Risk Analysis using Topic Models on Annual Reports
EDIC Semester Project I

Diego Antognini and Boi Faltings
Artificial Intelligence Laboratory
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland, 1015
Email: {firstName.lastName}@epfl.ch

***Abstract*—Annual reports are a way for companies to communicate their well-being to those whom it concerns like shareholders or stakeholders. In most reports, one or more sections are dedicated to the risks the enterprise might encounter in the future, which represents key information for people or other corporations relying on its performances. In this project, we first use different models on annual reports to learn topics which might embed risks and then express companies' risks as a mixture of topics. We then use clustering techniques in order to find companies sharing similar risks in the same sector or in a different area. Afterwards, we study how these topics might evolve through the time. Finally, we observe than Risk Factor section neither does not change significantly through the years nor are essentially shared among companies.**

1. Introduction

Publicly-traded companies have to communicate to the shareholders and stakeholders how their business is performing. To that purpose, enterprises typically write annual reports and share them to those whom it concerns or sometime make them publicly available. The latter contain plenty of information such as a description of the business, achievement highlights, comprehensive summary of the company's financial performances, properties, legal proceedings as well as risk factors which might have an impact on the company.

By reading these reports, shareholders and stakeholders can have the rough outline of a company's well-being and thus predict how it will be performing the next year as well as the financial risk of investment. This information allow them to gain knowledge about companies and have more accurate intuitions in order to make better decisions.

The most interesting sections in terms of significant risks[1] are these related to "Managements Discussion and Analysis of Financial Condition and Results of Operations", "Risk Factors" and "Quantitative and Qualitative Disclosures about Market Risk". They represent good indicators of how the company will perform during the next year. However, annual reports are getting longer as well as more redundant and are written in a way that make them complex to process (Dyer et al., 2017; Rekabsaz et al., 2017).

Our research focuses primarily on expressing risks associated to a company in a more concise and succinct way using topic models. The intuition behind this is that corporations working in the same sector or same industry share same products and thus, should also share similar risks. It emphasizes the use of topic modeling in order to express risks with a mixture of topics. Secondly, we investigate if topics are shared among different companies, as much as these operating in the same sector than these in a completely different area. Finally, we investigate if these topics are evolving through the years.

To this end, we try different topic models with different characteristics but the original Latent Dirichlet Allocation (LDA) (Blei et al., 2003) seems to work the best. Then we perform experiments to show that: (i) risks might be represented by topics (ii) these are not specially shared among companies and (iii) they do no vary significantly through the years due to report's Risk Factor section being mostly similar through the years for a same corporation.

The rest of this paper is organized as follows. Section 2 presents the dataset on which we perform our experiments. These are described in details in the Section 3. Finally, our work is concluded in Section 4.

2. Dataset and Preprocessing

We use 10-K annual reports of companies in the United States. These reports are mandatory for all publicly-traded corporations and are mandated by the U.S. Securities Exchange Commission (SEC)[2](Kogan et al., 2009). These annual reports contain usually plenty of information such as a description of the business, achievement highlights, comprehensive summary of the company's financial performances, properties, legal proceedings as well as risk factors which might have an impact on the company. We collect $175,092$ reports published over the period from January 1993 to June 2017 from $34,463$ different companies. This represents the biggest dataset of 10-K reports used so far up to our knowledge. Most works such as (Dyer et al., 2017;

1. An example on what sections Warren Buffett pays attention in annual reports: https://www.cnbc.com/2014/01/27/how-to-read-a-10-k-like-warren-buffet.html

2. http://sec.gov

TABLE 1: Some statistics of each dataset used after the preprocessing steps. The left side (Table 1a) is about section 1A and right side (Table 1b) about section 7A.

(a) Corpus containing the section 1A: "Risk Factors".

Year	# Words	# Documents	Words/Doc
1993	0.0	0	0
1994	10.7K	10	1,067
1995	24.3K	18	1,350
1996	215.4K	100	2,154
1997	751.4K	275	2,732
1998	1,088.6K	369	2,950
1999	1,396.8K	479	2,916
2000	1,612.5K	556	2,900
2001	1,950.4K	656	2,973
2002	1,985.5K	695	2,856
2003	2,392.5K	792	3,020
2004	2,276.9K	723	3,149
2005	2,262.0K	661	3,422
2006	8,442.6K	2,691	3,137
2007	8,510.9K	2,411	3,530
2008	12,013.3K	3,080	3,900
2009	13,695.9K	3,440	3,981
2010	14,830.6K	3,460	4,286
2011	20,442.3K	4,029	5,073
2012	19,032.8K	3,600	5,286
2013	16,798.1K	3,085	5,445
2014	15,061.1K	2,650	5,683
2015	13,888.8K	2,391	5,808
2016	12,474.7K	1,982	6,294
2017	9,768.9K	1,457	6,704
Total	180,927.2K	39,610	3,776

(b) Corpus containing the section 7A: "Quantitative and Qualitative Disclosures about Market Risk".

Year	# Words	# Documents	Words/Doc
1993	0.0	0	0
1994	0.0	0	0
1995	0.0	0	0
1996	0.0	0	0
1997	8.5K	25	339
1998	144.0K	527	273
1999	669.0K	2,507	266
2000	786.5K	2,965	265
2001	862.7K	3,199	269
2002	956.6K	3,272	292
2003	1,221.3K	3,936	310
2004	1,190.7K	3,593	331
2005	1,063.7K	3,034	350
2006	913.2K	2,478	368
2007	858.6K	2,315	370
2008	1,151.0K	3,069	375
2009	1,127.6K	2,918	386
2010	1,100.8K	2,737	402
2011	1,326.1K	2,863	463
2012	1,167.6K	2,310	505
2013	965.8K	1,956	493
2014	763.7K	1,603	476
2015	685.3K	1,442	475
2016	605.6K	1,282	472
2017	437.3K	934	468
Total	18,005.4K	48,965	378

Liu et al., 2016; Tsai and Wang, 2014, 2017, 2013) rely on the 10-K corpus of (Kogan et al., 2009) containing reports from 1996 to 2006 or more recently (Rekabsaz et al., 2017) who use reports over the period from 2006 to 2015.

For the perspective of risk analysis, two sections of 10-K reports are of special interests: section 1A known as "Risk Factors" and section 7 known as "Managements Discussion and Analysis of Financial Condition and Results of Operations" (MD&A), especially the subsection 7A "Quantitative and Qualitative Disclosures about Market Risk". Because section 1A describes all kind of possible risks the company might encounter in the future and section 7A is where the most important forward-looking content is very likely to be found according to (Kogan et al., 2009), we decide to filter out all other sections.

We write from scratch a hand-written Python script collecting reports from the EDGAR website[3] and gathering the following information for each of these: the fiscal year, the report release date and the section 1A as well as section 7a if they exist. These reports are plain text and not very well structured: companies have some freedom in which order they put the information and how they present it even though there are general instructions[4]. We create heuristic methods to extract relevant contents. These include finding relevant sections (1A or 7A) having a minimum as well as maximum number of tokens, knowing where to stop extracting their text to avoid adding content from the following section, etc. Moreover, some sections might be included by reference which we decide to ignore due to the complexity of finding the related external documents.

After having collected the documents, we tokenize the text and use as preprocessing: removing HTML tags, remnant markups, punctuations and multiple dashes, downcasing, replacing numbers by a special token and multiple spaces as well as new lines by a single space. Table 1 shows some statistics about the datasets. In total, there are $39,610$ ($\sim 22.62\%$) reports with a valid section 1A and respectively $48,965$ ($\sim 27.97\%$) for section 7A. If we have a closer look on the average number of words per document for section 1A in Table 1a, we observe that it is roughly increasing over the years. (Dyer et al., 2017) show that 10-K reports over the period $1996 - 2013$ tend to become longer, more redundant and more complex to understand. One of the main factor is related to the increasing length of the risk factor disclosures due to new SEC requirements. We confirm this observation and also note that unlike the section 1A, the section 7A is unaffected by this trend.

3. Topic Modeling

We investigate different topics models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Hierarchical Dirichlet Process (HDP) (Teh et al., 2005), spherical HDP (sHDP) (Batmanghelich et al., 2016), ProbLDA (Srivastava and Sutton, 2017), Topic Keyword Model (TKM) (Schnei-

3. https://www.sec.gov/Archives/edgar/daily-index/
4. https://www.sec.gov/files/form10-k.pdf

der, 2017), Gaussian LDA (Das et al., 2015), Latent Feature Topic Model (LFTM) (Nguyen et al., 2015) and LDA2vec (Moody, 2016).

LDA is a generative probabilistic model considering documents as a mixture of topics (known a priori) and topics as a mixture of words. Each topic has a multinomial distribution sampled from a Dirichlet distribution (with parameter α). Same applies for the word distribution given the topic (with parameter β). HDP is a variant of LDA where the number of topics is not known a priori and outputs a hierarchy of topics where the user is free to choose what level corresponds the best to its data. ProbLDA replaces the mixture model for words by a product of experts which should drastically improve topic coherence according to the authors. TKM has a different approach and propose a topic model based on keywords only: each word has a keyword score stating how likely it belongs to a specific topic. Moreover, words with a high score have an influence on their neighborhood. Finally, they also remove redundant topics by measuring the difference in their word-topic distributions using the symmetrized KullbackLeibler divergence. The other methods rely on incorporating an additional prior using word embeddings (Mikolov et al., 2013) learnt on a big external corpus. In this manner, it allows to have more semantic coherence between topics as it uses word embeddings which capture syntactic and semantic regularities in language. Gaussian LDA is a variant of LDA where the multinomial distribution of words given a topic is replaced by a gaussian distribution of word embeddings. sHDP behaves similarly to Gaussian LDA beside using a von Mises-Fisher distribution to model the density of words over a unit sphere rather than a Gaussian distribution. Finally, LFTM and LDA2vec include word representations into the two Dirichlet distributions.

For topic models relying on word embeddings, we compute them using the Skip-Gram model of (Joulin et al., 2016), with 600 dimensions and default values for the other parameters, because the existing pre-trained models have been trained on general corpora (e.g. Wikipedia, Google News, etc.). Therefore, these embeddings incorporate a general syntactic and semantic meaning and thus, might not be representative of financial text as specified in (Bodnaruk et al., 2015; Loughran and McDonald, 2011). As corpus, we concatenate all sections 1A with 7A (with preprocessing as in Section 2) and end up with 200 millions of tokens.

Unsurprisingly, as our corpus is substantially large, most topic models cannot be trained within a reasonable amount of time due to their high complexity. The only models which can be trained in a descent time (i.e. less than a week) are LDA, HDP and TKM. In order to compare their performances in an automatic way, we do not use the well-known perplexity metric because it has been shown that it is not correlated with human judgement (Newman et al., 2010). Consequently we opt for the metric C_V (Röder et al., 2015) which combines the indirect cosine measure with the normalized pointwise mutual information (NPMI) and a sliding window. The authors show that C_V is the most correlated with human judgement.

3.1. Topic Visualization

We train the three topic models and obtain the best one, with a significant higher C_V than the others, with LDA. In total, the best empirical number of topics with respect to the C_V score is 66 topics for the section 1A and 19 for the section 7A. A visualization obtained with (Sievert and Shirley, 2014) is shown in Figure 1. For each section, topics are projected into a two-dimensional space using multidimensional scaling. The diameter of the circles represent the marginal topic distribution, the larger the higher. As we can observe, topics from the section 1A are more spread and more fine-grained compared to these of the section 7A. As a result, topics in the risk factors section encapsulate more information because most of the topics have a small marginal topic distribution and thus are not shared by a large number of documents. In Figure 1.1, the largest diameter corresponds approximately to a marginal topic distribution of 10% and only 4 topics over 66 have such a value. All the topics with their top words are shown in the appendix in Figure 5. By looking the generated topics, many areas are represented such as pharmacy, business, customers, finance, infrastructure, etc. We can observe that there are only a few overlaps among the topics as seen in Figure 1.1. From now, we only consider our topic model computed on the section 1A.

3.2. Clustering in the topic space

Once the topic distribution vectors of the documents are computed, we also investigate if we could find clusters of different companies given their Risk Factors section. As a first attempt, we plot the representation of these reports (over the whole period) using tSNE (van der Maaten and Hinton, 2008) and project them in a two-dimension space. The Figure 2.1 shows the result. Years are represented with distinct colors. As a first observation, we can observe that no clear clusters appear. This reveals that through the years, there are no clear distinctions among topic distributions. Moreover, we can see two tiny clusters on bottom left of the figure and one on the top. Section 1A in these are almost identical besides a couple of words. For example, the blue cluster contains reports of *NCO GROUP* companies being different with respect to EDGAR because their CIK numbers (universal identification number from SEC) are not the same due to their business addresses[5]. Because this attempt does not show promising results, we also try by using the sectors of the companies. We collect those automatically using Yahoo categorization of sectors. It contains the following categories: Financial, Basic Materials, Utilities, etc. Only $\sim 22.7\%$ of the reports have an attached

5. Here are four of these points: https://www.sec.gov/cgi-bin/browse-edgar?CIK=1058624&Find=Search&owner=exclude&action=getcompany, https://www.sec.gov/cgi-bin/browse-edgar?CIK=1431674&Find=Search&owner=exclude&action=getcompany, https://www.sec.gov/cgi-bin/browse-edgar?CIK=1029319&owner=exclude&action=getcompany and https://www.sec.gov/cgi-bin/browse-edgar?CIK=1431659&owner=exclude&action=getcompany

Figure 1: Topic visualization for sections 1A and 7A, produced with (Sievert and Shirley, 2014). Topics are projected into a two-dimensional space using multidimensional scaling. The diameter of the circles represent the marginal topic distribution, the larger the higher.

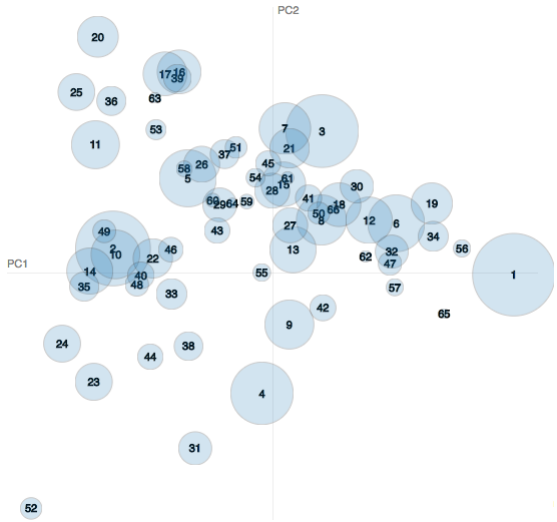

Figure 1.1: Topic visualization for section 1A.

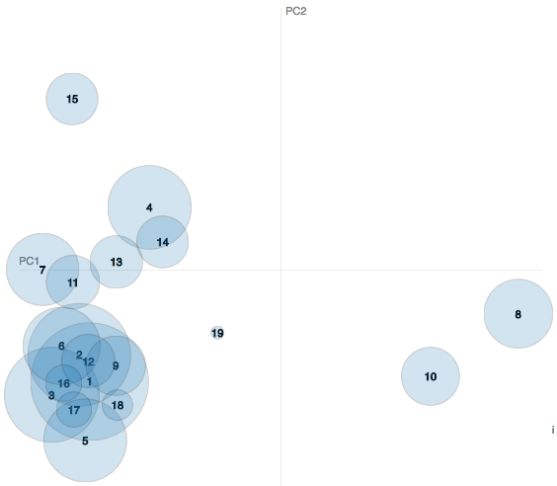

Figure 1.2: Topic visualization for section 7A.

sector. The result is shown in Figure 2.2 where each color is a different sector. Albeit there are few clusters popping (e.g. red one corresponds to the financial sector), most of the reports are scattered and thus, cannot isolate clearly similar reports.

3.3. Topic evolution through time

In light of these results, we experiment whether the topic distributions change over the time. To achieve this, we compute the mean topic vector per year from all reports in the same year. Figure 3.1 shows the similarity heat-map for each pair of the years' centroids. Surprisingly, except the years 1994 and 1995 which are far from the other centroids,

Figure 2: tSNE projection of the topic distribution from all reports having a Risk Factors section. On the left side, each color represents a year and on the right a sector. The latter plot is smaller because we can find the sector automatically only for a limited number of companies ($\sim 22.7\%$ of the reports).

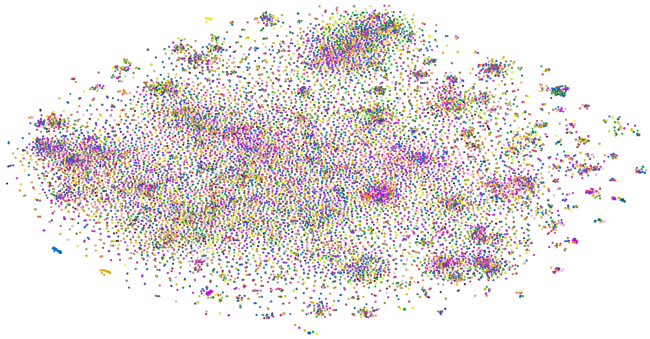

Figure 2.1: Each color represents a year.

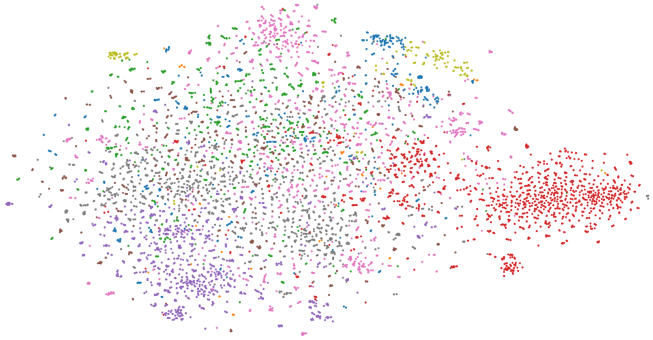

Figure 2.2: Each color represents a sector.

all the others are highly-correlated (above 0.95). This trends is confirmed by (Dyer et al., 2017) which shows that reports are getting longer, more redundant and boilerplate. (Rekabsaz et al., 2017) try a similar experiment but using volatility computed with the GARCH model (Bollerslev, 1986) and obtain a very diverse heat-map. This observation might support that topics are indeed too similar across reports no matter the year. We also investigate when reports' content change slightly and plot a dendrogram based on centroid vectors of the years in Figure 3.2. We observe that there exists cycles of 3-4 consecutive years where content of reports change slightly (at the scale of a percent).

Finally, we examine more closely how topics might change through the year. Unfortunately, due to a large corpus, we could not train a dynamic topic model such as Dynamic Topic Model (DTM) (Blei and Lafferty, 2006) in a reasonable time (i.e. less than a few weeks). Nevertheless, we tackle it by exploring three angles: the correlation between topics and years, topics and sectors using Yahoo categories and lastly, proportion of each topics in the topic centroids for all years. The Figure 4.1, 4.2 and 4.3 show respectively the results of these three cases. For the two firsts, we can remark that most topics with high correlation are shared with all years (respectively all sectors) and thus,

Figure 3: Figure 3.1: Cosine similarity heat-map between mean of topic vectors per year. Figure 3.2: Dendrogram of centroid vectors of the year with cosine distance. The two final clusters are linked together with a distance of 1.2 but this latter has been removed from the plot for clarity purpose.

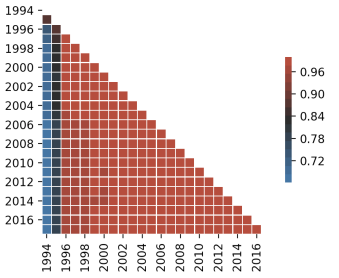

Figure 3.1

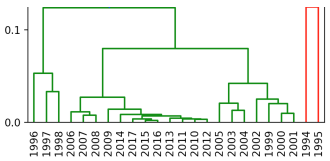

Figure 3.2

do not encapsulate much information. For the last, topics vary a lot in the first four years surely due to the number of reports available in the period $1994-1996$. Thereafter, topics are more or less stable up to 2017. We also do the same on individual companies and find similar trends. All these observations support the fact that topics might be not very representative of the evolution of risks through the years, probably because of the similar content in the Risk Factor section of annual reports.

4. Conclusion

In this project, we first show that risks from annual reports can be represented with topics. We start by trying different topic models wit different characteristics and obtain the best results with the basic LDA with 66 topics. These cover a wide range of areas from pharmacy to finance. Consequently, we are able to represent risks as a mixture of significant topics.

Secondly, we compare risk topics among different companies using our topic space and observe that there are nearly no clear clusters of enterprises sharing same risks. This concerns as much as corporations operating in the same sector than these in a completely different area.

Lastly, we study the evolution of topics through the years and note they do no vary significantly due to report's Risk Factor section being mostly similar through the time for a same corporation.

Our experiments confirm our hypothesis that Risk Factor section does not change significantly through the years for a given company. Moreover, we verify the general observations of (Dyer et al., 2017) who found that annual reports tend to become more longer, more redundant and boilerplate and finally do not bring much additional information year after year.

Figure 4: Figure 4.1: Cosine similarity heat-map between mean of topic and years. Figure 4.2: Same as before but with topics and sectors. Figure 4.3: Topic evolution through the years where each color corresponds to a topic and the y axis the proportion of each topic.

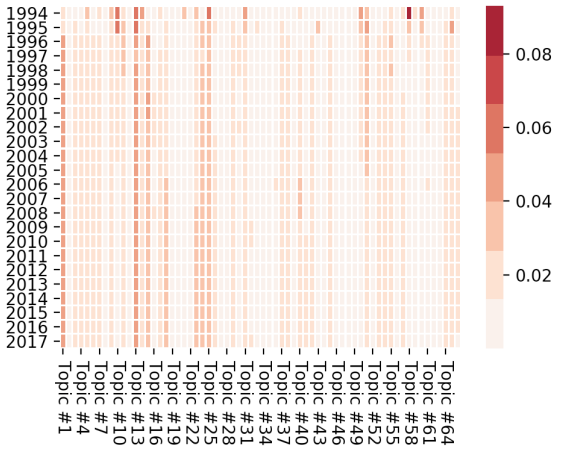

Figure 4.1

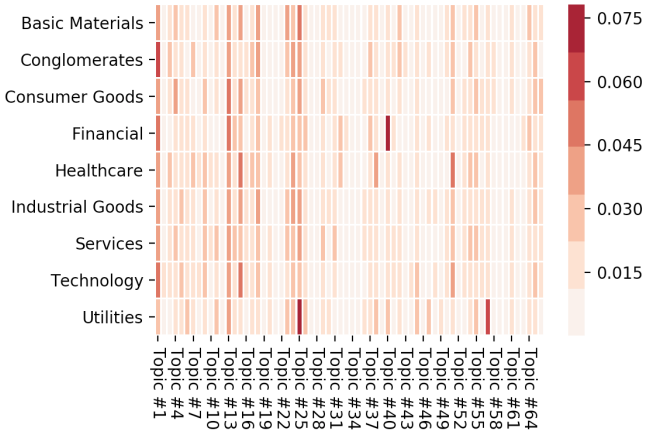

Figure 4.2

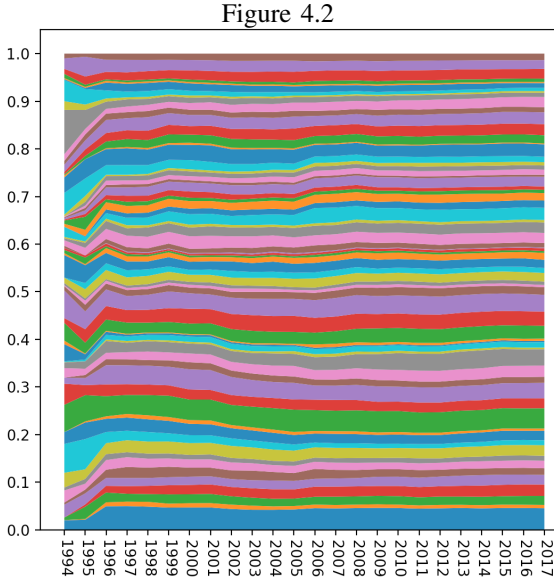

Figure 4.3

5. Acknowledgments

I would like to thank Prof. Boi Faltings for all his valuable help throughout this project.

References

Batmanghelich, K., Saeedi, A., Narasimhan, K., and Gershman, S. (2016). Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bodnaruk, A., Loughran, T., and McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4):623–646.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.

Das, R., Zaheer, M., and Dyer, C. (2015). Gaussian lda for topic models with word embeddings. In *ACL (1)*, pages 795–804. The Association for Computer Linguistics.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017). The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics*, 64(2):221 – 245.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics.

Liu, Y.-W., Liu, L.-C., Wang, C.-J., and Tsai, M.-F. (2016). Fin10k: A web-based information system for financial report analysis and visualization. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2441–2444, New York, NY, USA. ACM.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Mikolov, T., Yih, S. W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics.

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Rekabsaz, N., Lupu, M., Baklanov, A., Hanbury, A., Dür, A., and Anderson, L. (2017). Volatility prediction using financial disclosures sentiments with word embedding-based ir models. *arXiv preprint arXiv:1702.01978*.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Schneider, J. (2017). Topic modeling based on keywords and context. *arXiv preprint arXiv:1710.02650*.

Sievert, C. and Shirley, K. E. (2014). Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. *International Conference on Learning Representations*.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392.

Tsai, M. and Wang, C. (2014). Financial keyword expansion via continuous word vector representations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1453–1458.

Tsai, M. and Wang, C. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, 257(1):243–250.

Tsai, M.-F. and Wang, C.-J. (2013). Risk ranking from financial reports. In *European Conference on Information Retrieval*, pages 804–807. Springer.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Appendix

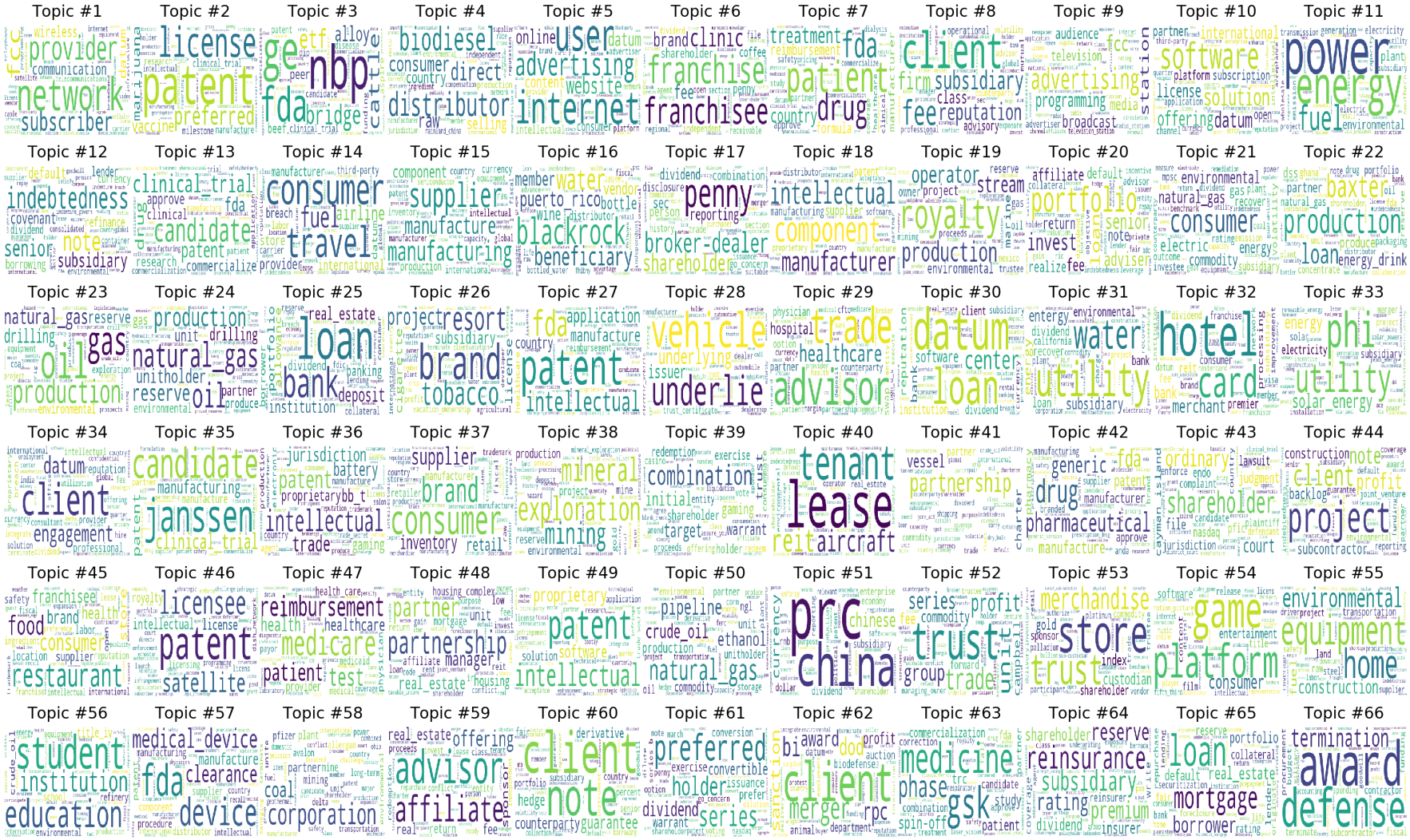

Figure 5: The 66 topics from the Risk Factors section obtained with LDA.