https://doi.org/10.51408/issi2025_127

Old but Not Obsolete: Bag-of-Words vs. Embeddings in Topic Modeling

Jean-Charles Lamirel[1], Francis Lareau[2], Christophe Malaterre[3]

[1] *jean-charles.lamirel@loria.fr*
Université de Strasbourg, SYNALP-LORIA,
615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France (France)

[2] *lareau.francis@courrier.uqam.ca*
Université de Sherbrooke, Dept of Philosophy,
2500, boul. de l'Université, Sherbrooke (QC) J1K 2R1 (Canada),
Université du Québec à Montréal, Dept of Philosophy & CIRST,
455 bd. René-Lévesque Est, Montréal (QC) H3C 3P8 (Canada)

[3] *malaterre.christophe@uqam.ca*
Université du Québec à Montréal, Dept of Philosophy & CIRST,
455 bd. René-Lévesque Est, Montréal (QC) H3C 3P8 (Canada)

Abstract

Topic modeling techniques, including classical Bag-of-Words (BOW)-based methods like Latent Dirichlet Allocation (LDA) and emerging embedding-based models such as Top2Vec and BERTopic, are pivotal for uncovering latent themes in text corpora. This study builds upon previous work on an alternative BOW-based approach relying on feature maximization, CFMf, addressing limitations and extending comparisons along multiple metrics. Using a corpus of philosophy of science research articles (*N*=16,917), we evaluate LDA, CFMf, Top2Vec, and BERTopic across coherence, diversity, and recall metrics while also qualitatively examining top-word interpretability. Results reveal distinct trade-offs: while Top2Vec excels in coherence and diversity, it underperforms in recall and interpretability; BERTopic marginally outperforms LDA in coherence but not recall; CFMf balances these dimensions, outperforming others in coherence and diversity. Findings highlight the enduring relevance of BOW-based models and emphasize the modularity of topic modeling pipelines, advocating for hybrid approaches that integrate optimal components for improved performance.

Introduction

Topic modeling is a cornerstone in computational text analysis, aiming to uncover hidden themes in large corpora. Classical approaches, such as Latent Dirichlet Allocation (LDA), rely on statistical methods based on the Bag-of-Words (BOW) representation. Recently, embedding-based models such as Top2Vec and BERTopic have emerged as promising alternatives. In prior research, we highlighted the performance of a novel BOW-based method, Clustering and Feature Maximization with F1-measure (CFMf), though limitations remained, notably the generation of marginal topics with high document counts (Lamirel et al., 2024). The present study builds upon this work by addressing three objectives. First, we aim to mitigate the residual defects of CFMf. Second, we extend our comparative framework to include transformer-based models like BERTopic, which leverage Large Language Models (LLMs) and long-text embeddings. Finally, we investigate the modular nature of topic modeling, hypothesizing that combining the best components of various

approaches may yield a hybrid, high-performing model. Using a corpus of 16,917 philosophy of science research articles, we evaluate LDA, CFMf, Top2Vec, and BERTopic across multiple performance metrics, including coherence, diversity, and recall measures.

Methods overview

Topic models rely on a broad range of approaches to reveal hidden themes in extensive text corpora. Focusing on LDA, CFMf, Top2Vec, and BERTopic, we briefly describe these approaches notably in terms of text preprocessing, vectorization, clustering, ranking of documents and of words.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a generative statistical model that considers each document as a mixture of topics, each being a mixture of words with specific probabilities. It involves estimating Dirichlet distributions using techniques like Gibbs sampling or variational inference. It starts with tokenization, converting documents into word tokens, then representing them as Bag-Of-Words (BOW) vectors that quantify the tokens in each document. LDA's probabilistic clustering enables ranking of documents and words.

CFMf combines Feature Maximization (Lamirel et al., 2016) for feature selection, based on the F-measure, and Growing Neural Gas (GNG) for neural clustering (Fritzke, 1994). GNG is a winner-take-most algorithm that can utilize various metrics to capture a dataset's topology. To address a text size clustering bias observed when using the classical Euclidean metric (Lamirel et al., 2024), an angular metric is now deployed by renormalizing the cluster's prototype vectors during each learning step. GNG, like LDA, requires the number of topics beforehand. Key steps involve tokenization and BOW vectorization with a normalized TFIDF scheme. GNG clusters documents into topics, while the F-measure ranks words within topics. Cosine distance between topic's prototypes and documents is used for ranking documents.

Top2Vec original model (Angelov, 2020) utilizes Doc2Vec (Le & Mikolov, 2014) for semantic embedding of words and documents. Using HDBSCAN clustering technique (Campello et al., 2013), dense clusters emerge based on data density without the need to specify the number of topics. Each cluster is represented by its centroid taken as the average of cluster document embeddings. Top2Vec considers clusters as topics, using cosine similarity to centroids for reassigning ambiguous documents identified by HDBSAN. Key steps include tokenization and word/document embedding representation.

BERTopic original model (Grootendorst, 2022) employs transformer models like BERT (Devlin et al., 2019) to create deep contextual embeddings. HDBSCAN clusters documents using these embeddings. A BOW representation is used to rank words and documents through class-based TFIDF scores (c-TFIDF). Ambiguous cases from HDBSCAN are reassigned via cosine similarity between c-TFIDF representations. The process entails tokenization and dual vector representation: transformer-based for clustering and BOW-based for topic reassignment and document/word ranking.

Experimental protocol

The dataset comprised the complete collection of 16,917 full-text research articles from eight leading philosophy of science journals, as curated by Malaterre and Lareau (2022) and covering the period from 1930 to 2017. The corpus underwent standard preprocessing steps: tokenization, part-of-speech tagging, and lemmatization (TreeTagger package (Schmid, 1994) with Penn TreeBank (Marcus et al., 1993)). Words appearing in fewer than 50 sentences were excluded; only nouns, verbs, adverbs, and adjectives were retained. Documents were then vectorized to produce term-document matrices (TDMs) based on word frequencies for LDA and BERTopic, and on normalized TFIDF for CFMf.

LDA modeling was conducted via a Python API and used a word frequency TDM. CFMf was implemented with custom C and C++ code, using a normalized TFIDF TDM. Top2Vec was executed using a Python API, with the preprocessed corpus transformed by Doc2Vec serving as input. For BERTopic, full-text documents were used as inputs for generating document embeddings through a start-of-the-art method noted for its best average score on the Massive Text Embedding Benchmark Leaderboard: the stella model (stella_en_1.5B_v5) based on Alibaba-NLP and supporting the representation of long texts (131,072 tokens or more). BERTopic standard pipeline was performed with a Python API, using also the TDM for word ranking and outlier reassignment.[1]

Models were built for a number of topics from $K = 5$ to 100. For LDA and CFMf, predetermined values were chosen to sample this interval. For Top2Vec and BERTopic, specific values for the parameter corresponding to minimum cluster size were chosen through trial and error to generate models with different K values. Note that the terms "cluster", "class", or "topic" are used interchangeably. CFMf, Top2Vec, and BERTopic perform crisp clustering of documents and extract top-terms representing topics shared by documents of the same clusters. In contrast, LDA considers documents as probability distributions over topics; crisp clustering is obtained by grouping documents based on their dominant topic.

To compare model performance along complementary dimensions, four measures were used: (i) coherence, which indicates the extent to which top words in each topic are more meaningful when considered together (we used the coherence C_V of Röder et al. (2015)); (ii) topic diversity, which measures the distinctness of topic top words (expressed as the ratio of the number of unique top words in all topics by the total number of top words in all topics); (iii) a measure we call "micro inner recall" (*mIR*) which indicates the extent to which topic top words are found, on average, in the topic documents; and (iv) and "micro joint inner recall" (*mJIR*), which indicates how

[1] LDA Python API: https://github.com/lda-project/lda; CFMf implemented with custom C and C++ code available upon request (plans are to translate this method into Python and transfer it to GitHub in the near future); Top2Vec Python API https://github.com/ddangelov/Top2Vec; Doc2Vec Gensim implementation: https://github.com/piskvorky/gensim; BERTopic API: https://maartengr.github.io/BERTopic/api/bertopic.html; stella model stella_en_1.5B_v5, https://huggingface.co/dunzhang/stella_en_1.5B_v5.

well the top words of the clusters can all jointly recall the documents associated with these clusters. The latter two can be expressed as:

$mIR = \frac{1}{W \times |D|} \sum_{c=1}^{K} \sum_{i \in Top_W[c]} |\{d \in c \mid d[i] \neq 0\}|$ $mJIR = \frac{1}{|D|} \sum_{c=1}^{K} |\{d \in c \mid \exists\, i, i \in Top_W[c] \mid d[i] \neq 0\}|$

where W is the number of top words chosen as description of any cluster c, $|D|$ is the number of documents in the corpus, $Top_W[c]$ is the set of the top W words describing topic c, $d[i]$ represents the presence/absence of word i in the document d.

To gain qualitative insights into the relative topical coverage of the models and facilitate top-word comparison, clusters generated by CFMf, Top2vec and BERTopic were aligned to previously interpreted LDA topics (K=25) based on maximum number of shared documents.

Results

Results of the coherence measures across models show that coherence increases as a function of the number of topics K, reaching some sort of plateau after 50 topics for BERTopic or even earlier around 20-30 topics for the other three models (Fig. 1A). This indicates that topic top words tend to be specific to more narrowly defined clusters as K increases. Of the four approaches, Top2Vec displays the highest coherence at about 0.8 from K=20 onward. CFMf follows with coherence above 0.7 also from K=20 onward. While LDA exhibits the lowest coherence scores, reaching a plateau of about 0.55 from K=20 onward, it is slightly outperformed by BERTopic at lower K values and more significantly at higher K values.

As the number of topics K increases, topic diversity tends to decrease (Fig. 1B), which is to be expected since increasing the number of topics simultaneously increases the likelihood of overlap between top-words. Highest topic diversity—typically above 0.95—is obtained by Top2Vec across all values of K. CFMf ranks second, with diversity measures decreasing from about 0.9 below K=20 to 0.8 after. LDA and BERTopic reach about the same bottom value of about 0.65 after K=30, though LDA outperforms BERTopic for lower K values.

If one were to evaluate the models solely on coherence and diversity, then Top2Vec would come on top. Yet, the two measures of inner recall show a radically different perspective. In terms of micro inner recall—which is the average capability of topic top words to recall their sets of topic documents—, Top2Vec displays by far the lowest scores, below 0.35 for all K values (Fig. 1C). On the other hand, LDA exhibits the highest scores, consistently above 0.8. BERTopic follows, with *mIR* values decreasing from about 0.8 for K=10 to 0.7 for K=100. As for CFMf, it exhibits *mIR* scores that reach a plateau of about 0.6 from K=25 onward.

Measures of joint inner recall, which is the capability by all top words to jointly recall all corpus documents, single out Top2Vec as the least well-performing approach (Fig. 1D). Indeed, while *mJIR* scores for LDA, BERTopic and CFMf all reach about 1, *mJIR* measures for Top2Vec reach a plateau of about 0.9, starting from 0.7 to 0.8 scores for K values below 25. This shows the inability of top-words generated by Top2Vec to recall a remaining fraction (about 10%) of the corpus, even when increasing the number of topics or clusters.

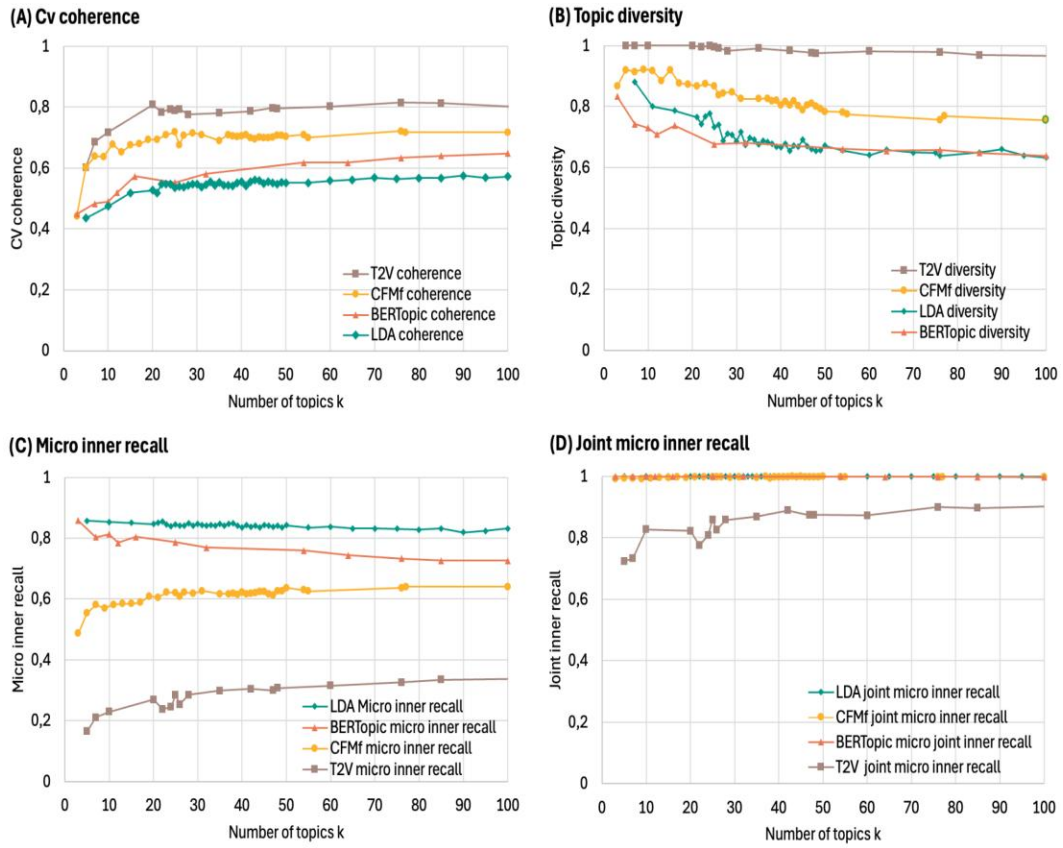

Fig. 1. Performance comparisons between topic models. (A) C_V coherence, (B) Topic diversity, (C) Micro inner recall *mIR*, (D) Micro joint inner recall *mJIR* (for $W = 10$ top-words).

Overall, the four approaches pick out topics that have a good descriptive similarity in terms of top words (Table 1). Yet nuances exist. Most striking is the weaker interpretability of Top2Vec top words, for instance for cluster (21) which mentions author names and technical disciplinary terms, or for cluster (16) which is about causation without naming it but mentioning author names. LDA, CFMf and BERTopic fare better in this respect. While CFMf still mentions author names in some topics—e.g. (8), (17), and (18)—they are fewer than Top2Vec and tend to be well aligned with easily interpretable topics. CFMf top words also tend to convey meaningful interpretations often more precise than LDA, e.g. distinguishing between relativity (2) and quantum mechanics (22), as Top2Vec and BERTopic also do. As for BERTopic, top words are also conducive to clear interpretations, although some of them remain very generic. Note some shifts in the overall balance of topics compared to the LDA model, with fewer topics related to philosophy of language and logic (0, 21, 20) and more topics related to rational decision (14, 18, 19) and especially philosophy of physics (17, 22, 13, 6, 5). What remains to be investigated is whether such changes are also related to changes in the relative proportions of the topics (as expressed in topical percentages or numbers of documents sorted by

dominant topics in LDA or number of cluster documents in BERTopic, CFMf and Top2Vec).

Table 1. Comparison of the top-words for *K*=25. LDA topic colors/labels as in (Malaterre & Lareau, 2022); for CFMf, Top2Vec and BERTopic, colors based on closest LDA topics; numbers are IDs; due to space reasons, only the top 4 words are listed, with abbreviations.

LDA		CFMf		Top2Vec		BERTopic	
Formal	set; function; relation; define.nition	(2)	proba.; propos.; theorem; condition.	(21)	sneed; balzer; moulines	(0)	sentence; truth; language; set
Language	language; sentence; term; meaning	(8)	carnap; wittgenstein; schlick; neurath	(8)	prerequisite; mortgage; exist.; false.	(21)	vague.ness; borderline; predicate
Mathematical	mathematical.tics; number; proof	(6)	math.; axiom; proof; geometry	(24)	nominalistic.ally; indispens.; colyvan	(20)	belief; agent; revision; model
Sentence	sentence; context; use; say	(1)	sentence; language; quine; speaker	(19)	quine.ean; synonymy; gavagai	(7)	belief.ve; know.ledge; epistemic
Truth	logic; truth; true; proposition	(0)	logic; modal; proposition; predicate	(1)	quantifier.cation; provable; semantics	(10)	theory; realist.m; scientific
Arguments	argument; claim; say; question	(9)	belief; epistemic; justif.; doxastic	(20)	ditmarsch; bisimilar; baltag; fagin	(24)	law; nature; generaliz.; statement
Knowledge	belief; knowledge; epistemic; know	(14)	realist.ism; fraassen; putnam	(6)	justified.cation; reliabilist; bivs	(1)	proba.; hypothesis; evidence
Sc.-theory	theory; scientific; empirical; realism	(3)	confirm.ation; hypothesis; inductive	(14)	anti.realist; pessimistic; nma	(14)	game; player; strategy; equilibrium
Confirmation	law; hypothesis; statement; evidence	(11)	proba.; bayes.; frequency; chance	(11)	longino; feminist; kourany; funding	(18)	moral; reason; action; normative
Experiment	datum; experiment; value; use	(10)	agent; game; player; utility	(0)	bayes.ians; probability; finetti	(19)	economic.s; theory; price
Probability	probability; measure; value; give	(24)	selection; pop.; fitness; evolutionary	(18)	morally.ty; baier; utilitarian.ism	(4)	selection; gene; organism; pop.
Agent-	agent; action; decision; game	(23)	gene; cell; organism; protein	(23)	replicat.; payoff; huttegger; signalers	(15)	function; teleological; artefact; goal
Evolution	selection; pop.; organism; gene	(12)	brain; cognitive; machine; mental	(5)	gene.notype; phenotype.pic; allele	(2)	mental; property; state; cognitive
Mind	behavior; state; mental; action	(4)	visual; perception; perceptual; color	(4)	neural.roscience; processing; cortex	(23)	information; dretske; signal
Neuroscience	system; inform.; process; cognitive	(15)	cause.ation; event; intervention	(16)	spirtes; intervention; scheines; pearl	(8)	cause.al; event; variable
Perception	object; experience; perception; color	(7)	model; simulation; datum; measure	(12)	idealize.ation; batterman; approxim.	(12)	model; repres.; system; target
Causation	cause.ation; event; effect	(5)	law; explanation.tory; hempel	(15)	mereolog.; markosian; truthmaking	(11)	explanation.tory; understanding; law
Explanation	model; explanation.tory; account	(19)	entropy; energy; atom; chemical	(17)	nonreductive; kim; physicalist.m	(17)	chemical, chemist.ry; substance
Property	property; world; object; relation	(2)	spacetime; einstein; relativity; clock	(22)	macro.microstate; microcanonical	(22)	measure.ment; scale; quantity
Particles	theory; energy; law; particle	(22)	quantum; particle; measure.; wave	(7)	inertial; spacetime; relativity.istic	(13)	entropy; time; system; state
Quantum	time; state; space; quantum	(17)	kant; newton; galileo; motion	(9)	eigenstate.s; quantum; superpos.	(6)	time; space; theory; relativity
Classics	motion; body; force; newton	(16)	science.tific; philosophy; history	(2)	seventeenth; newton; descartes	(5)	quantum; state; particle; measure.
History	work; time; man; history	(13)	moral; man; emotion; god	(13)	spiritual; conscience; dostoevsky	(9)	newton; motion; galileo; body
Philosophy	world; nature; knowledge; concept	(21)	social.ciety; science; economic	(3)	mankind; lundberg; society; civiliz.	(16)	theory.retical; term; model
Social	science.tific; social; research	(18)	kuhn; popper; laudan; lakatos	(10)	laudan; kuhn; kuhnian; lakatos	(3)	science.tific; theory; research

Discussion

Limitations of the study may concern the corpus used, especially its preprocessing quality and residual noise. Another limitation is our focus on four topic modeling approaches—many others remaining unexplored—and a set of metrics that only cast particular perspectives and all show obvious weaknesses. Nevertheless, the findings revealed significant trade-offs in performance. For example, Top2Vec excels in coherence and diversity but performs poorly in recall and interpretability. LDA and BERTopic perform well in recall but less so in coherence and diversity, favoring broader coverage. CFMf appears to balance these trade-offs effectively.

The study highlighted distinct advantages and drawbacks of the four approaches. Contrary to BOW-based approaches, embedding-based models like Top2Vec and BERTopic rely on text-representation learning: Doc2Vec requires a substantial amount of text to be effective while transformer-based models depend on the very large datasets used for training. Clustering methods also differ significantly. While the BOW-based approaches we tested require choosing the number of clusters beforehand, this can only be done indirectly for embedding-based methods using HDBSAN like Top2Vec and BERTopic, making it more difficult to identify an optimum model based on specific metrics. Also, while LDA performs fuzzy clustering, the other three approaches crisp-cluster documents and interpret clusters

as topics. As a result, handling ambiguity varies among methods. LDA represents documents as probability distributions over topics while Top2Vec and BERTopic rely on HDBSAN for document clustering and outlier detection and deploy specific approaches for outlier reassignment. The angular clustering adaptation implemented with CFMf solved the problem of outlier classes with high document count, and future work will evaluate a more fine-grained outlier reassignment strategy which could also impact small classes.

Extraction of top-words also vary significantly. While LDA simultaneously optimizes probability distributions for topics in documents and for words in topics, the other three approaches extract top-words in a second step after document clustering, for instance through word-topic embeddings distance for Top2Vec, c-TFIDF for BERTopic or Feature Maximization for CFMf. Future work will more systematically explore word ranking and topic profiling using word intrusion tasks.

Conclusion

Overall, the comparative study we conducted shows contrasting results for BOW-based models and embedding-based models. No single approach uniformly outperforms others across all metrics and top-word interpretability, underscoring the need for multiple evaluation perspectives: while Top2Vec reaches highest coherence and diversity scores, it falls behind in terms of recall and qualitative interpretability; BERTopic only slightly outperforms LDA in terms of coherence and diversity, but not recall; as for CFMf with its angular clustering adaptation, it appears to strike a balance between the different metrics, outperforming both LDA and BERTopic in terms of coherence and diversity, though not recall, and generating top-words with high interpretability. These findings show that statistical BOW-based models, far from being obsolete, stand the ground against recent embedding-based methods. They also reveal critical insights into the modularity of topic modeling pipelines.

Acknowledgments

J.-C.L. acknowledges funding from ANRT. F.L. acknowledges funding from Canada Social Sciences and Humanities Research Council (Postdoctoral Fellowships 756-2024-0557, Grant 430-2018-00899). C.M. acknowledges funding from Canada Social Sciences and Humanities Research Council (Grant 430-2018-00899) and Canada Research Chairs (CRC-950-230795).

References

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics* (arXiv:2008.09470). arXiv.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 7819, pp. 160–172). Springer Berlin Heidelberg.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T.

Solorio (Eds.), *Proc. of the 2019 Conf of the North Am. Chap. of the ACL* (pp. 4171–4186). ACL

Fritzke, B. (1994). A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems*, *7*.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv.

Lamirel, J.-C., Dugué, N., & Cuxac, P. (2016). New efficient clustering quality indexes. *2016 International Joint Conference on Neural Networks (IJCNN)*, 3649–3657.

Lamirel, J.-C., Lareau, F., & Malaterre, C. (2024). CFMf topic-model: Comparison with LDA and Top2Vec. *Scientometrics*. *129*, 6387–6405

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, 1188–1196.

Malaterre, C., & Lareau, F. (2022). The early days of contemporary philosophy of science. *Synthese*, *200*(3), 242.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, 44–49.