Characterization of Idea Relations in Text: Investigation with Topic Modelling and Structural Topic Modelling

Yijie Yao, Qitian Hu, Shicheng Liu
Team: Mission Inspiring God
The University of Chicago
{yjyao, jasonhu, shicheng2000}@uchicago.edu

Abstract

We present a study on characterizing relations between ideas generated from topic modelling, inspired by Tan et al. (2017). **TODO**

1 Introduction

1.1 Overview

1.2 Contribution

2 Computational Framework

2.1 Topic Modelling

2.2 Structural Topic Modelling

The raw topic model (raw TM) only adds topic and document as internal organization of the corpus, but in practice, we usually have a much larger range of metadata. Structural topic modeling (sTM (Roberts et al., 2013b)) allows us to add document-level metadata on which topical prevalence or topical content is of interest. The main idea is to use generalized linear models as priors and then condition on metadata of the document.

Based on three variants of LDA (Blei et al., 2003a), correlated topic model (CTM, (Blei et al., 2007)), the Dirichlet-Multinomial Regression topic model (DMR, (Mimno and McCallum, 2008)) and the Sparse Additive Generative (SAGE, (Taddy, 2013)) topic model.

The logistic normal prior for topics in the standard CTM is replaced by a logistic-normal linear model. The design matrix for the covariates X allows for flexible forms of the covariates that the prior could be conditioned on.

The distribution over words is replaced with a multinomial logistic function such that a token's distribution is jointly determined by topic, covariates, and topic-covariate interaction. We adopted the R implementation provided by the authors (Roberts et al., 2019).

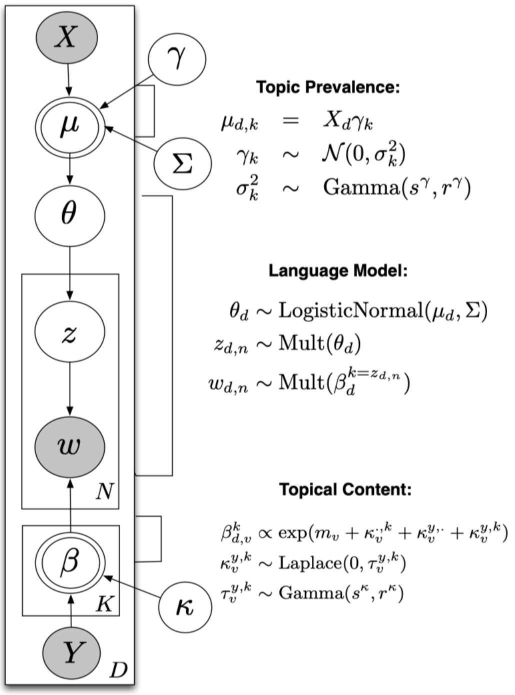
X
γ
μ
Σ
θ
z
w
N
β
K
κ
Y
D
Topic Prevalence:
$\mu_{d,k} = X_d\gamma_k$
$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$
$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$
Language Model:
$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$
$z_{d,n} \sim \text{Mult}(\theta_d)$
$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$
Topical Content:
$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$
$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$
$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$

Figure 1: Plate Diagram for the Structural Topic Model

A more specific description of the generative process is described here.

First, draw the document-level attention to each topic from a logistic-normal generalized linear model based on a vector of document covariates X_d.

$$\vec{\theta}_d \mid X_d\gamma, \Sigma \sim LogisticNormal\,(\mu = X_d\gamma, \Sigma)$$

where X_d is a 1-by- p vector, γ is a p -by- $K-1$ matrix of coefficients and Σ is $K-1$ -by$K-1$ covariance matrix. Second, given a document-level content covariate y_d, form the document-specific distribution over words representing each topic (k) using the baseline word distribution (m), the topic specific deviation $\kappa_k^{(\mathrm{t})}$, the covariate group deviation $\kappa_{y_d}^{(\mathrm{c})}$ and the interaction between the two

$\kappa^{(\mathrm{i})}_{y_d,k})$

$$\beta_{d,k} \propto \exp\left(m + \kappa^{(\mathrm{t})}_{k} + \kappa^{(\mathrm{c})}_{y_d} + \kappa^{(\mathrm{i})}_{y_d,k}\right)$$

m, and each $\kappa^{(\mathrm{t})}_{k}$, $\kappa^{(\mathrm{c})}_{y_d}$ and $\kappa^{(\mathrm{i})}_{y_d,k}$ are V -length vectors containing one entry per word in the vocabulary. When no convent covariate is present β can be formed as $\beta_{d,k} \propto \exp\left(m + \kappa^{(\mathrm{t})}_{k}\right)$ or simply point estimated (this latter behavior is the default).

Third, for each word in the document, $(n \in 1, \ldots, N_d)$ - Draw word's topic assignment based on the document-specific distribution over topics.

$$z_{d,n} \mid \vec{\theta}_d \sim Multinomial\left(\vec{\theta}_d\right)$$

- Conditional on the topic chosen, draw an observed word from that topic.

$$w_{d,n} \mid z_{d,n}, \beta_{d,k=z_{d,n}} \sim Multinomial\left(\beta_{d,k=z_{d,n}}\right)$$

2.3 Characterization of Idea Relations

Given a set of corpora, the above two methods will identify each corpus with a list of ideas. We then need a methodology to rigorously examine the relations between these ideas. One main contribution of Tan et al. (2017) is the proposal of a novel method to accomplish exactly this. The authors focus on two dimensions to characterize each relation between two ideas:

1. cooccurrence reveals to what extent two ideas tend to occur in the same contexts;
2. similarity between the relative prevalence of ideas over time reveals how two ideas relate in terms of popularity or coverage.

where the first dimension is captured by empirical pointwise mutual information (PMI) and the second by correlation between normalized document frequency of ideas. For details on this computation, please refer to the original paper.

In presence of these two dimensions, each relation between two ideas can then be put into one of the four categories below:

1. **Friendship**: correlated over time, likely to cooccur
2. **Head-to-head**: anti-correlated over time, unlikely to cooccur
3. Tryst: anti-correlated over time, likely to cooccur
4. **Arms-race**: correlated over time, unlikely to cooccur

We employ the same framework to capture relations between ideas in this study.

3 Result Replication

In Tan et al. (2017), the authors explored the ACL dataset consisting of papers from ACL, NAACL, EMNLP, and TACL from 1980 to 2014. This dataset consists of 4.8K papers, of which a processed version is made available.

Using the open-sourced code and dataset in the initial paper, we are able to replicate results that are very similar to ones presented in the paper.

Ori.	Rep.	First	Second
		Friendship	
1	1	word alignment	machine translation
		Arms-race	
1	1	sentiment analysis	machine translation
2	3	sentiment analysis	word alignment
23	38	rule,forest methods	discourse (coherence)
		Head-to-head	
1	3	discourse (coherence)	machine translation
7	Missing	discourse (coherence)	word alignment
38	Missing	rule,forest methods	sentiment analysis
		Tryst	
5	3	rule,forest methods	machine translation

Table 1: Result replication of the ACL-related data presented in Figure 8 of Tan et al. (2017). Column '**Ori.**' denotes the ranking of the relation as presented in Tan et al. (2017) while column '**Rep.**' denotes the ranking of the same relation in our experiment.

Running the **topic** method as described in Chenhao's work (Tan et al., 2017), we obtain a set of topics that are *associated* with particular topic words as determined by LDA (Blei et al., 2003b). The code outputs the top 50 relations for each of the proposed idea relations: **Friendship**, **Head-to-head**,

Identified Topic	Associated Topic Words
machine translation	translation, phrase, source, statistical machine, translation., system, mt, target, reordering, sentence
sentiment analysis	sentiment, opinion, positive, negative, polarity, reviews, review, words, aspect, product
word alignment	alignment, word, alignments, aligned, sentence, pairs, paraphrases, words, paraphrase, pair
discourse (coherence)	discourse, text, structure, relations, two, coherence, relation, focus, discourse., cue
rule,forest methods	rules, rule, rules., derivation, rules, , set, figure, derivations, forest, synchronous

Table 2: The set of topic words for each relation that appears in Table 1.

Tryst, **Arms-race**. Each of the two topics in these relations are represented by a set of words. We then need to manually name these topics and identify their associations.

Results of our experiment are shown in Table 1, which compares our result with the presented result. For details on which topic words are identified with what relations, see Table 2. Table 1 shows that for the most part, our results line up well with the ones presented in the original paper with slight differences (most notably on the two **Head-to-head** relations that were not found in the Top 50 results in our run). This could be due to randomness introduced in the topic modelling process but needs further investigation.

Notably, the association between identified topic and associated topic words also differs slightly with that presented in the original paper. In caption of Figure 8, Tan et al. (2017) identifies the **rule, forest methods** topic with 'rule, grammar, derivation, span, algorithm, forest, parsing, figure, set, string', which differs from the result of our experiment - 'rules, rule, rules., derivation, rules, ,set, figure, derivations, forest, synchronous'.

In our final report, we will also provide our replicated version of Figure 9 in Tan et al. (2017). This needs to be fine-tuned because the figures generated directly from LaTeX script are too corase for presentation.

In conclusion, we have successfully replicated the ACL results presented in Tan et al. (2017). The presented results mostly match up with our experiment.

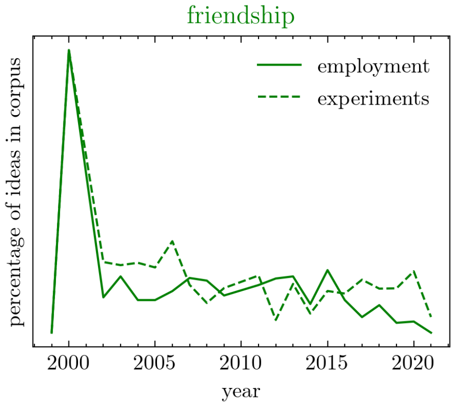

Figure 2: **Friendship** relation between 'employment' and 'experiments' in AER

4 Exploratory Studies

4.1 American Economics Review [NEW]

Chenhao suggests the idea of analyzing the American Economic Review, one of the most prestigious journals of Economics. AER journal is a representative of the intellectual history of the field. We were able to download over 4500+ journal articles from the 1999 to the 2021 from the AER official website. Here are some notable findings:

In Fig. 2, we see a **friendship** relation between "employment" and "experiments". The "employment" points to the topics including "employees, punishment, savings, contribution, default, monitoring, money, treatment, future, behavior"; the "experiments" points to the topics including "subjects, experiment, subject, rule, sub, experiments, rules, experimental, games, complexity". This friendship relation corresponds exactly to the relationship between the subject of economic research and its popular research methods. When the economists are studying the topics related to behaviors of employees under various motives within firms, it is usually conducted with some methods within the realm of experimental economics. So the **friendship** seems to be a fitting narrative here.

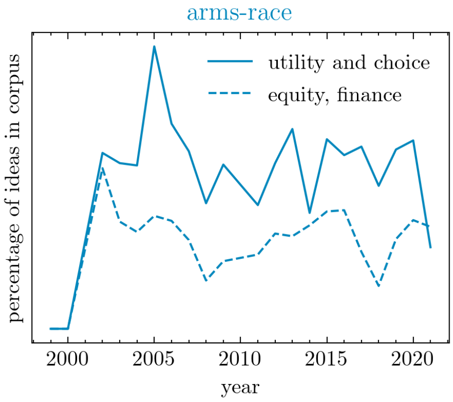

Figure 3: **arms-race** relation between 'utility and choice" and 'equity, finance' in AER

In Fig. 3, we observe an **arms-race** between utility and choice versus equity, finance. The "utility and choice" here stands for consumer preference, utility, and decision theory. The "equity, finance" stands for the topic of financial economics, which includes the study of "equity, returns, asset, stock, return, investors, investment, wealth, mortgage". By observing these two topics to unlikely to cooccur, we can propose the hypothesis is that when a paper uses an approach of a theoretical model of consumer preference and utility, it is unlikely to come across with the topics of financial economics because the latter one increasingly more relies on empirical methods instead of theoretical reasoning.

Another interesting question we can ask is based on a **head-to-head** relation between "price auction bidding" and "ethnicity, immigrants" shown in 5 . The first one refers to the topic of "price, auction, and bidding mechanisms", and the second one refers to the topic of "ethnic groups and mobility of immigrants". We do not know the exact reason why these two topics are unlikely to cooccur and anti-correlated. One possibility is that the bidding and auction mechanism designed for the immigration path is never an object of study because such policies are losing favors and they are unlikely to happen together. But the question of why they are anti-correlated remained a myth.

We do want to point out although we can extract many notable relations from our results, the results could have been better. We saw left-end towards the years 1999 and 2000 seem to contain one or many outliers in many graphs. Because we have to obtain the AER dataset, we had to initially extract DOI number from the AER websites and download all the corresponding papers from JSTOR and other related sites. we had encountered some issues when converting PDFs to textual corpora with PDFMiner. Among 4500+ original papers, we eliminated about 1000 papers with surprisingly high percentage gibberish words. To further improve the results of the exploratory studies on AER, we would need to fine-tune our corpus preprocessing and reexamine our method of converting PDFs to txt files and keep greater number of journals with a high percentage of meaningful words.

4.2 People's Daily (REN MIN RI BAO)

With the co

5 Comparison between sTM and raw TM

We choose to use People's Daily to compare the two topic models, because

1. The textual quality of the corpus is better
2. It is large and thick enough to demonstrate the difference
3. It has a long time frame so structural topic model might be effective

Although we used high-performance 128G-memory server, structural topic model could not be run probably, probably due to the limit of R language. Thus, we removed every word that appears in less than 10 documents or appear in more than half of all the documents. A brief comparison with the topics using the original corpus using raw topic model reveals that the quality of topics is not severely affected, and major historical themes are correctly identified.

We compare raw topic model and structural topic from 3 aspects. Firstly, we interpret the topics and check if important historical themes are captured. We discovered that both models accurately captured the themes of anti-imperialism, Chairman Mao and cultural revolution, socialism and communism, United Nations and diplomacy, Korea and Vietnam wars, market economy, firm and investment, etc. sTM tends to capture more detailed topics, like Israel and Palestine, India and Pakistan, and municipal constructions, while these topics were not captured by raw TM.

Secondly, to quantify the topic difference, we calculated the average number of characters of each word selected by the two topic models. Note that each Chinese word is composed by a number of characters, and usually the longer the word is, the more detailed and specific it is. The average word length for raw TM is 2.09 (std = 0.789), while that of sTM is 3.28 (std = 0.591). sTM tends to capture more detailed, specific words.

Third, we visualize the time range of documents in each topic. As the figure shows, while raw TM tends to capture topics that capture documents around 1976, sTM topics could vary in the mean year of related documents. Time range of Raw TM topics have standard deviation centered at 17 year, while sTM topics could be very specific to a period (low standard deviation) or vice versa.

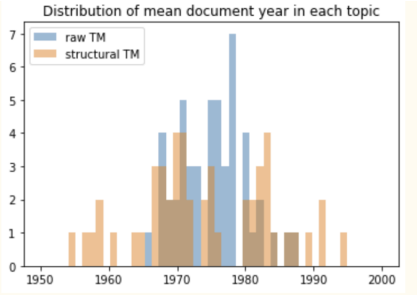

Figure 4: Distribution of mean document year for each topic

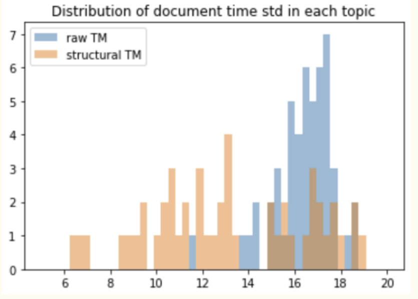

Figure 5: Distribution of document year standard deviation for each topic

To sum up, raw TM topics are more general and usually capture documents in a broader period. With metadata on the temporal change of prevalence of topics, sTM captures more specific topics of specific historical periods, but also long periods of time.

6 Related Work

We now present three strands of related studies in addition to what we have discussed.

6.1 Trends in ideas

Most previous studies before Chenhao's work focus on the trends of single ideas over an extended period of time. For example, Rule et al. (2015) presents a number of methods that reveal change in contents that masks continuity in the articulation of major governance task through analyzing State of Union Addresses from 1790 to 2014. Hall et al. (2008) applies unsupervised topic modeling to ACL anthology to analyze historical trends in the field of Computational Linguistics from 1978 to 206. There are not many papers out there studying the trends and dynamics between two ideas until (Tan et al., 2017). So our work attempted to further apply Chenhao's work to different fields such as Economics and different language of Mandarin Chinese.

6.2 Applications in Finance and Economics

There is growing popularity of application of Natural Language Processing in field of Economics and Finance mostly for predictive modelling and sentiment analysis based on Economics and Financial news. For example, Sesen et al. (2018) showcase different methods of getting inference from NLP models and evaluate their predictive performance based on different sources of news data. Vicari and Gaspari (2020) uses Deep Learning methods and specifically LSTM to analyze daily news headline from 2008 and 2016 for the purpose of developing optimal algorithmic trading strategy. We had very different vision: we attempt to apply idea relations analysis on the academic field of Economics with the goal of proposing social science hypothesis.

6.3 Structural Topic Modellings

Roberts et al. (2013a) develops the structural topic model that incorporate corpus structure and metadata in to the standard model and shows the model's use in two applied problems: the analysis of open-ended responses in a survey experiment about immigration policy, and understanding differing media coverage of China's rise. Since Roberts' paper, there is a growing number of application of structural topic modelling method. For example, Kuhn (2018) describes the application of structural topic modeling to Aviation Safety Reporting System data, and this application effectively reveals previously unreported issues and connections.

7 Conclusion

8 Application on Economic News [DELETE]

We are interested in how idea relations evolve in the news that reflect the US economy. We applied the open-sourced code in Tan et al. (2017) to a dataset of economic news. This data set (crowdflower, 2015) consists of news articles that are regarded relevant to the US economy, spanning from 1951 to 2014. The dataset was originally used for sentiment analysis.

8.1 Observation I: Economic Intuition Confirmed

The topics we found seem to confirm to the basic economic intuition. Here are some examples:

"inflation, prices, price, consumer, increase, increases, rate, rise, rising, higher" **and** "would, house, bill, congress, senate, committee, legislation, plan, proposed, members" are in a **head-to-head** relation (ranked 47). We see that the first group points to topics of rising price/inflation, and the second group points to the topic of government. These two topics are anti-correlated over time. This head-to-head relation makes sense because usually governments including the US can employ a contractionary monetary policy to fight inflation. So such anti-correlation seems to show such head-to-head tension.

"sales, auto, retailers, stores, retail, consumers, car, industry, cars, consumer" **and** "government, program, new, help, federal, make, programs, health, financial, could" are in a **arms-race** relation (ranked 19). The first group points to the topic of auto industry, and the second group points to topic of government assistance program. These two topic correlate over time but unlikely to co-occur. Because when the auto industry is at its prime time, the industry does not need government assistance. When the government assistance comes to play greater role, it usually indicates a struggling auto industry. They are indeed correlated over time but unlikely to co-occur.

"economic, world, global, markets, china, countries, european, international, asian, financial" **and** "billion, deficit, billion, , record, last, exports, trade deficit, imports, billion, year" are in a **tryst** relation (ranked 26). The first group indicates topic of trade with Asia and Europe. The second group points to trade deficit. These two topics are likely to co-occur, because when there is trade then there is trade deficit on one side. It also confirms their anti-correlated relationship, because a high US trade deficit would lead to protectionism in trade that will negatively impact the already-existing trade patterns and volumes.

8.2 Observation II : Problematic Observations and Probable Causes

There are too many groups of words that point to similar topics. In the example below, all three topics look the same. In the ACL results that we replicated, each group of words seems to point to a unique topic.

1. 'market, average, today, new york, , stock, new, high, week, list, gains' (appear in Rank 1 relation in **Friendship**)

2. 'stock, market, points, new york, dow jones, volume, new york, , stocks, average, point' (appear in Rank 1 relation in **Friendship**)

3. 'index, dow jones, stocks, industrial average, investors, fell, rose, points, , stock, points' (appear in Rank 1 relation in **Head-to-head**

Although there are differences in words like dow jones, stock, and fell, but they are all closely linked to the stock market and are too similar to provide an meaningful interpretations. Here are some reasons for this. First, it might be due to the nature of short economic news. The topics of economic briefings are very limited and this intrinsic lack of diversity lead to a lack of diversity in topics. Second, the lack of diversity of topics might be caused by the length of the individual article. Each economic news here is significantly shorter than an ACL paper. So the conciseness of the article doesn't allow the topics related to stock markets to dive into deeper subtopics. Due to these reasons, we will move to more complex corpus in our next step.

8.3 Related Works

We now present three strands of related studies in addition to what we have discussed.

8.3.1 Trends in ideas

Most previous studies before Chenhao's work focus on the trends of single ideas over an extended period of time. For example, Rule et al. (2015) presents a number of methods that reveal change in contents that masks continuity in the articulation of major governance task through analyzing State of Union Addresses from 1790 to 2014. Hall et al. (2008) applies unsupervised topi modeling to ACL anthology to analyze historical trends in the field of Computational Linguistics from 1978 to 206. There are not many papers out there studying the trends and dynamics between two ideas until (Tan et al., 2017). So our work attempted to further apply Chenhao's work to different fields such as Economics and different language of Mandarin Chinese.

8.3.2 Applications in Finance and Economics

There is growing popularity of application of Natural Language Processing in field of Economics and Finance mostly for predictive modelling and sentiment analysis based on Economics and Financial news. For example, Sesen et al. (2018) showcase different methods of getting inference from NLP models and evaluate their predictive performance based on different sources of news data. Vicari and Gaspari (2020) uses Deep Learning methods and specifically LSTM to analyze daily news headline from 2008 and 2016 for the purpose of developing optimal algorithmic trading strategy. We had very different vision: we attempt to apply idea relations analysis on the academic field of Economics with the goal of proposing social science hypothesis.

8.3.3 Structural Topic Modellings

Roberts et al. (2013a) develops the structural topic model that incorporate corpus structure and metadata in to the standard model and shows the model's use in two applied problems: the analysis of open-ended responses in a survey experiment about immigration policy, and understanding differing media coverage of China's rise. Since Roberts' paper, there is a growing number of application of structural topic modelling method. For example, Kuhn (2018) describes the application of structural topic modeling to Aviation Safety Reporting System data, and this application effectively reveals previously unreported issues and connections.

References

David M Blei, John D Lafferty, et al. 2007. A correlated topic model of science. *The annals of applied statistics*, 1(1):17–35.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003a. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003b. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

crowdflower. 2015. Economic news article tone. *data.world*.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. page 363–371.

Kenneth D. Kuhn. 2018. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, 87:105–122.

David M Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, volume 24, pages 411–418. Citeseer.

Margaret Roberts, Brandon Stewart, Dustin Tingley, and Edoardo Airoldi. 2013a. The structural topic model and applied social science. *Neural Information Processing Society*.

Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2019. Stm: An r package for structural topic models. *Journal of Statistical Software*, 91(1):1–40.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013b. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. 2015. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112(35):10837–10844.

M. Berkan Sesen, Yazann Romahi, and Victor Li. 2018. Natural language processing of financial news. *Big Data and Machine Learning in Quantitative Investment*, page Chapter 10.

Matt Taddy. 2013. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Chenhao Tan, Dallas Card, and Noah A. Smith. 2017. Friendships, rivalries, and trysts: Characterizing relations between ideas in texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Mattia Vicari and Mauro Gaspari. 2020. Analysis of news sentiments using natural language processing and deep learning. *AI & SOCIETY*.