Text Processing, pp. 435–446. Springer.

[9] Maclean Gaulin. 2017. Risk Fact or Fiction: The Information Content of Risk Factor Disclosures. Ph.D. Dissertation, Jones Graduate School of Business Rice University Houston, Texas

[10] Nicolas Goix, Nicolas Drougard, Romain Brault and Mael Chiapin. 2016. One Class Splitting Criteria for Random Forests, In Proceedings of Machine Learning Research 77:343–358, 2017

[11] David Guthrie, Louise Guthrie, Ben Allison, Yorick Wilks. 2007. Unsupervised Anomaly Detection. In Proceedings of IJCAI-07

[12] David Guthrie. 2008. Unsupervised Detection of Anomalous Text. PhD Dissertation, University of Sheffield

[13] Douglas M. Hawkins. 1980. Identification of Outliers. Chapman and Hall, New York, London 1980.

[14] Ole-Kristian Hope, Danqui Hu and Hai Lu. 2016. The Benefits of Specific Risk-Factor Disclosures. Review of Accounting Studies. Research Paper no. 2016-49

[15] Ramakrishnan Kannan, Hyenkyun Woo, Charu C. Aggarwal and Haesun Park. 2017. Outlier Detection for Text Data. In Proceedings of the 2017 SIAM International Conference on Data Mining

[16] Todd. D. Kravet and Volkan Muslu. 2011. Textual Disclosures and Investors' Risk Perceptions. Review of Accounting Studies 18, pp. 1088-1122

[17] Hans-Peter Kriegel, Matthias Schubert and Arthur Zimek. 2008. Angle-based outlier detection in high-dimensional data. In Proceedings KDD'08, pages 444–452.

[18] K. Krippendorff. 2004. Content Analysis: An Introduction to Its Methodology, 3rd ed. Thousand Oaks, CA, USA: Sage

[19] Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou. 2008. Isolation Forest. In Proceedings of ICDM 2008

[20] J. B. MacQueen. 1967. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.

[21] Larry M. Manevitz and Malik Yousef. 2001. One-Class SVMs for Document Classification. Journal of Machine Learning Research 2 (2001) 139-154

[22] Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 405 (2): 442–451

[23] Karen K. Nelson and A. C. Pritchard. 2016. Carrot or Stick? The Shift from Voluntary to Mandatory Disclosure of Risk Factors. Journal of Empirical Legal Studies, 13(2), pp.222-297

[24] Ruts, I and P. Rousseuw. 1996. Computing depth contours of bivariate point clouds. Computational Statistics and Data Analysis, 23:153-168, 1996.

[25] Bernhard Scholkopf, John C. Platt, John Shaw-Taylor, Alex J. Smola and Robert C. Williamson. 1999. Estimating the Support of a High Dimensional Distribution. Microsoft Research Technical Report MSR-TR-99-87

[26] C. Zhou, R. C. Paffenroth. 2017. Anomaly Detection with Robust Deep Autoencoders. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). ACM, New York, NY, USA, 665-674.