Text mining arXiv: a look through quantitative finance papers

Michele Leonardo Bianchi[a,1]

[a]*Financial Stability Directorate, Bank of Italy, Rome, Italy*

This version: April 8, 2024

Abstract. This paper explores articles hosted on the arXiv preprint server with the aim to uncover valuable insights hidden in this vast collection of research. Employing text mining techniques and through the application of natural language processing methods, we examine the contents of quantitative finance papers posted in arXiv from 1997 to 2022. We extract and analyze crucial information from the entire documents, including the references, to understand the topics trends over time and to find out the most cited researchers and journals on this domain. Additionally, we compare numerous algorithms to perform topic modeling, including state-of-the-art approaches.

Key words: quantitative finance, text mining, natural language processing, unsupervised clustering, topic modeling, research trends, named entity recognition.

1 Introduction

Quantitative finance is a field of finance that studies mathematical and statistical models and applies them to financial markets and investments, for pricing, risk management, and portfolio allocation. These models are needed to analyze financial data, to find the price of financial instruments and to measure their risk (see Vogl (2022) and Bianchi et al. (2023)). Readers are referred to Derman (2011) for an insightful exploration of the role of models in finance and to Ippoliti (2021) for some philosophical remarks on mathematics and finance.

The world of finance is always moving forward even in times of crisis. Innovations in finance come from the development of new financial services, products or technologies. Research trends in quantitative finance are driven not only by innovations, but also by structural changes in financial markets or by changes in regulation (Cesa (2017) and Carmona (2022)). When a structural change occurs some models are not anymore able to explain the phenomena observed in the market, consequently quants and researchers start working on new models. Examples of such changes are when the implied volatility smile appeared in 1987 (see Derman and Miller (2016)) or the Euribor-OIS spread materialized in 2007 (see Bianchetti and Carlicchi (2012)). Research activities driven by new products are, for instance,

[1]The author thanks arXiv, ChatGPT, and Google Scholar for use of their open access interoperability and Sabina Marchetti for her comments and suggestions. This publication should not be reported as representing the views of the Bank of Italy. The views expressed are those of the author and do not necessarily reflect those of the Bank of Italy.

the development of pricing models for interest rate and equity derivatives started in '90s, the structuring of credit products in the early 2000s, or the recent research trend on cryptocurrencies. New technologies applied to finance are namely the increasing role of big data and the advent of machine learning techniques. Regulation have an impact on the development of new quantitative tools for measuring, managing and monitoring financial risks (e.g. the Basel Accords).

In this paper we explore the arXiv preprint server, the dominant open-access preprint repository for scholarly papers in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. The papers in arXiv are not peer-reviewed but there are advantages in submitting to this repository, mainly to disseminate a paper without waiting for the peer review and publishing process, which can be slow (see Huisman and Smits (2017)). The arXiv collection provides a unique source of data to conduct various studies, including bibliometric, trend, and citation network analyses (see Clement et al. (2019)). It is a valuable resource for advancing scientific knowledge and conducting research on research, often referred to as *meta-research*. For example, Eger et al. (2019) and Viet and Kravets (2019) perform trend detection on computer science papers stored in arXiv, Lin et al. (2020) conduct a case study of computer science preprints submitted to arXiv from 2008 to 2017 to quantify how many preprints have eventually been printed in peer-reviewed venues, Tan et al. (2021) explore the images of around 1.5 million of papers held in the repository, Okamura (2022) investigates the citations of more than 1.5 million preprints on arXiv to study the evolution of collective attention on scientific knowledge, Bohara et al. (2023) train a state-of-art classification approach, and Fatima et al. (2023) design an algorithm to help researchers to perform systematic literature reviews.

We study all papers on quantitative finance, a small portion of the entire arXiv containing more than two millions of works at the time of writing. The choice is also motivated by our experience in this domain and by scientific curiosity.

The code is run on a standard desktop environment, without recurring to a big cluster. Scaling to a large number of papers may be not trivial. Dealing with a large amount of data requires significant computing resources, including processing power and memory, to manipulate and analyze the data efficiently. It is not simple, and maybe even impossible, to explore more than two million of papers with a standard desktop environment like ours.

The studies of papers on finance topics is not new in the literature. Burton et al. (2020) review the history of a well-known journal in this field and highlight its growth in terms of productivity and impact. The authors present a bibliometric analysis and identify key contributors, themes, and co-authorship patterns and suggest future research directions. Ali and Bashir (2022) conduct a systematic literature review and a bibliometric analysis on around 3,000 articles on asset pricing sourced from the top 50 finance and economics journals, spanning a 47-year period from 1973 to 2020. As observed by the authors, the exclusion of certain publications may potentially offer an alternative perspective on the landscape of existing asset pricing research. By using bibliometric and network analysis techniques, including the Bibliometrix Tool of Aria and Cuccurullo (2017), Sharma et al. (2024) investigate more than 4,000 papers on option pricing appeared from 1973 to 2019.

They follow the procedure suggested by Donthu et al. (2021). Their study aims to pinpoint high-quality research publications, to discern trends in research, to evaluate the contributions of prominent researchers, to assess contributions from different geographic regions and institutions, and, ultimately, to examine the interconnectedness among these aspects. The works of Burton et al. (2020), Ali and Bashir (2022) and Sharma et al. (2024) are focused on asset pricing or on a specific journal, their corpus is obtained by searching in the Scopus database some specific keywords, and the bibliometric analysis relies on VOS viewer (see Van Eck and Waltman (2010)) and Gephi (see Bastian et al. (2009)).

Our study explores all papers on quantitative finance collected in arXiv till the end of 2022 (around 16,000) and it considers text mining techniques implemented in Python to extract information directly from the portable document format (pdf) files containing the full text of the papers, excluding images, without relying on ad-hoc software or proprietary databases. Westergaard et al. (2018) found that examining the full text of documents significantly improved text mining compared to studies that only explored information collected from abstracts.[2] Their finding highlights the importance of using complete textual content for more comprehensive and accurate text mining and analysis.

The main objectives of our work are twofold. First, we explore the topics of the quantitative finance papers collected in arXiv in order to describe the evolution of topics over time. After having evaluated the performance of various clustering algorithms, we investigate on which themes researchers have focused their attention in the period from 1997 to 2022. Second, we try to understand who are the most prominent authors and journals in this field. Both analyses are performed with data mining techniques and without actually reading the papers.

The remainder of the paper is organized as follows. First, we provide a brief description of the data analyzed in this work (Section 2). Then, in Section 3 the preprocessing phase is discussed by offering further insights on the papers analyzed in our work. In Section 4 we compare various clustering algorithms and, after having selected the best performer, we explore, by splitting our corpus in 30 clusters, the evolution of topics over time. Finally, in Section 5 we describe an entity extraction process to investigate authors and journals with the largest number of occurrences in the corpus considered in this work. Section 6 concludes.

2 Data description

In this section we provide a description of the papers analyzed in this work. As observed above, there are various domains in arXiv (i.e. physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics) and each domain has is own categories. The categories within the quantitative finance domain are the following:

- computational finance (q-fin.CP) includes Monte Carlo, PDE, lattice and other numerical methods with applications to financial modeling;

[2] As a crosscheck, we conduct the analysis on both abstracts and full texts. The analysis using the full text data shows better results.

- economics (q-fin.EC) is an alias for econ.GN and it analyses micro and macro economics, international economics, theory of the firm, labor economics, and other economic topics outside finance;
- general finance (q-fin.GN) is focused on the development of general quantitative methodologies with applications in finance;
- mathematical finance (q-fin.MF) examines mathematical and analytical methods of finance, including stochastic, probabilistic and functional analysis, algebraic, geometric and other methods;
- portfolio management (q-fin.PM) deals with security selection and optimization, capital allocation, investment strategies and performance measurement;
- pricing of securities (q-fin.PR) discusses valuation and hedging of financial securities, their derivatives, and structured products;
- risk management (q-fin.RM) is about risk measurement and management of financial risks in trading, banking, insurance, corporate and other applications;
- statistical finance (q-fin.ST) includes statistical, econometric and econophysics analyses with applications to financial markets and economic data;
- trading and market microstructure (q-fin.TR) studies market microstructure, liquidity, exchange and auction design, automated trading, agent-based modeling and market-making.

These categories are assigned by the authors when they submit their papers. Even if it is possible to select multiple couples of domain-category belonging to more than one domain, we select as reference category only the first category within the quantitative finance domain. Figure 1 shows the numbers of papers on quantitative finance submitted to arXiv between 1997 and 2022. The increase in the last three years is mainly due to the q-fin.EC category.

As observed in Section 1, the code is implemented in Python and it is run under Ubuntu 22.04 on a desktop with an AMD Ryzen 5 5600g processor with 32GB of RAM. As we will describe in the following, numerous packages are considered.

As far as the collection process is concerned, we retrieve data from arXiv by selecting all categories within quantitative finance (i.e. q-fin). We collect articles metadata and pdf files for all articles from 1997 to 2022 for a total of around 16,000 articles (18GB of data).

While the metadata are obtained through *urllib.request* and *feedparser*, the pdf files are downloaded by means of the *arxiv* package. The metadata can be collected by following the suggestions provided in the arXiv web-pages. They are a fundamental input of the analysis and include the link to the paper main web-page, from where it is possible to extract the paper identification code (*id*, e.g. 2005.06390). The metadata contain information like authors names, paper title, primary category, submission and last update dates, abstract, and publication data when available (e.g. digital object identifier, DOI). Subsequent updates of

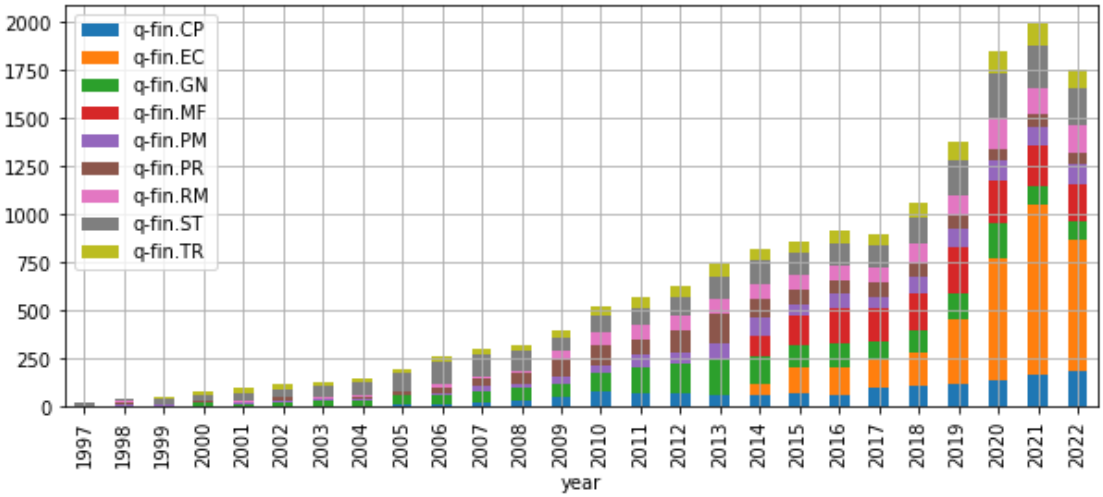

Figure 1: Categories by year.

the papers can be stored in the repository and for this reason there is the version number at the end of the paper *id*.

Since a paper could be assigned to multiple categories, a web-scraping tool written in Python allows us to retrieve from the paper main web-page the list of all categories of each single paper. We select from this list the subset of categories within the quantitative finance domain. Thus we assign as reference category of a paper the first category appearing in this subset. Starting from this list, we are able to filter and analyze all papers in the nine categories within q-fin.

The *pdftotext* package allows us to extract the text from pdf files. Each paper becomes a single (long) string. As discussed in Section 3, the length of these strings vary accross papers (see Figure 5), also because some documents are not papers (e.g. there are both theses and books). As a first assessment of the corpus, for each document we estimate the readability of the papers through *textstat*. As shown in Figure 2 the Flesch reading ease score (see DuBay (2004)) is on average equal to 65.7 (plain English), the lower and upper quartile are 59.91 (fairly difficult to read, but not far from the plain English) and 71.95 (fairly easy to read), respectively, and 99 per cent of the papers are in the range from 40.28 (difficult to read) and 88.20 (easy to read). There is only one paper with a negative value, but this is caused by the text contained in the figures. All other papers are above 17.17, that is above the *extremely difficult to read* level.

3 Text preprocessing

This section describes the text processing steps. The preprocessing phase is performed with *nltk*:(1) we split the text in tokens; (2) we extract the numbers representing years in the text;[3] (3) we identify all strings containing alphabet letters, and we refer to them as *words* even in the case they do not belong to the English

[3]We assume these numbers have 4 digits. We do not explore this data in the empirical analysis.

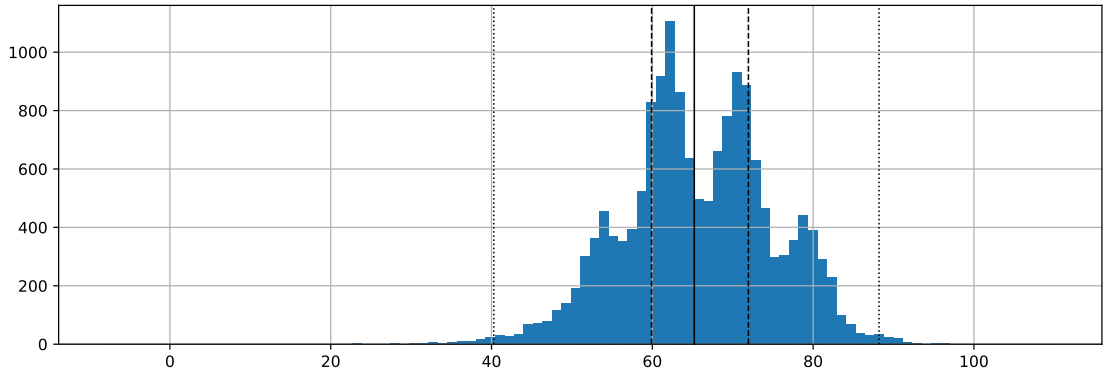

Figure 2: The Flesch reading ease score is reported. The vertical line represents the median, the dashed line is the quantile of level 0.25 (0.75), and the dotted line is the quantile of level 0.005 (0.995).

vocabulary; this step allows also to remove some symbols which are not recognized as letters in text analysis. Then, (4) we remove all stopwords and all words with length less than 3 characters; we also check whether there are words with more than 25 characters (quite uncommon in English); (5) we conduct a lemmatization by means of a part-of-speech tagger considering nouns, verbs, adjectives and adverbs; (6) we check if the paper is written in English by means of the *langdetect* package and we discard all non English papers. Both the extraction phase and the preliminary text analysis is parallelized by means of the *multiprocessing* package. We refer to the output of this first preprocessing phase as *lemmatized data*.

Thus, we analyze the frequency of the words across the whole corpus and we remove all words appearing less than 25 times. We discard also some words frequently used in writing papers on quantitative finance and which do not help in understanding the topic of the paper. The list of these words includes, for example, "proof" and "theorem", verbs commonly used in mathematical sentences (e.g.,"assume","satisfy", and "define"), mathematical functions (e.g. "min", and "log") and adverbs. The complete list is available upon request. In Figure 3 the list of the top 100 most frequent words obtained after this cleaning phase and their percentage of appearance is shown. The word "model" is extremely frequent (one every 100 words). The word "http" is also quite common, indicating that the papers full texts contain numerous internet links.

After a first preprocessing phase, we conduct an n-gram analysis by considering the *Phrases* model of the *gensim* package. We ignore all words and bigrams with total collected count over the entire corpus lower than 250 and set the score threshold equal to 10. We find the bigrams and then, to find trigrams and fourgrams, we apply again the same model to the transformed corpus including bigrams. This approach allows us to have a better ex-post understanding of the corpus which is full of n-grams (e.g. Monte Carlo simulation, Eisenberg and Noe, or bank balance sheet). The wordclouds of the main bigrams and trigrams (fourgrams) are depicted in Figure 4.

It should be noted that some topic modeling algorithms analyzed in Section 4 do not need text preprocessing. In those cases the input is just a single list

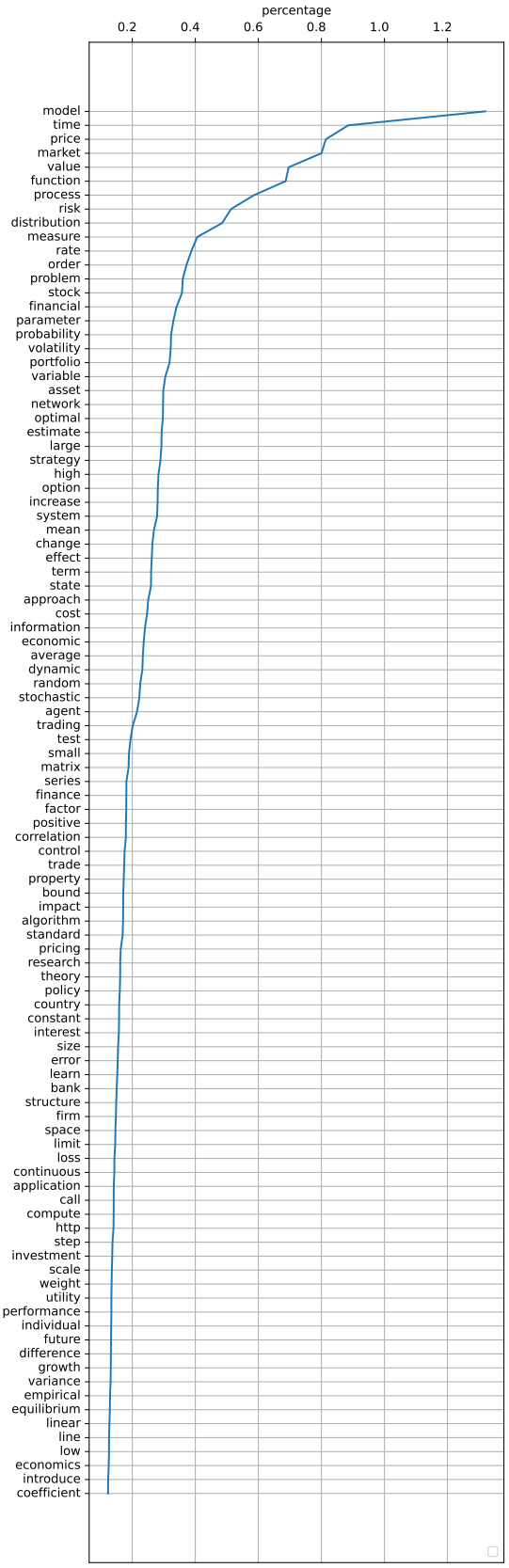

Figure 3: Frequent words of the corpus and their percentage of appearance.

Figure 4: Bigrams and tri(four)grams word clouds based on frequency with parameters *min count* equal to 250 and *threshold* equal to 10.

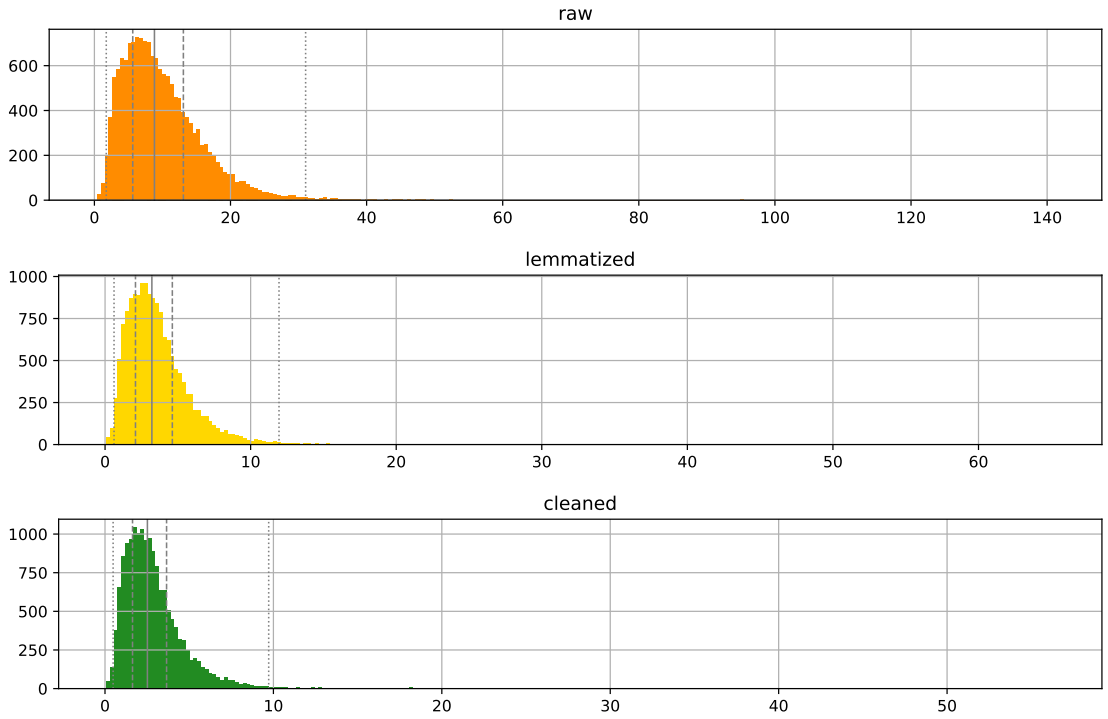

Figure 5: Papers length, in terms of number of words, for raw, lemmatized and cleaned data. The x-axis values are in thousands. The scale of x-axis varies across the three datasets. The vertical line represents the median, the dashed line is the quantile of level 0.25 (0.75), and the dotted line is the quantile of level 0.01 (0.99).

containing the whole paper text. While we refer to this latter input as *raw data*, we define the output of the preprocessing as *cleaned data*.

In Figure 5 we show how the length of the papers varies over the three data preprocessing phases. Starting from the raw data, containing all words and symbols, after a preliminary cleaning step we obtain the lemmatized data and then, after the last cleaning steps, the cleaned data. The number of words varies from a median value of 8,824 words for the raw data to around 2,518 words for the cleaned ones.

4 Topics trend

Now we are in the position to perform a topics trend analysis. We employ a topic modeling approach to identify the subjects discussed in the documents examined in this study, and then we observe how these topics change over time. Topic modeling refers to a class of statistical methods used to determine which subjects are prevalent in a given corpus. In Section 4.1 we select the best performing model among some selected approaches presented in the related literature. We evaluate these approaches by assessing their ability to accurately match the nine q-fin categories that researchers assign to their work when submitting it to arXiv. Then in Section 4.2, after having split the papers into 30 clusters, each one representing a specific topic, we discuss the evolution of research trends over time.

4.1 Algorithms performance on full texts

Topic modeling algorithms are widely used in natural language processing and text mining to uncover latent thematic structures in a collection of documents. Different algorithms have been developed, each with its own strengths and limitations (see Sethia et al. (2022)). The choice of a topic modeling algorithm depends on specific factors, such as the desired level of topic granularity and computational constraints. Some algorithms may require substantial computational resources and large amounts of training data. Here, we compare different algorithms by looking at some standard performance measures.

Since is not simple to assess the performance of different topic modeling algorithms (see Rüdiger et al. (2022)), we start by comparing the clusters assigned by each algorithm on the entire corpus to the nine clusters defined by the q-fin categories described in Section 2, that is the categories that researchers assign to their work during the submission process to arXiv. By exploiting a Bayesian optimization strategy, Terragni et al. (2021) present a framework for training, analyzing, and comparing topic models where the competitor models are trained by searching for their optimal hyperparameter configuration, for a given metric and dataset. Here we consider a simpler approach in which we compare the models by looking at some metrics. Evaluating topic modeling algorithms typically involves the use of performance measures, and it is important to note that different algorithms can yield varying results across these metrics.

We consider the following models.

- *K-means*. We perform a clustering analysis by considering the K-means algorithm implemented in *scikit-learn*. This algorithm groups data points into K clusters by minimizing the distance between data points and their cluster center. The document word matrix is created through the *CountVectorizer* function, that converts the corpus to a matrix of token counts. We ignore terms that have a document frequency strictly higher than 75%. T

- *LDA*. By considering the same document word matrix analyzed with the K-means algorithm, we perform topic modeling with the latent Dirichlet allocation (LDA). LDA is a well-known unsupervised learning algorithm. As observed in the seminal work of Blei et al. (2003), the basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. We study two different implementations of LDA (i.e. *scikit-learn* and *gensim*).

- *Word2Vec*. We train a word embedding model (i.e. Word2Vec) and then we perform a clustering analysis by considering again the K-means approach. An embedding is a low-dimensional space into which high-dimensional vectors are projected. Machine learning on large inputs like sparse vectors representing words is easiser if embeddings are considered. Ideally, an embedding captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space. The Word2Vec neural network introduces distributed word representations that capture syntactic and semantic word relationships (see Mikolov et al. (2013)). More in details, we

generate document vectors using the trained Word2Vec model, that is, we get numerical vectors for each word in a document, and then the document vector is the weighted average of the vectors. Thus, the K-means algorithm is applied to the matrix representing the corpus. We consider the Word2Vec model implemented in *gensim*.

- *Doc2Vec*. We create a vectorised representation of each document through the Doc2Vec model and then we perform a clustering analysis by considering the K-means approach. The Doc2Vec extends Word2Vec and it can learn distributed representations of varying lengths of text, from sentences to documents (see Le and Mikolov (2014)). We consider the Doc2Vec model implemented in *gensim*.

- *Top2Vec*. We study the Top2Vec model, an unsupervised learning algorithm that finds topic vectors in a space of jointly embedded document and word vectors (see Angelov (2020)). This algorithm directly detects topics by performing the following steps. First, embedding vectors for documents and words are generated. Second, a dimensionality reduction on the vectors is implemented. Third, the vectors are clustered and topics are assigned. This algorithm is implemented in an ad-hoc library named *Top2Vec* and it automatically provides information on the number of topics, topic size, and words representing the topics.

- *BERTopic*. We study a BERTopic model, which is similar to Top2Vec in terms of algorithmic structure and uses BERT as an embedder. As described in the seminal work of Grootendorst (2022), from the clusters of documents, topic representations are extracted using a custom class-based variation of term frequency-inverse document frequency (TF-IDF). This is the main difference with respect to Top2Vec. The algorithm is implemented in an ad-hoc library named *BERTopic*. The main downside with working with large documents, as in our case, is that information will be ignored if the documents are too long. A limited number of tokens are treated and anything longer is cut off. Since we are dealing with large documents, to work around this issue, we first split each documents in chunks of 300 tokens, thus we fit the model on these chunks. BERTopic does not allow one to directly select the number of topics, for this reason on the first step we obtain a number of topics much larger than the desired one. Since we obtain for each chunk the corresponding topic, we have for each document a list of possibly different topics and the length of these lists varies across documents (i.e. the length of a single list depends on the length of the corresponding document). To cluster this list of lists of topics, we consider each integer representing a topic as a word. Thus we use the Word2Vec algorithm described above to find similarities between these list of topics. Each topic label, that is the number representing the topic, is treated as a string, and Word2Vec transforms it into a numerical vector. We then apply K-means clustering to group these lists based on their similarity in the vector space. The resulting clusters reveal relationships and patterns among these lists and allow us to select the number of clusters we need for our purposes.

In theory both BERTopic and Top2Vec should use raw data since these algorithms rely on an embedding approach, and keeping the original structure of the text is of paramount importance (see Egger and Yu (2022)). However, raw data extracted from quantitative finance papers have a considerable amounts of formulas, symbols and numbers that may affect the algorithms performance. For this reason, we consider both raw and cleaned data. Additionally, for these state-of-the-art algorithms the number of extracted topics tends be large, but the algorithms offer the possibility to reduce the number of topics and of outliers, that can be larger than expected. The parameters of a BERTopic model have to be carefully chosen to avoid memory issues. Alternatively, it is possible to perform an online topic modeling, that is the model is trained incrementally from a mini-batch of instances. This result in a less resource demanding approach in terms of memory and CPU usage, but it generates less rich and less comprehensive outputs and for these reasons we do not consider this incremental approach here.

In Table 1 we report the following similarity measures between true and predicted cluster labels: (1) the rand score (RS) is defined as the ratio between the number of agreeing pairs and the number of pairs, and it ranges between 0 and 1, where 1 stands for perfect match; (2) the adjusted rand score (ARS), that is the rand score adjusted for chances, has a value close to 0 for random labeling, independently of the number of clusters and samples, and exactly 1 when the clusterings are identical (up to a permutation), however is bounded below by -0.5 for especially discordant clusterings; (3) the mutual info score (MI) is independent of the absolute values of the labels (i.e. a permutation of the cluster labels does not change the value of the score); (4) the normalized mutual information (NMI) is a normalization of the MI to scale the results between 0 (no mutual information) and 1 (perfect correlation); (5) cluster accuracy (CA) is based on the Hungarian algorithm to find the optimal matching between true and predicted cluster labels; (6) to compute the purity score (PS), each cluster is assigned to the class which is most frequent in the cluster, and the similarity measure is obtained by counting the number of correctly assigned papers and dividing by the number of observations. This latter score increases as the number of clusters increases and for this reason, it cannot be used as a trade off between the number of clusters and clustering quality, that is to find the optimal number of clusters.

The measures presented in Table 1 demonstrates that the Doc2Vec approach, when coupled with K-means clustering on cleaned data, outperforms other models. This is evident from the higher MI and PS measures it achieves compared to its competitors. Moreover, Doc2Vec exhibits practical advantages, as it is straightforward to implement and significantly reduces computing time when compared to more advanced techniques like BERTopic. Interpreting the results of the Doc2Vec approach is simple, as it allows for the easy identification of the most representative documents by retrieving the centroid vectors of each cluster. The Word2Vec approach, when coupled with K-means clustering on cleaned data, shows also a good performance.

It is worth noting that there are no significant differences in performance measures when applying either the Top2Vec or BERTopic methods to raw or cleaned data. This could be attributed to the fact that raw data contain mathematical formulas that do not contribute substantial additional information, even if, as shown

	RS	ARS	MI	NMI	CA	PS
K-means	0.570	0.029	0.232	0.136	0.271	0.297
LDA scikit-learn	0.823	0.194	0.608	0.284	0.376	0.460
LDA gensim	0.788	0.085	0.276	0.131	0.275	0.314
Word2Vec K-means	**0.832**	0.200	0.613	0.283	0.371	0.427
Doc2Vec K-means	0.831	0.220	**0.699**	0.325	0.388	**0.490**
Top2Vec raw	0.810	0.195	0.501	0.239	0.365	0.404
Top2Vec cleaned	0.811	0.206	0.530	**0.387**	0.253	0.416
BERTopic raw	0.826	0.238	0.608	0.289	**0.436**	0.458
BERTopic cleaned	0.821	**0.239**	0.574	0.276	0.398	0.429

Table 1: Algorithms performance. We report the rand score (RS), the adjusted rand score (ARS), the mutual info score (MI), the normalized mutual info score (NIM), the cluster accuracy (CA), and the purity score (PS).

in Figure 5, raw data have a larger number of words. This finding indicating an equivalence between raw or cleaned data could be also attributed to the relatively simple structure commonly found in quantitative finance papers. It is important to note that these findings may not generalize to papers or books with more intricate and complex text structures and without formulas.

LDA implemented in *scikit-learn* has a better performance than the LDA implementation in *gensim*. The plain K-means does not show satisfactory results, even if the algorithm can be implement without a great effort.

Finally, as an overall assessment, it is important to highlight that the performance metrics reported in Table 1 do not demonstrate particularly impressive results. This could partly stem from the wide-ranging nature of each q-fin category, encompassing numerous subtopics and arguments. Conversely, some papers can be classified under multiple categories, as it is not always obvious how to select a single definitive category for a given work.

4.2 Empirical study

As shown in Section 4.1, the best performing model is Doc2Vec with K-means clustering applied on cleaned data. This model is considered to have a better understanding of the topics discussed in the quantitative finance papers analyzed in this work. To obtain the desired number of topics we perform again a K-means clustering analysis. To extract the most representative documents, we retrieve the centroid vectors of each cluster. These centroids represent the average position of all document embeddings assigned to a particular cluster. For each cluster centroid, we find the nearest neighbors among the original Doc2Vec embeddings. These nearest neighbors are the documents that are closest to the centroid in the embedding space and can be considered as the main documents of that cluster. Thus we select for each cluster the 20 most representative documents and we find a label for the topic on the basis of the documents titles. Note that the underlying meanings of the topics are subject to human interpretation. However, also this phase is automated by asking the topic to ChatGPT (GPT-3.5) after having provided the list of 20 titles (see Ebinezer (2023)).

Given the size of our sample, a reduction down to 30 topics can be considered a good compromise for a topic analysis. The selected number of topics strikes a

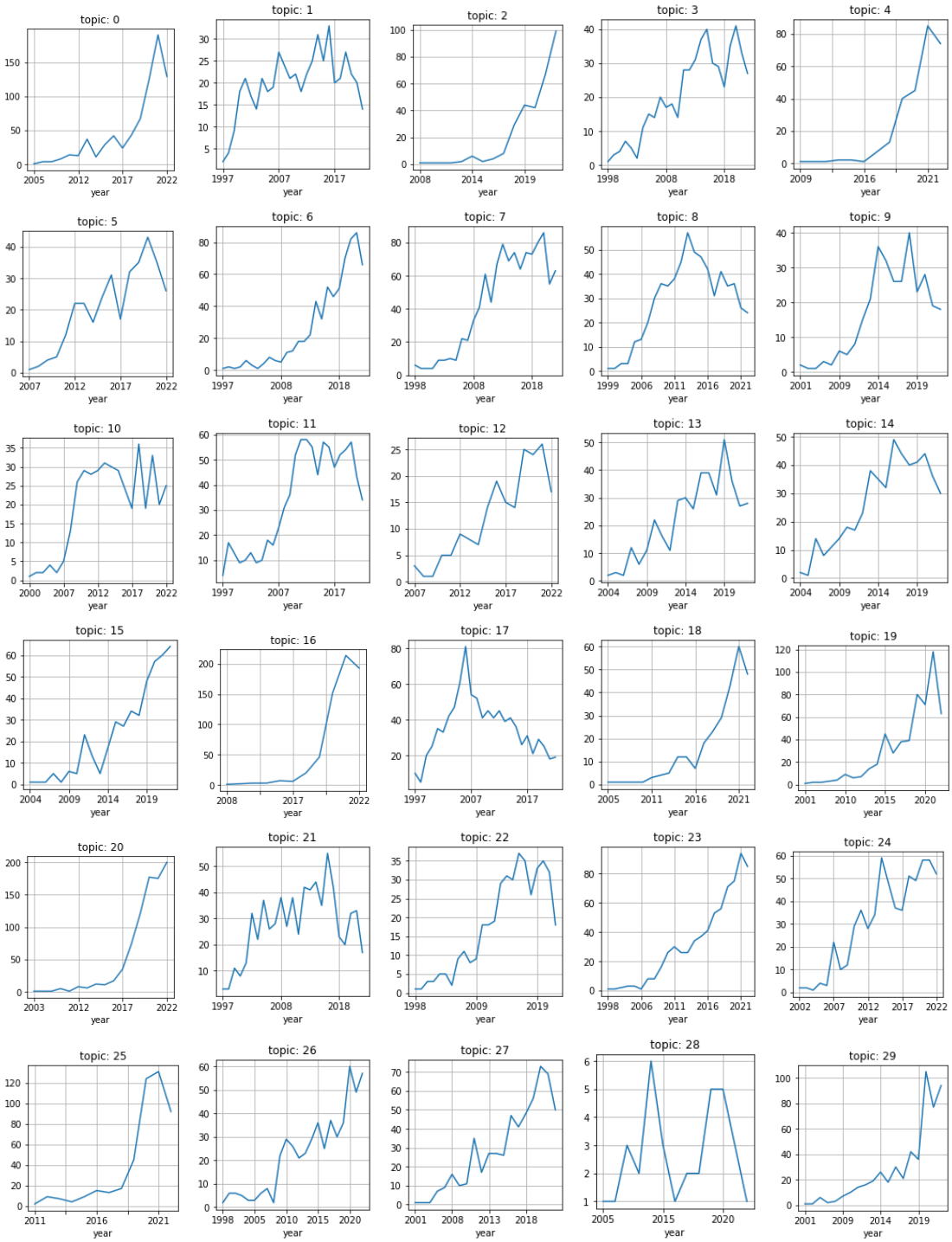

Figure 6: Topics trend by year across the sample of around 16,000 papers in the q-fin categories.

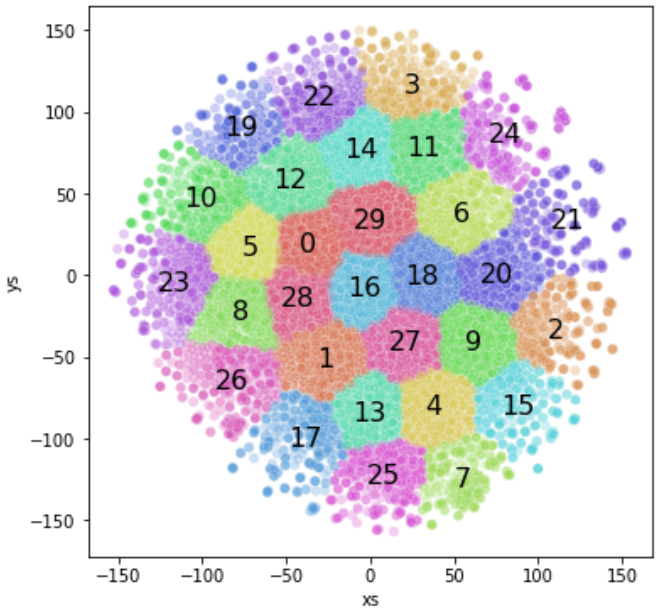

Figure 7: We show the topics clusters projected on the 2-dimensional space through the standard t-distributed stochastic neighbor embedding (t-SNE) algorithm. The numbers represent the topics reported in Table 2.

good balance between ensuring a sufficient quantity of documents for each topic and maintaining the desired level of granularity. This approach allows us to extract meaningful insights from the data while avoiding an excessive division of content that could affect our ability to identify overarching trends and patterns. As shown in Figure 6, most of the topics have an increasing trend (topic 28 seems to be the only exception). This is also motivated by the growth in the number of papers shown in Figure 1. For each topic, the list of 20 titles is the input for the question we ask to the ChatGPT chatbot. The topics label (i.e. the ChatGPT reply to the question) together with the title of the most representative paper are reported in Table 2.

It is interesting to observe that topics related to decentralized finance and blockchain technology (2) and stock price prediction with deep learning and news sentiment analysis (20) show a remarkable increase. This is also true for health, policy, and social impact studies, represented by the topic 16, and diverse perspectives in education, innovation, and economic development (0). Both topics are oriented towards economics. These topics trends depend also by the introduction in 2014 of the q-fin.EC category within the arXiv quantitative finance papers. Classical quantitative finance subjects like portfolio optimization techniques and strategies (6), stochastic volatility modeling and option pricing (7), game theory and strategic decision-making (19) high-order numerical methods for option pricing in finance (23) as well as, new themes appeared in the literature in the last years, like deep reinforcement learning in stock trading and portfolio management (4) and environmental and economic impacts of mobility technologies (25) have attracted the interest of researchers in the analyzed period. The representativeness of topic 28 is limited, mainly because the number of papers in this cluster is low, and one

could merge it with another cluster (i.e. 8).

It is clear that some topics are more related to economics than to finance. This also depends on the presence of the q-fin.EC category. For articles in this category there is not always a flawless alignment with quantitative finance.

To visualize the clusters, in Figure 7 the documents vectors are projected in the 2-dimensional space through the standard t-distributed stochastic neighbor embedding (t-SNE) algorithm. The larger the distance between topics, the more distinct the papers in those topics are in the original high-dimensional space. Conversely, it is also possible to have a better view on how close are topics each other (e.g. 8 and 28). It appears that the most specialized or narrowly-focused topics tend to occupy peripheral positions, while themes that are more aligned with economics are positioned closer to the center.

5 Extracting authors and journals

In this section we extract the author surnames by means of the *spacy* package. More in details, starting from the raw data, we perform a named-entities recognition. Since this approach extract both first names and surnames, we remove all first names by checking if these names are included in a list of about 67,000 first names. It should be noted that the number of occurrence of the surname of the author in a paper strongly depends on the citation style. Surnames are always reported in the references, but they do not necessary appear in the main text of a paper. Additionally, even if the author-date style is widely used (i.e. the citation in the text consists of the authors name and year of publication), the surname of the first author appears more frequently (e.g. Bianchi is more probable than Tassinari, even if Bianchi and Tassinari are coauthors of the same papers, together with other coauthors).

The algorithm is able to find the names and surnames occurring in the text. These are included in the PERSON entity types. The first 100 authors by number of occurrences in the corpus are selected. In order to have additional information on these authors, we obtain topics, number of citations, h-index and i10-index from Google Scholar (see Table 3). It should be noted that not all authors are registered in Google Scholar, even if they made a significant contribution to the field (e.g. Markowitz) or there are authors with the same surname and belonging to the same research field (see also Figure 3 in Sharma et al. (2024)). This is the case for some researchers we find in our corpus (e.g. Zhou, Bayraktar and Chakrabarti). For these last authors is not simple to find a perfect match in Google Scholar even if their number of occurrences is generally high.[4]

The algorithm is also able to find the most cited journals, included in the ORG entity types: Journal of Finance (4490 occurrences)*, Mathematical Finance (3785), Journal of Financial Economics (3325)*, Physica A (3137)*, Quantitative Finance (3044), Econometrica (2473)*, Journal of Econometrics (1971), American

[4]For the reasons described above we exclude from Table 3 the following researcher: Zhou, Markowitz, Peng, Jacod, Merton, Guo, Lo, Follmer, Yor, Almgren, Embrechts, Bayraktar, Artzner, Weber, Jarrow, Feng, Samuelson, Tang, Chakrabarti, Glasserman, Tsallis, Leung, Sato, Zariphopoulou, Kramkov, Karoui, Cizeau, Cao and Christensen.

Economic Review (1878), Insurance Mathematics and Economics (1667), Review of Financial Studies (2636)*, Journal of Banking and Finance (1542)*, Physical Review (1538), Journal of Economic Dynamics (1289), Energy (1267), Operations Research (1242), The Quarterly Journal of Economics (1238)*, Journal of the American Statistical Association (1204), Management Science (1160), European Journal of Operational Research (1066), Quantum (1043), IEEE Transactions (996), Journal of Political Economy (990), Journal of Economic Theory (977), Energy Economics (946), International Journal of Theoretical and Applied Finance (946), SIAM Journal on Financial Mathematics (888), Science (865), Expert Systems with Applications (845), Applied Mathematical Finance (743), Finance and Stochastics (736), Mathematics of Operations Research (652), PLoS (602), The Annals of Applied Probability (593), Stochastic Processes and their Applications (525), Energy Policy (520), International Journal of Forecasting (520), The European Physical Journal (514), Journal of Empirical Finance (510), Journal of Risk (509). We consider only journals with more than 500 occurrences and we exclude publishing houses. The journals with the asterisk symbol are identified with more than one name. It should be noted that some well-known journals are slightly below 500 occurrences. Furthermore, it is worth noting that papers with a strong mathematical focus tend to receive significantly fewer citations compared to papers that lean more towards economics or finance.

It is important to acknowledge that while the arXiv repository serves as a valuable resource for scholarly papers, it may not encompass the entirety of the quantitative finance research landscape. While the repository strives to be inclusive and comprehensive, there may be variations in the representation of scholars from different countries. Some scholars may have a relatively higher presence due to their active participation in submitting their research to arXiv. The platform content is reliant on authors voluntarily submitting their work, which introduces the possibility of selection bias. As a result, some authors and their contributions may not be represented. Therefore, our analysis and conclusions should be interpreted within the context of the available arXiv data, recognizing that there may be additional research and authors in the field of quantitative finance who have chosen alternative avenues for publishing their work. The same observation is true for the findings described in Section 4.2.

It is possible that influential scholars may not be as consistently represented or that, for various reasons, they have not regularly submitted their work to arXiv (see also Metelko and Maver (2023)). In the study of Sharma et al. (2024), focused on option pricing, some well-known authors are cited but they do not appear among the first 100 authors in our analysis. This discrepancy could be influenced by factors such as publication preferences, possible copyright issues, or institutional practices that may vary across different academic communities.

As a final remark, in our view, researchers in quantitative finance should consider submitting their work to arXiv due to the potential benefits it offers (see also Mishkin et al. (2020)). The delay between arXiv posting and journal publication, which can sometimes be more than a year, underscores the importance of submitting preprints to the repository. By doing so, researchers can help the community to understand research trends in their field more promptly, while also accelerating the dissemination of their own findings. This approach aligns with the findings of

Wang et al. (2020), who show that rapid and open dissemination through preprints helps scholarly and scientific communication, enhancing the reception of findings in the field. In the meanwhile, as a possible alternative to gain a comprehensive understanding of the entire landscape, future studies may consider incorporating other reputable academic databases and journals to ensure a more holistic exploration of quantitative finance research and its authors. Network approaches would also help to identify cliques and highly connected groups of authors. Insights from network clusters could further improve the understanding of the textual data investigated in this work.

6 Conclusions

In this study, we explore the field of quantitative finance through an analysis of papers in the arXiv repository. Our objectives are twofold: first, we investigate the evolution of topics over time, second, we identify prominent authors and journals in this domain. By employing data mining techniques, we achieve these goals without reading the papers individually.

The preprocessing phase, when needed, ensures the suitability of the data for subsequent analyses. Topic modeling helps in gaining insights and understanding the main themes and trends within our large dataset. By applying topic modeling algorithms, we identify the best performer and examine the temporal evolution of quantitative finance topics. This analysis reveals the changing research trends and highlights the emergence and decline of various topics over time.

Furthermore, we conduct an entity extraction process to identify influential authors and journals in the field. Through quantifying authors and journals occurrences, we shed light on the researchers who have made notable contributions to quantitative finance.

Our study demonstrates the power of data mining techniques in uncovering insights from a large-scale preprint repository. Our work not only showcases the power of data mining but also highlights the continued growth and dynamism of quantitative finance as a discipline. The techniques explored in this work can assist researchers in exploring and identifying novel research topics, discovering connections between different research areas, and staying up-to-date with the latest developments in the field. Furthermore, our methodology may serve as a roadmap for future studies on broader datasets or in other scientific domains utilizing text mining techniques. Although scaling to a larger number of papers may pose challenges, our approach provides valuable insights.

Finally, we believe that quantitative finance researchers should consider sharing their work on arXiv to potentially accelerate the dissemination and impact of their findings and to enhance the community understanding of research trends.

References

A. Ali and H.A. Bashir. Bibliometric study on asset pricing. *Qualitative Research in Financial Markets*, 14(3):433–460, 2022.

D. Angelov. Top2vec: Distributed representations of topics. `https://arxiv.org/abs/2008.09470`, 2020.

M. Aria and C. Cuccurullo. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4):959–975, 2017.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362, 2009.

M. Bianchetti and M. Carlicchi. Interest rates after the credit crunch: multiple-curve vanilla derivatives and SABR. `https://arxiv.org/abs/1103.2567`, 2012.

M.L. Bianchi, G.L. Tassinari, and F.J. Fabozzi. Fat and heavy tails in asset management. *The Journal of Portfolio Management*, 2023.

D. M Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

K. Bohara, A. Shakya, and B. Debb Pande. Fine-tuning of RoBERTa for document classification of arXiv dataset. In G. Shakya, S.and Papakostas and K.A. Kamel, editors, *Mobile Computing and Sustainable Informatics*, pages 243–255, Singapore, 2023. Springer Nature Singapore.

B. Burton, S. Kumar, and N. Pandey. Twenty-five years of The European Journal of Finance (EJF): A retrospective analysis. *The European Journal of Finance*, 26(18):1817–1841, 2020.

R. Carmona. The influence of economic research on financial mathematics: Evidence from the last 25 years. *Finance and Stochastics*, 26(1):85–101, 2022.

M. Cesa. A brief history of quantitative finance. *Probability, Uncertainty and Quantitative Risk*, 2(1):1–16, 2017.

C.B. Clement, M. Bierbaum, K.P. O'Keeffe, and A.A. Alemi. On the use of arXiv as a dataset. `https://arxiv.org/abs/1905.00075`, 2019.

E. Derman. *Models behaving badly: Why confusing illusion with reality can lead to disaster, on Wall Street and in life*. Wiley, 2011.

E. Derman and M.B. Miller. *The volatility smile*. Wiley, 2016.

N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W.M. Lim. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133:285–296, 2021.

W.H. DuBay. The principles of readability. *ERIC*, 2004.

S. Ebinezer. Transform your topic modeling with ChatGPT: Cutting-edge NLP. https://medium.com/, 2023.

S. Eger, C. Li, F. Netzer, and I. Gurevych. Predicting research trends from arXiv. https://arxiv.org/abs/1903.02831, 2019.

R. Egger and J. Yu. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in sociology*, 7, 2022.

R. Fatima, A. Yasin, L. Liu, J. Wang, and W. Afzal. Retrieving arXiv, SocArXiv, and SSRN metadata for initial review screening. *Information and Software Technology*, 161:107251, 2023. ISSN 0950-5849.

M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://arxiv.org/abs/2203.05794, 2022.

J. Huisman and J. Smits. Duration and quality of the peer review process: the author's perspective. *Scientometrics*, 113(1):633–650, 2017.

E. Ippoliti. Mathematics and finance: Some philosophical remarks. *Topoi*, 40: 771–781, 2021.

Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

J. Lin, Y. Yu, Y. Zhou, Z. Zhou, and X. Shi. How many preprints have actually been printed and why: a case study of computer science preprints on arXiv. *Scientometrics*, 124(1):555–574, 2020.

Z. Metelko and J. Maver. Exploring arXiv usage habits among Slovenian scientists. *Journal of Documentation*, 79(7):72–94, 2023.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781, 2013.

D. Mishkin, A. Tabb, and J. Matas. ArXiving before submission helps everyone. https://arxiv.org/abs/2010.05365, 2020.

K. Okamura. Scientometric engineering: Exploring citation dynamics via arXiv eprints. *Quantitative Science Studies*, 3(1):122–146, 2022.

M. Rüdiger, D. Antons, A. M Joshi, and T.O. Salge. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *Plos one*, 17(4): e0266325, 2022.

K. Sethia, M. Saxena, M. Goyal, and R.K. Yadav. Framework for topic modeling using BERT, LDA and K-Means. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 2204–2208. IEEE, 2022.

P. Sharma, D.K. Sharma, and P. Gupta. Review of research on option pricing: A bibliometric analysis. *Qualitative Research in Financial Markets*, 16(1):159–182, 2024.

K. Tan, A. Munster, and A. Mackenzie. Images of the arXiv: Reconfiguring large scientific image datasets. *Journal of Cultural Analytics*, 3:1–41, 2021.

S. Terragni, E. Fersini, B.G. Galuzzi, P. Tropeano, and A. Candelieri. OCTIS: comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, 2021.

N. Van Eck and L. Waltman. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2010.

N.T. Viet and A.G. Kravets. Analyzing recent research trends of computer science from academic open-access digital library. In *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 31–36, 2019.

M. Vogl. Quantitative modelling frontiers: a literature review on the evolution in financial and risk modelling after the financial crisis (2008–2019). *SN Business & Economics*, 2(12):183, 2022.

Z. Wang, Y. Chen, and W. Glänzel. Preprints as accelerator of scholarly communication: An empirical analysis in Mathematics. *Journal of Informetrics*, 14(4): 101097, 2020.

D. Westergaard, H.H. Stærfeldt, C. Tønsberg, L.J. Jensen, and S. Brunak. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Computational Biology*, 14(2), 2018.

number	label	title of the most representative paper
0	diverse perspectives in education, innovation, and economic development	Perspectives in public and university sector co-operation in the change of higher education model in Hungary, in light of China's experience
1	modeling financial market dynamics	Comment on: Thermal model for adaptive competition in a market
2	decentralized finance and blockchain technology	Understanding the maker protocol
3	correlation analysis in financial markets and networks	Random matrix theory and cross-correlations in global financial indices and local stock market indices
4	deep reinforcement learning in stock trading and portfolio management	Practical deep reinforcement learning approach for stock trading optimal market making by reinforcement learning
5	optimal trading and portfolio liquidation strategies in financial markets	An FBSDE approach to market impact games with stochastic parameters
6	portfolio optimization techniques and strategies	Seven sins in portfolio optimization
7	stochastic volatility modeling and option pricing	On the uniqueness of classical solutions of Cauchy problems
8	asset pricing, investment, and arbitrage in financial markets	Characterization of arbitrage-free markets
9	network analysis of financial contagion and systemic risk	Clearing algorithms and network centrality
10	counterparty risk and valuation adjustments in financial derivatives	Collateral margining in arbitrage-free counterparty valuation adjustment including re-hypotecation and netting
11	quantum models in finance and option pricing	Sornette-Ide model for markets: Trader expectations as imaginary part
12	valuation and risk management in annuity and insurance products	A policyholder's utility indifference valuation model for the guaranteed annuity option
13	optimal dividend strategies in stochastic control and risk management	Optimal dividends problem with a terminal value for spectrally positive Levy processes
14	risk measures and utility maximization under model uncertainty	On the C-property and w^*-representations of risk measures
15	economic complexity, networks, and trade patterns	Economic complexity and growth: Can value-added exports better explain the link?
16	health, policy, and social impact studies	Ramadan and infants health outcomes
17	statistical analysis of financial markets and volatility	Volatility distribution in the S&P500 Stock Index
18	renewable energy economics and electricity market dynamics	On wholesale electricity prices and market values in a carbon-neutral energy system
19	game theory and strategic decision-making	Simultaneous auctions for complementary goods
20	stock price prediction with deep learning and news sentiment analysis	Stock Prediction: a method based on extraction of news features and recurrent neural networks
21	kinetic wealth exchange models in economics	Gibbs versus non-Gibbs distributions in money dynamics
22	market order flow and price impact	Order flow and price formation
23	high-order numerical methods for option pricing in finance	High-order compact finite difference scheme for option pricing in stochastic volatility with contemporaneous jump models
24	optimal investment and consumption in financial models with constraints	Recursive utility optimization with concave coefficients
25	environmental and economic impacts of mobility technologies	A review on energy, environmental, and sustainability implications of connected and automated vehicles
26	advanced risk measures in financial modeling	Generating unfavourable VaR scenarios with patchwork copulas
27	Bayesian models for financial tail risk forecasting	A semi-parametric realized joint value-at-risk and expected shortfall regression framework
28	pricing and modeling options in stochastic volatility models with jumps	Semi-analytical pricing of barrier options in the time-dependent Heston model
29	economic growth and market dynamics	Uncovering volatility dynamics in daily REIT returns

Table 2: For each topic we report the label extracted from ChatGPT and the title of the most representative paper.

author_id	occurences	name	topics	citations	h-index	i10-index
mG07_6k	4505.0	Ioannis Karatzas	['stochastic analysis', 'stochastic control', 'mathematical finance']	35724	60	127
58amEmw	4441.0	Jean-Philippe Bouchaud	['statistical mechanics', 'disordered systems', 'random matrices', 'quantitative finance', 'agent based models']	49794	105	351
ahLm1v0	3256.0	Walter Schachermayer	[]	15891	56	124
HGsSmMA	3135.0	Didier Sornette	['cooperation', 'organization', 'patterns', 'prediction']	53371	112	587
CqFCQVE	2794.0	Alexander Schied	['probability theory', 'stochastic processes', 'mathematical finance']	16780	35	71
2QOp9_M	2457.0	Peter Carr	['financial engineering', 'quantitative finance', 'mathematical finance', 'derivatives', 'volatility']	24060	62	117
mVF1X_U	2371.0	Freddy Delbaen	['mathematik', 'ökonomie']	25557	50	90
ElAtiUs	2258.0	Darrell Duffie	['finance', 'central banking']	58442	88	142
fq7BQos	2042.0	Dilip B. Madan	['mathematical finance', 'general equilibrium theory']	25131	55	156
8abFiFM	1774.0	Robert Engle	['finance and econometrics']	189678	118	229
GU9HgNA	1743.0	Quanquan Gu	['statistical machine learning', 'nonconvex optimization', 'deep learning theory', 'reinforcement learning', 'ai for science']	14915	56	161
Q7N-rCk	1672.0	Rosario Nunzio Mantegna	['econophysics', 'statistical physics', 'complex systems', 'financial markets', 'information filtering']	28269	67	139
QsYYhSE	1629.0	Søren Johansen	['matamatical statistics', 'econometrics']	97728	65	132
vZA2pjw	1627.0	Benoît B. Mandelbrot	['mathematics', 'fractals', 'economics', 'information theory', 'fluid dynamics']	142895	96	319
Lf1kf1Q	1505.0	Mario Coccia	['evolution of technology', 'scientific change', 'social dynamics', 'complex adaptive systems', 'environment & covid-19']	19376	106	228
9HXRjPk	1444.0	Damir Filipovic	['quantitative finance', 'quantitative risk management']	6910	39	73
6_INHZI	1389.0	Fabrizio Lillo	['quantitative finance', 'statistical mechanics', 'data science']	10900	51	121
mGpnlA8	1218.0	Touzi Nizar	['stochastic control', 'mathematical finance', 'monte carlo methods']	11831	56	120
rp-3Yoo	1187.0	Barry Williams	['banks and banking', 'bank risk', 'multinational', 'banking']	2342	17	23
MZNxzRY	1161.0	Huyên Pham	['mathematical finance', 'stochastic control', 'numerical probabilities']	9967	54	135
3HhvEUc	1147.0	Yuri Kabanov	['mathematical finance', 'mathematics']	6297	38	77
-YEPo1E	1143.0	Wing-Keung Wong	['financial economics', 'econometrics', 'investment theory', 'risk management', 'operational research']	14440	65	274
GyPrRgc	1138.0	Swarn Chatterjee	['financial planning', 'wealth management', 'financial literacy', 'household finance', 'behavioral finance']	2719	28	54
RZid9X8	1075.0	Guido Caldarelli	['network theory', 'network science', 'statistical physics', 'complex systems']	24165	71	191
ImhakoA	1075.0	Daniel Kahneman	[]	519507	158	369
zO_tShM	1050.0	Marek Rutkowski	['mathematical finance', 'stochastic processes']	7559	30	67
7NJ7Ax8	1039.0	Patrick Cheridito	[]	5400	34	59
nyfza90	1019.0	Volker Schmidt	['virtual materials testing', 'statistical learning', 'image analysis', 'spatial stochastic modeling', 'monte carlo simulation']	11896	52	230
x4vtSxI	1017.0	Rene Carmona	['stochastic analysis', 'financial mathematics', 'financial engineering']	17474	59	139
kukA0Lc	999.0	Yoshua Bengio	['machine learning', 'deep learning', 'artificial intelligence']	656874	222	763
vQ0_nz8	989.0	Emmanuel Bacry	['self-similarity', 'multifractal', 'stochastic modeling', 'statistical finance', 'financial time-series modelization']	11937	47	69
1XwLUrc	980.0	Jim Gatheral	['volatility modeling', 'market microstructure', 'algorithmic trading']	6126	30	42
3HwRbiQ	955.0	Jerome Friedman	[]	283058	95	197
e2Xowj0	900.0	Neil Shephard	['econometrics', 'economics', 'statistics', 'financial econometrics', 'finance']	42035	69	140
pEnxwCM	887.0	Victor M. Yakovenko	['condensed matter theory', 'econophysics']	8891	44	101
a11vssU	845.0	Constantinos Kardaras	['stochastic analysis', 'probability', 'mathematical finance']	1557	20	31
79htA7g	838.0	Bent Flyvbjerg	['project management', 'management', 'infrastructure', 'planning', 'cities']	73264	70	152
zH1qBSo	834.0	Albert Shiryaev	['probability theory']	35521	59	163
QVb4LGI	815.0	Andrey Itkin	['mathematical finance', 'computational finance', 'derivatives', 'quantitative finance', 'machine learning']	709	14	19
Zuhod6s	813.0	Yong Deng	['uncertainty', 'deng entropy', 'information volume', 'random permutation set', 'chaos and fractal']	23189	81	335
bWlZ3-Y	810.0	Eric Jacquier	[]	4458	19	25
GKthQJQ	804.0	Peter K. Friz	['rough path theory', 'stochastic analysis', 'pdes', 'finance']	5678	39	82
2qTa_4U	794.0	Francis Diebold	['economics', 'econometrics', 'time series', 'statistics']	76159	97	175
utY1nTo	794.0	Matteo Marsili	['statistical mechanics', 'stochastic processes', 'collective phenomena in socio-economic systems', 'networks', 'complex systems']	10093	50	139
ZpG_cJw	783.0	Robert Tibshirani	['statistics', 'data science', 'machine learning']	460493	172	525
65wdZxA	780.0	Damiano Brigo	['probability', 'mathematical finance', 'stochastic analysis', 'signal processing', 'differential geometry and statistics']	9663	42	114
bxJe87s	780.0	Marco Frittelli	['financial mathematics', 'mathematical finance', 'probability']	3795	24	33
aVju7cI	771.0	Monique Jeanblanc	['mathématiques financières']	10256	51	110
-iOn6uI	769.0	Aurélien Alfonsi	[]	2731	22	30
P_LECrk	750.0	Tomasz R. Bielecki	['mathematical finance', 'stochastic processes', 'stochastic control', 'stochastic analysis', 'probability']	6901	39	89
5sQ0Fag	729.0	Ajit Singh	[]	21592	60	360
6quAJUE	706.0	Josef Teichmann	['mathematical finance', 'machine learning in finance', 'rough analysis']	3019	30	67
58amEmw	705.0	Jean-Philippe Bouchaud	['statistical mechanics', 'disordered systems', 'random matrices', 'quantitative finance', 'agent based models']	49794	105	351
JicYPdA	691.0	Geoffrey Hinton	['machine learning', 'psychology', 'artificial intelligence', 'cognitive science', 'computer science']	687453	180	436
i2MC67A	679.0	J.F. Muzy	['multifractal analysis', 'econophysics', 'turbulence']	13417	55	83
K9yGky8	678.0	Andreas Kyprianou	['probability theory', 'applied mathematics']	7946	44	99
aCSds20	670.0	Xavier Gabaix	['economics', 'finance']	30926	56	75

author_id	occurences	name	topics	citations	h-index	i10-index
G-WPCrM	667.0	Diego Garlaschelli	['network theory', 'econophysics', 'sociophysics', 'statistical physics']	7394	41	73
YTCnA4E	664.0	Eduardo Schwartz	['finance']	44652	81	141
A0ISJPU	664.0	Steven Shreve	['probability', 'financial mathematics']	35605	42	70
fFFOHec	660.0	Alexander McNeil	[]	24473	41	60
dYwbc9s	659.0	Guido Imbens	['causal inference', 'econometrics']	90765	95	169
OQK4DDY	657.0	Peter Forsyth	['scientific computing', 'computational finance', 'numerical solution of pdes']	10617	58	143
c1wQ9_k	655.0	Daojian Zeng	['natural language processing']	4877	13	17
Vs7kOf4	645.0	Marcel Nutz	['optimal transport', 'mathematical finance', 'game theory']	2339	30	43
vjc1kF0	640.0	Francesca Biagini	['financial and insurance mathematics', 'stochastic calculus', 'probability']	2598	20	42
zGJKZpk	629.0	Marianne Bertrand	[]	64693	66	119
nEfnJZM	628.0	Vadim Linetsky	['mathematical finance', 'financial economics']	5260	39	63
Bekg2Qo	621.0	Joel Shapiro	['financial intermediation', 'regulation of financial institutions', 'corporate governance', 'industrial organization']	3449	15	16
KDhGvNQ	611.0	Johanna Ziegel	['statistical forecasting', 'risk measures', 'postitive definite functions', 'stereology', 'copulas']	2088	19	31
r5PHkCs	610.0	Thomas Guhr	['theoretical physics']	8395	39	104

Table 3: Number of authors occurrences and Google Scholar metrics as of May 2023.