Text Mining for Big Data Analysis in Financial Sector: A Literature Review

Pejić Bach, Mirjana; Krstić, Živko; Seljan, Sanja; Turulja, Lejla

sustainability

MDPI

Review

Text Mining for Big Data Analysis in Financial Sector: A Literature Review

Mirjana Pejić Bach [1,*], Živko Krstić [2], Sanja Seljan [3] and Lejla Turulja [4]

[1] Faculty of Economics & Business, University of Zagreb, 10000 Zagreb, Croatia
[2] Atomic Intelligence, 10000 Zagreb, Croatia; zivko.krstic@live.com
[3] Faculty of Humanities and Social Sciences, Information and Communication Sciences, University of Zagreb, 10000 Zagreb, Croatia; sanja.seljan@ffzg.hr
[4] School of Economics and Business, University of Sarajevo, 71000 Sarajevo, Bosna i Hercegovina; lejla.turulja@efsa.unsa.ba
* Correspondence: mpejic@efzg.hr

Received: 15 January 2019; Accepted: 21 February 2019; Published: 28 February 2019

Abstract: Big data technologies have a strong impact on different industries, starting from the last decade, which continues nowadays, with the tendency to become omnipresent. The financial sector, as most of the other sectors, concentrated their operating activities mostly on structured data investigation. However, with the support of big data technologies, information stored in diverse sources of semi-structured and unstructured data could be harvested. Recent research and practice indicate that such information can be interesting for the decision-making process. Questions about how and to what extent research on data mining in the financial sector has developed and which tools are used for these purposes remains largely unexplored. This study aims to answer three research questions: (i) What is the intellectual core of the field; (ii) Which techniques are used in the financial sector for textual mining, especially in the era of the Internet, big data, and social media; (iii) Which data sources are the most often used for text mining in the financial sector, and for which purposes? In order to answer these questions, a qualitative analysis of literature is carried out using a systematic literature review, citation and co-citation analysis.

Keywords: big data; text mining; financial sector; data science; language

1. Introduction

The financial sector generates a vast amount of data like customer data, logs from their financial products, transaction data that can be used in order to support decision making, together with external data, like social media data and data from websites. Finacle Connect (2018) [1] indicates the top 10 technologies for financial industries, including the rise of API economy, cloud business enablement, blockchain for banking, and usage of artificial intelligence. Turner et al. (2012) [2] in the Executive Report prepared for the IBM Institute for Business Value indicate that 71% of banking and financial institutions use big data analytics for generating a competitive advantage relevant for their organizations. The same authors state that in 2010 there were 36% of such banking and financial institution, indicating the increase of 97% in two years. This increase points out the relevance of big data technologies in today's business for long-standing business challenges in the banking and financial sector. Applications of big data in the financial sector are various, including social media analysis, web analytics, risk management, fraud detection, and security intelligence. One of the possible roads to extract information from the vast amount of big data is text mining or text analytics (Pejic-Bach et al., 2019) [3].

The aim of text mining (also referred to as text data mining and text analytics) is to analyze textual document (including emails, reviews, plain texts, web pages, reports, and official documents) in order

to extract the data, transform it into information and make it useful for various types of decision making. Text mining encompasses linguistic, statistical, and machine learning techniques, which can be used, in its final stage for analysis, visualization (via maps, charts, mind maps), for integration with structured data in databases or warehouses, for machine learning, etc. Although conducted on unstructured text, one of the first tasks is to organize it and structure in the way suitable for further qualitative and quantitative analysis. Text mining extracts relevant words (N-grams) and relationships between them in order to categorize them and make conclusions relevant to a business problem or scientific inquiry. In other words, the goal of text mining is the extraction of knowledge and patterns from various text documents (Zhai, Velivelli, and Yu, 2004) [4]. Some of usual text mining undertakings include classification and clustering of phrases or topics, named entity recognition, information extraction (Yehia, Ibrahim, and Abulkhair, 2016) [5], sentiment analysis (Schumaker, Zhang, Huang, and Chen, 2012 [6]; Nakayama et al., 2018 [7]), keyword extraction, natural language processing (NLP) (Ong, Chen, Sung, and Zhu, 2005 [8]; Klopotan, Zoroja and Meško, 2018 [9]) including tagging, parsing, topic detection, etc. Herráez, Bustamante and Saura (2017) [10] used text mining in order to extract topics using content analysis of e-commerce organizations, while Reyes-Menendez et al. (2018) [11] used text mining in order to extract topics from social media and to classify them according to sentiments.

Increased interest has been paid to multilingual text mining in order to get insight into information across languages.

Text mining has gained its popularity with big data resources when analyzing big data in the financial sector. This way, financial organizations can identify valuable information from customer opinions, corporate documents, and posts on social networks, e-mails, call logs, detect customer churn, fraud or risks, etc. In order to highlight the various aspects of the use of textual mining in banking and finance, this study aims to answer three research questions: (i) What is the intellectual core of the field? (ii) Which text mining techniques are used in the financial sector for textual mining, especially in the era of the Internet, big data, and social media? (iii) Which data sources are the most often used for text mining in the financial sector, and for which purposes?

In order to answer these questions, a qualitative literature analysis is conducted using a systematic literature review, citation, and co-citation analysis. These methodologies allow mapping and analysis of the evolution of the scientific field (Batistič, Černe, and Vogel, 2017) [12]. Besides, in order to consider second and third research question, the paper provides an overview typical text mining techniques used in the financial sector and analyses them according to the type of data sources used, as well as according to their typical business applications.

This paper aims to contribute to both theory and practice. Through the citation and co-citation analysis and the answers to the first and second research questions, primary theoretical contributions reflect in summing up the conclusions and research trends of the field. In addition, the second and third research question offers practical contributions through a summarized overview of the presenting relevant text mining techniques according to data sources used and typical applications.

The paper is structured as follows. The introductory part examines the impact of big data analysis on the financial sector with an emphasis on text data analysis. In the second chapter, a survey of similar literature reviews focusing on data mining and text mining applications in finance has been presented, which is used for developing research questions. The third chapter presents the methodology used for conducting the research, and the steps of the research process were presented. The third part presents the results of citation and co-citation analysis. The fourth chapter presents various text mining techniques used in finance. The fifth chapter provides the analysis of data sources used and typical applications for text mining in finance. Finally, conclusions are given, and further research is proposed.

2. Research Questions

Since the emergence of data mining, as the advanced data manipulation, processing, and modeling approach, the interest in its usage in finance has grown exponentially, which generated the need for

literature reviews that could provide a focused outline to advantages and disadvantages of data mining utilization in finance for the researchers and practitioners (Zhang and Zhou, 2004) [13]. Numerous literature reviews were conducting focusing to different aspects of data mining applications in various fields of finance, such as stock markets predictions (Hajizadeh, Ardakani and Shahrabi, 2010) [14], financial fraud detection (Ngai et al., 2011) [15], and financial risk analysis (Jin, Wang and Zeng, 2018) [16].

Text documents that are an abundant and dominant source of relevant information in the business domain are unstructured, which is an obstacle in the fast processing of the information stored in them. Therefore, text mining as the automated approach to the analysis of various text documents emerged as the attempt to make use of text-based unstructured information. In spite of the importance of text mining for finance, only recently literature surveys investigated the utilization of text mining for financial applications. Some of the literature reviews only sporadically included text mining applications. Ngai et al. (2011) [15] developed a classification framework for analyzing data mining applications in financial fraud detection, such as classification, clustering, visualization and outlier detection, regression, and prediction. They identified one text mining application in their analysis, using Naïve Bayes text mining in order to employees likely to conduct fraud (Holton, 2009) [17]. Gray and Debreceny (2014) [18] designed a research taxonomy for fraud detection in financial statement analysis, identifying the following usage of text mining in this area: deception analysis, as a relevant tool in fraud detection.

Review papers that focus solely on text mining rarely investigate financial applications. Sun, Luo and Chen (2017) [19] developed a review of natural language processing applications for text mining in order to extract opinions. In their review, they focused on various approaches, such as comparative opinion mining and deep learning. Nassirtoussi et al. (2014) [20] developed a theoretical and practical review of applications of text mining for market prediction, which focuses mainly to online sentiment analysis using social media and news texts, and their utilization for prediction of FOREX market and stock exchange markets.

Kumar and Ravi (2016) [21] conducted a literature review of text mining applications in finance. They surveyed papers published from 2000 to 2016 and developed the following groups of text mining applications: FOREX and stock market forecasts, customer relationship management applications, as well as security applications, focusing to cybersecurity. In their analysis, they focused on text mining algorithms, such as decision trees, neural networks, linear regression and logistic regression. In their work, they do not provide citation and co-citation analysis. Systematic analysis of text sources utilized for text mining is not provided in their research.

Based on this review of similar research, the following gaps are identified. First, due to the exponential growth of text mining utilization in the field of finance, there is a growing need to provide an up-to-date review, tracking the cutting-edge state-of-the-art research. Therefore, the first research question has been outlined as: What is the intellectual core of the field? aiming to provide the answer with the longer research period (from 2000 to 2019), and using citation and co-citation analysis. Second, the aim of the paper is to detect the most used text mining techniques, taking into account the most recent advances in the field., such as big data. Therefore, the second research question is posed as Which text mining techniques are used in the financial sector for textual mining, especially in the era of the Internet, big data and social media? Finally, in order to capture the venues for the future research as well as to provide the practitioners with the outlook on how to use text sources available to them online, and in their organizations, the third question is posed as: Which data sources are the most often used for text mining in the financial sector, and for which purposes?

3. Methodology

In order to provide the answer to the research questions of this study, multiple research methodologies, such as bibliometric techniques co-citation and citation analysis, and systematic literature review (SLR). SLR is outperforming informal literature review "with respect to the

planning for literature review, the design of search string, sources to be searched, publication inclusion and exclusion criteria, publication quality assessment and the data extraction process" (Niazi, 2015, p. 845) [22], since SLR refers to "identifying, assessing, and interpreting available research studies with the purpose to provide answers to the research question" (Wahono, 2015, p. 1) [23]. Following the recommendations of Wahono (2015) [23] and Wang et al. (2016) [24], the research steps were created and used in our SLR analysis (Figure 1).

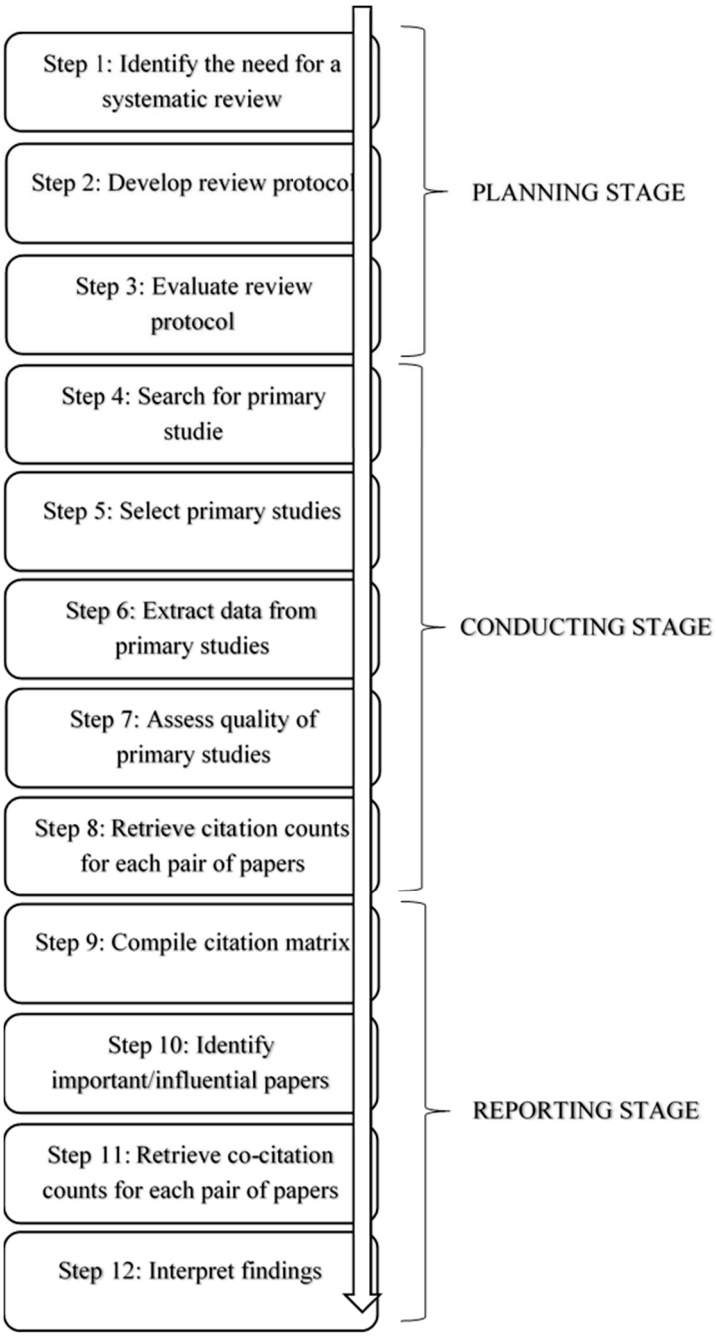

Figure 1. Research steps used in this study.

First, based on the analysis of previous literature reviews of text mining and data mining in finance (as presented above), the need for a systematic review in that area has been identified (Phase 1). The review protocol (Phase 2) and the evaluation of the review protocol (Phase 3) are related to the search strategy and study selection process (Wahono, 2015) [23]. The review protocol

is evaluated in relation to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard for writing SLRs (Moher et al., 2006) [25]. Following the established bibliometric protocol, the data for the research is acquired by searching publications using the search string (*"text mining" AND finan**) to be found in all fields, for the period from 2004 (Phase 4). The search was conducted in October 2018, in the Social Science Citation Index (SSCI), Science Citation Index Expanded (SCI-EXPANDED) and Emerging Sources Citation Index (ESCI) databases, and it generated 345 items. After a manual analysis of the literature, the final list of 123 publications was used for this analysis (Phase 5). This phase has been presented in detail in Figure 2. Data were extracted from primary studies by manual analysis and extraction of the most used text-mining techniques (Phase 6 and Phase 7). Bibexcel and Pajek software are employed to identify important papers and conduct co-citation analysis of papers indexed in Web of Science (Batistič, Černe and Vogel, 2017) [12], for the purpose of conducting Phase 8 and Phase 9. Citation and co-citation analysis were used in order to detect the most relevant work in the field (Phase 10, and Phase 11). The most cited papers were discussed in relation to the identified text mining techniques in the financial sector (Phase 12).

Figure 2 presents the PRISMA flow diagram, outlining the process of development of the final list of papers included in the analysis, following the practice used in numerous systematic literature research, such as Saura et al. (2017) [26]. Initially, 345 papers were identified by searching Web of Knowledge with the search term *"text mining" AND financ**. In order to select papers that focus to the utilization of text mining to finance, titles and abstracts of 345 initially extracted papers were examined. Among them, 186 papers were not related to the topic, e.g., used phrase text mining in a different context or mention the term finance sporadically—this approach resulted in 159 potentially adequate papers. Authors have read the full text of these papers and excluded additional 59 papers that are not related or are vaguely related to the research topic, or which provide an only shallow description of text mining approach. Additional 23 papers were tracked by snowballing approach, using references of papers. Therefore, the final list of 123 papers is developed, which is the focus of our analysis. Rest of the paper focuses on these 123 papers that are focused on the use of text mining in various applications. Appendix A provides the list of papers included in the literature review.

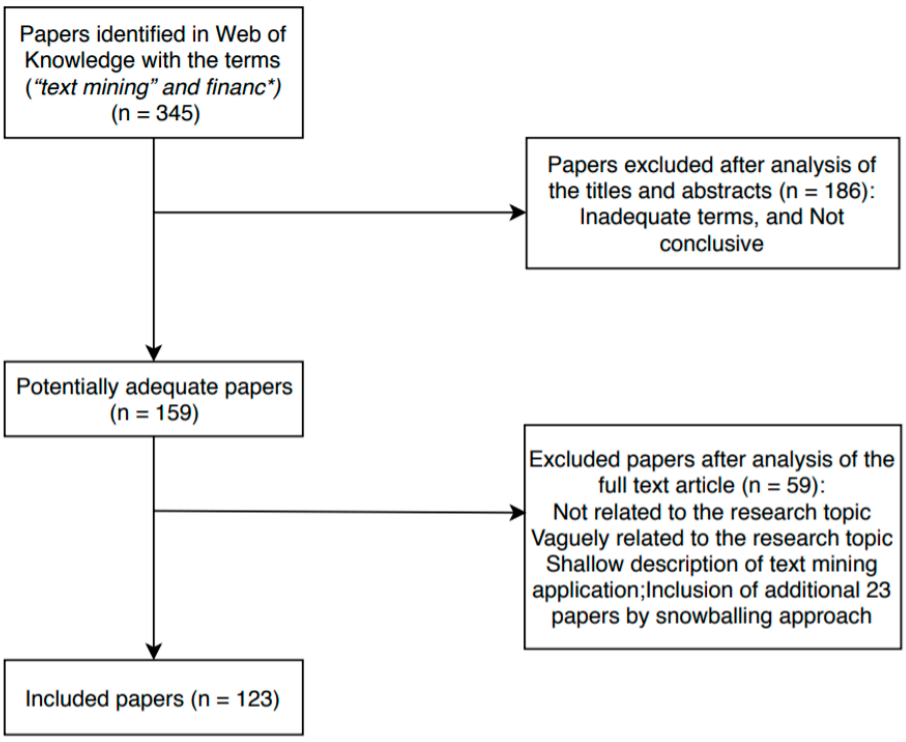

Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.

4. Citation and Co-Citation Analysis

RQ1: *What is the intellectual core of the field?*

Throughout the whole period, an exploratory focus is evident, i.e., financial forecasting or prediction of financial markets. A number of papers were initially low. In the period from 2001 to 2009 less than 5 papers were published per year, while in the period from 2011 to 2014 less than 10 papers were published per year. However, the number of papers increased to 15 in 2015, followed by an exponential increase in the following years (14 papers in 2016, 18 papers in 2017, 27 papers in 2018). It can be presumed that this increase is the result of the overall interest of the financial community in various data analysis approaches, which generated the creation of diverse solutions covered by FinTech umbrella (Arner, Barberis and Buckley, 2015) [27]. Figure 3 presents the number of citations on this topic over the examined period with the continuous increase in the number of published studies.

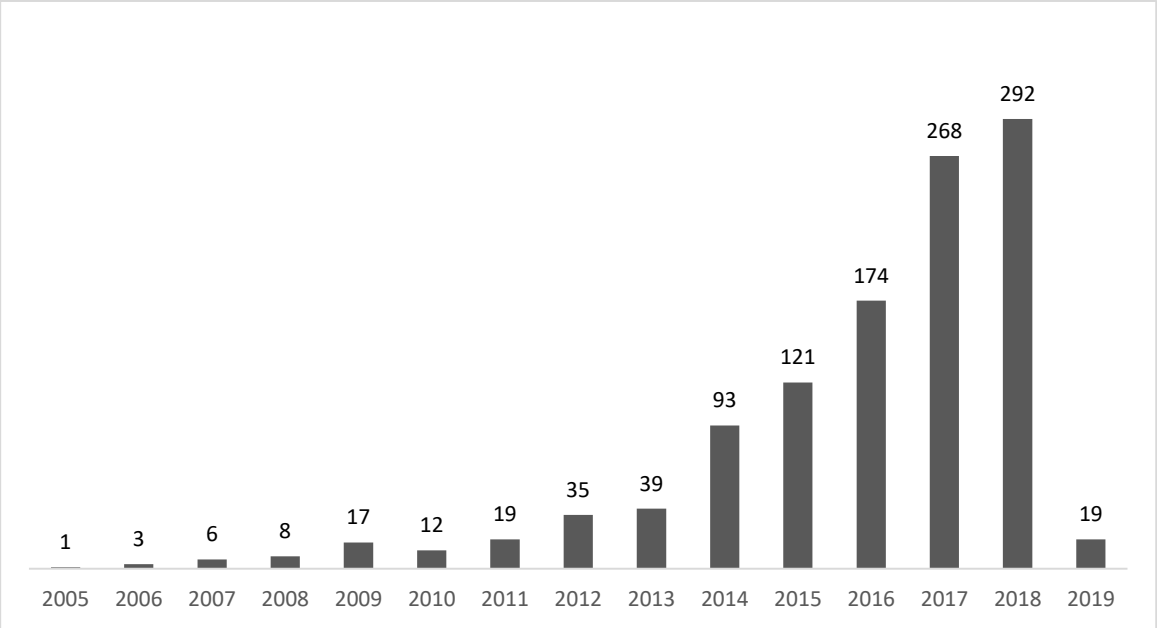

Figure 3. Total number of citations in the field on the Web of Science (Social Science Citation Index (SSCI), Science Citation Index Expanded (SCI-EXPANDED), and Emerging Sources Citation Index (ESCI).

A citation network of all papers has been created in order to identify the most important documents in the network, and therefore the trends of the field research. First, a citation analysis of identified papers has been conducted. Citing occurs when one study refers to other papers as source ones, and the citation analysis provides information on papers and their sources and provides the insight into the overall citations number (Wang et al., 2016) [24]. Co-citations are considered as citing two papers together in different paper. The co-citation analysis identifies the sources of the search in the observed academic area, by the detection of the most relevant papers, as well as generating indicators of the paper impact, since the number of citations are co-related with the relevance of the work in the particular academic community (Batistič, Černe, and Vogel, 2017) [12]. The observed number of citations of papers is 1107, while the average citations number is 9.

Table 1 lists the most cited papers (>50 citations). Based on the analysis, it can be concluded that most cited studies are Nassirtoussi, Wah, Aghabozorgi, and Ngo (2014) [20], Humpherys, Moffitt, Burns, Burgoon, and Felix (2011) [28], Schumaker, Zhang, Huang, and Chen (2012) [6], Vemuri (2017) [29], Ong, Chen, Sung, and Zhu (2005) [8], and (Coussement and Van den Poel, 2008) [30].

The research focus of these studies is presented in Table 1.

Table 1. Citation analysis of the most cited papers in the Web of Science.

Title	Authors, Publication Year	Total Citations	Objectives
Text mining for market prediction: A systematic review	Nassirtoussi, Wah, Aghabozorgi, and Ngo (2014) [20]	97	The study presents an overview of studies related to the market forecast based on online text mining and creates an outlook to the main elements. In addition, the paper presented a scomparison of all systems with the identification of the main differentiating factors.
Identification of fraudulent financial statements using linguistic credibility analysis	Humpherys, Moffitt, Burns, Burgoon, and Felix (2011) [28]	73	The study analyses corporate fraud detection "through a unique application of existing text-mining methods on the Management's Discussion and Analysis and tests for linguistic differences between fraudulent and non-fraudulent MD & As" (Humpherys, Moffitt, Burns, Burgoon, and Felix, 2011, p. 585).
Evaluating sentiment in financial news articles	Schumaker et al. (2012) [6]	71	The study deals with the choice of words and tones used in newspaper articles and their relation to the movements in stock prices.
Newsmap: a knowledge map for online news	Ong et al. (2005) [8]	62	This study focuses on the automatic generation of a NewsMap knowledge hierarchy map, based on online news, especially in finance and healthcare.
Integrating the voice of customers through call center emails into a decision support system for chum prediction	(Coussement and Van den Poel, 2008) [30]	51	The study analyzes the effect of adding text information to the churn prediction system that uses only traditional marketing information.

On the other hand, "the importance of a paper can be determined by its influence in the citation network, which can be measured by two indexes, degree centrality, and betweenness centrality. Degree centrality is measured as the number of direct ties that a node in the network has, while betweenness centrality implies the extent to which one node exists on the shortest path between other nodes" (Wang et al., 2016, p. 35) [24]. In this regard, the co-citation analysis has been conducted, and the co-citation network plotted in Figure 4 was created.

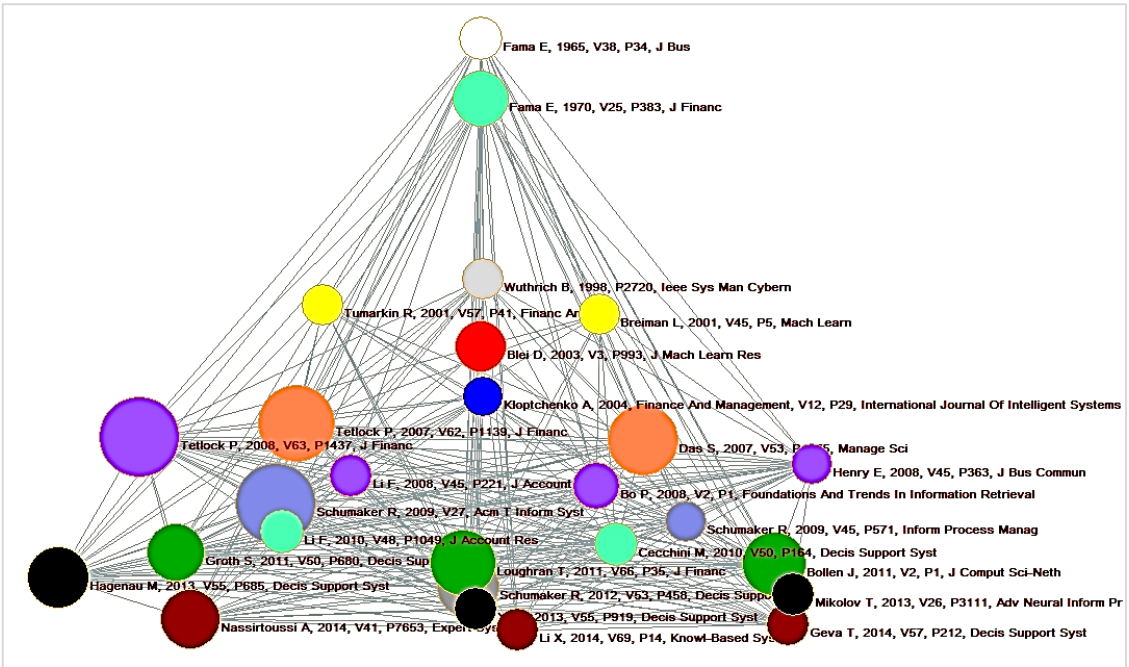

Figure 4. Co-citation network presenting historical evolution of the field.

Following the usual practice, the density of the network is calculated, which outlines the connections between the network nodes. If density is above 0.5, it is considered as high (Abrahamson and Rosenkopf, 1997) [31]. In this research, the network density is 0.754 that indicates that the links between the papers are quite abundant. According to degree centrality and betweenness

centrality measures, the most influential papers are those published by Schumaker and Chen (2009) [32], Schumaker et al. (2012) [6], Tetlock (2007) [33], Loughran and McDonald (2010) [34], Bollen, Mao, and Zeng (2011) [35], Pang and Lee (2008) [36], and Hagenau, Liebmann, and Neumann (2013) [37]. They will be briefly discussed in order to detect the research trends in the field.

In addition to most cited studies, studies with the highest degrees centrality deal with similar topics. Thus, Schumaker and Chen (2009, p. 1) [32] "examine a predictive machine learning approach for financial news articles analysis using words and phrases representations." Tetlock (2007, p. 1139) [33] analyses the relationship "between the media and the stock market using daily content from the Wall Street Journal column", while Loughran and McDonald (2010, p. 36) [34] develop a negative word list that "reflects tone in financial text and link the lists to returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings". Furthermore, Bollen et al. (2011) [35] analyze the text content of daily Twitter feeds and assess its utilization for stock-market predictions. It is similar to Hagenau et al. (2013) [37] who examine if financial news could be used for the prediction of stock prices. These studies and authors represent the field's intellectual core and can be considered the most influential in the field. Besides, the bibliographic co-citation analysis points to two studies as a source of the field dealing with the portfolio analysis model for a stable Paretian market and with the theoretical and empirical literature on the efficient market models. In other words, intellectual core papers are related to financial markets rather than to text mining. The oldest cited paper dealing with text mining of web data was published in 1998 with the aim of presenting techniques for the exploitation of textual financial news and analysis results (Wuthrich et al., 1998) [38]. Then, Tumarkin and Whitelaw (2013, p. 41) [39] analyzed the "relationship between Internet message board activity and abnormal stock returns and trading volume." After 2003, there is a growing interest in this research topic.

5. Text Mining Techniques in Finance

RQ2: *Which text mining techniques are used in the financial sector for textual mining, especially in the era of the Internet, big data and social media?*

To address the second question, a comprehensive search has been conducted using the same string, but in other databases. In order to narrow the research, and given the importance of big data for the information era, this research is focused on big data environments rather than simple text mining technologies. Big data denotes the vast and complex amount of data in structured documents, semi-structured documents (documents having structure, but differentiating among themselves, e.g., XML documents, HTML files) and unstructured type of documents (in terms of record layout, embedded metadata). Mathew (2012) [40] points out several issues in big data analytics: diversification of data types along with the vast amount of data, more changes and uncertainty, more unanticipated questions, and real-time needs and decision-making.

Many companies and institutions have a large amount of data, which can be exploited and used to extract data and create information linked to new knowledge. In this research process, text mining can help in customer analytics, marketing opportunities, for fraud prevention, to improve operational activities and to develop new business models. When it comes to financial institutions, two primary data sources can be used for text mining: external and internal data sources. Internal data can be transaction data, log data, application data. External data can be from any social media and website. In the digital era, big data is generated from many increased sources, including online clicks, mobile transactions, social media, and data generated through sensor networks. (Yehia et al., 2016) [5]. While financial institutions already have some inputs from their customers in some form, benefits of big data technologies are to compare institution with competitors with same objective metric (usually machine learning generated metric). In this sense, text mining can be integrated into the business intelligence process and business applications, which represents the promising target. Fan et al. (2006) [41] suggest future development in the integration of data and text mining used to

discover hidden information in various types of documents collected from various resources, with the purpose of improved decision-making solutions.

In order to classify the text mining techniques which are the most relevant for big data analytics in the financial sector, expert panel approach was used in order to provide the classification of techniques that reflect the usage of text mining in practice. Four experts were selected from the big data specialized tech-companies that deliver the service to various financial organizations (e.g., banks, insurance companies, and stock-market exchanges). This approach is following the recent practice that proposes the inclusion of from-the-field experts in order to gain results which are tightly related to practical usage of techniques on a day-to-day basis (Best et al., 2009 [42]; Bannan-Ritland, 2003 [43]). Experts have selected the following text-mining techniques used in big data analytics as the most relevant for financial sector: keyword extraction, named entity recognition, gender prediction, sentiment analysis, topic extraction, and social network analysis. In the rest of the chapter, the above-mentioned techniques are presented, and for each technique, selected studies dealing with this technique are presented.

5.1. Keyword Extraction

With new technologies and analysis in recent times and especially in the case of big data analytics with vast volumes of new data coming from different sources, there is a need for keyword extraction. Table 2 presents the selected papers dealing with the "keyword extraction" technique in the financial sector.

Table 2. Selected papers dealing with the "keyword extraction" technique in the financial sector.

Authors	Research
Hasan and Ng (2014) [44]	Automatic keyword extraction
Roh, Jeong, and Yoon (2017) [45]	Multilayered keyword extraction methodology for structuring technological information through natural language processing (NLP), h purpose: discovering trends in patent analysis, technology classification or knowledge flow among technologies
Eler et al. (2018) [46]	Pre-processing steps with impact on text mining techniques: lowercasing, deletions, stemming/ lemmatization, PoS (Part-of-Speech) tagging, parsing

Keyword extraction plays a key role in text mining financial applications. The simple form is when a list of keywords is needed in order to extract related comments and articles from an external source. More complex, but sophisticated usage, would be to use automatic keyword extraction (Hasan and Ng, 2014) [44]. This field gains vast interest in the past several years since volumes of data are growing and every document or comment cannot be read sequentially. The goal is to extract "sequence of words," called N-grams, through a semi-automated process. However, this process does require manual validation and comparison with the reference model, i.e., "gold standard" in order to assess the quality of the tool. Quality of terminology has gained importance regarding costs, user perceptions, customer satisfaction.

To summarize keyword extraction has 4 approaches (Bharti et al., 2017) [47]: statistical (term frequency, inverse document frequency), linguistic (WordNet, n-Gram, PoS (Part-of-Speech) patterns), machine learning (Naïve Bayes) and hybrid approach (some combination of previous three approaches.

In most of the cases, pre-processing techniques are needed, starting from the corpus collection, where documents can be collected from one or more sources, depending on various criteria, including corpus size and domain. A possible step could be deduplication of documents or articles, preformatting, possibly scanned and converted by OCR (optical character recognition). Once having a text in the appropriate format, the text is tokenized by which text is characterized as a list of words, numbers, signs, and punctuation and treated as "bag of words." Pre-processing steps through various methods have a substantial impact on text mining techniques (Eler et al., 2018) [46]. Through lowercasing the whole text (all tokens) are converted into lowercase, where some mistakes can happen (e.g., converting

of abbreviation US into us as a pronoun). To reduce noise in the text, there are various techniques like deletion of double spaces, numbers, names (if needed), punctuation, rare words, and stop-words The next step in reducing dimensionality is the introduction of stemming or lemmatization tasks on keywords in order to gather all variations of specific keywords (example: bank, banking, banks -> bank). Lemmatization uses PoS (Part-of-Speech) tagging to identify grammatical categories. This feature can be useful in the parsing algorithms to detect the correct POS word or to extract the sequence of words (N-grams). Many text mining tools use stemming which uses cutting of affixes (banking) -> bank+ing. This feature is useful in later use and especially for mention counting or online presence metric. This metric data practically counts how many specific keyword names exist for a particular page name or username. This online presence metric can be used for institution comparison in the financial sector (ex. bank1 v bank2).

Text mining can also be used in discovering trends in patent analysis, technology classification or knowledge flow among technologies as in Roh, Jeong, and Yoon (2017) [45]. They proposed multilayered keyword extraction methodology for structuring technological information through NLp. As pure keyword, extraction has deficiencies such as omitting meaningful keywords, they suggested to "meaningful keyword sets related to technological information. Firstly, they analyzed the characteristics of technological information" (Roh, Jeong, and Yoon, 2017, p. 1) [45], structured it by information type and then performed keyword extractions in each type through NLp.

5.2. Named Entity Recognition

Named entity recognition represents one of the key phases in text mining (Saju and Shaja, 2017) [48] used on large corpora of data, which can be used in information retrieval and extraction and further in NLP, machine translation and question–answering system, speech recognition, natural language generation, chatbots conversation, machine learning, document indexing, image recognition, etc. Many industries use named entity recognition on big data sets. Most of named entity recognition techniques use methods of machine learning, which requires large amounts of data in order to train a good classifying algorithm. Table 3 presents selected papers dealing with the "named entity recognition" technique.

Table 3. Selected papers dealing with the "named entity recognition" technique.

Authors	Research
Alvarado, Verspoor, and Baldwin (2015) [49]	Named entity recognition analysis on financial documentation and publicly available non-financial data set to extract information of risk assessment
Ritter et al. (2011) [50]	Supervised approach for named entity recognition

Named entity recognition is a process, which labels a Name—i.e., a sequence of words in documents, which denote email, amounts-currency, company/bank/institution name, brand name, city-state name, time, or others (Grishman and Sundheim, 1996) [51]. The three universally accepted name entities are a person, location, and organization. Named entity recognition consists mainly of two steps: detection of names in the text and classification by the type of entity, but also discovering relationships among entities. In the detection process, problems of segmentation can appear (e.g., National Bank of Croatia which is a single name, instead Croatia being a location), followed by classification, depending on annotated corpora. Named entity recognition could have its business value in industrial applications, as in bank transaction details, to detect contracts, e-mails, machine translation, question answering, spell checking, etc. Alvarado, Verspoor, and Baldwin (2015) [49] conducted named entity recognition analysis on financial documentation and publicly available non-financial data set to extract information of risk assessment.

5.3. Gender Prediction

Information about gender is often useful, especially when the emphasis of analysis is marketing planning and/or better understanding of customers. Table 4 presents selected papers dealing with the "gender" prediction technique in the financial sector.

Table 4. Selected papers dealing with the "gender prediction" technique.

Authors	Research
Phuong and Phuong (2014) [52]	Users' gender based on browsing history, important for marketing and personalization
Kucukyilmaz et al. (2006) [53]	Gender prediction in computer-mediated-communication/ chatbots
Lotto (2018) [54]	Gender prediction to predict financial inclusion, compared with traditional banking services

The simple approach to solving this problem is to make a dictionary of female and male names and then match that dictionary with usernames. This can be the right approach if fast results are needed. Still, when social media and websites are analyzed, the problem arises related to the number of accounts from different organizations, bots, and fake accounts with random names. In that case, the presented approach will only recognize what is in the dictionary, thus lowering the probability of recognizing the gender of the customer.

To solve this limitation next step would be to use natural language processing models such as "bag of words" and n-grams or a combination of both. This approach analyses word usage and differences between them and the difference between styles. Disadvantage again occurs in case of the data extracted from social media. Features used for this classification task are (Zhang and Zhang, 2010) [55]: words (authors suggest that binary representation is more effective—word exist or not in document), average word or sentence length, POS tags (noun, verb, adjective, and adverb), word factor analysis—finding groups of similar work (there are 20 lists—example of conversation list is known, care, friend, saying). Information gain can be used as feature selection and with SVM as a classifier, with the accuracy above 72%.

Friedmann and Lowengart (2016) [56] conducted an analysis to explain gender differences when choosing banking services. Galli and Rossi (2014) [57] performed research on gender in the credit market for 7 European countries in the period of financial crisis. Other authors used gender prediction using various text mining sources, such as browsing history (Phuong and Phuong, 2014) [52], and chatbots (Kucukyilmaz et al., 2006) [53].

5.4. Sentiment Analysis

Sentiment analysis or opinion analysis is used in the financial sector to identify the "voice of customers." Table 5 presents selected papers dealing with the "sentiment analysis" technique in the financial sector.

Table 5. Selected papers dealing with the "sentiment analysis" technique in the financial sector.

Authors	Research
Pang and Lee (2008) [36]	Sentiment analysis for determination of writer's attitude towards the specific topic
Nopp and Hanbury (2015) [58]	Sentiment analysis to detect risks in the banking system
Narayanan et al. (2013) [59]	Algorithms with correct feature selection and noise removal process

Sentiment analysis (Pang and Lee, 2008) [36] refers to text analysis or natural language processing techniques, which helps the determination of a writer's attitude towards a specific topic. Usage of sentiment analysis is frequent in the financial domain. Nopp and Hanbury (2015) [58] used sentiment analysis to detect risks in the banking system. Srivastava and Gopalkrishnan (2015) [60] analyzed

sentiments for the banking sector in order to assess the functioning of the bank. These narratives are created and disseminated in social interaction.

There are several approaches to build an accurate sentiment model. Some approaches address this problem from natural language processing view, other from machine learning view or, in current years, more specifically, as a deep learning problem. The first approach, based on natural language processing, is to build a dictionary of known negative and positive words. For this task, only extreme polarities and word that can be correctly associated with the polarity are needed. Based on the developed dictionary, the sentiment is calculated by a simple count of words found in a specific document from our dictionaries. Polarity with more discovered words "Wins" and text is then classified. The next approach, based on machine learning, is about creating a large data set, containing documents that are first classified manually (by a human). Based on the classification, the machine-learning model can be developed, that can provide the rules for automated classification. A problem can be addressed as the classification of two classes (positive or negative) or more (e.g., range from 1–5 for sentiment intensity). Features can be unigrams, bigrams, or a combination of both (Go et al., 2009) [61]. Document term matrix is built, based on our features and values in this matrix, which can be either frequency like "TF (Term Frequency), TF-IDF (Term Frequency-Inverse Document frequency), or binary representation" (Hussin, 2004, p. 158) [62]. In the big data architectures, the machine-learning model can be used on batch data but also in real-time data in order to perform real-time classification. Accuracy can be greater than 80% even with simple algorithms with correct feature selection and remove noises from the data (Narayanan et al., 2013) [59].

The last approach, based on deep learning, the sentiment analysis would be performed using word embeddings, such as word2vec, GloVe (Zhang et al., 2018) [63]. Word embeddings are used to represent words as vectors. With this technique, similar words can be mapped to nearby points in continuous vector space. Deep learning is an improvement from other approaches and especially in sentiment classification of relatively small documents (tweets, comments).

5.5. Topic Extraction

Topic modeling or topic prediction/ extraction is based on the number and distribution of terms across documents by counting the probability of belonging to a certain topic. Table 6 presents selected papers dealing with the "topic extraction" technique in the financial sector.

Table 6. Selected papers dealing with the "topic extraction" technique in the financial sector.

Authors	Research
Moro et al. (2015) [64]	Topic detection of a large number of manuscripts using text mining techniques when detecting terms belonging to business intelligence and banking domains (dirlecht allocation model), topics: credit banking, risk, fraud detection, credit approval and bankruptcy
Zhao et al. (2011) [65]	Social media as a source of entity-oriented topics, unsupervised machine learning approach
Lee and So Young (2017) [66]	Framework to identify the rise and fall of emerging topics in the financial industry

Moro et al. (2015, p. 1314) [64] performed topic detection of "a large number of manuscripts using text-mining techniques when detecting terms belonging to business intelligence and banking domains". They used latent dirlecht allocation model to detect topics, by using a dictionary of terms in order to detect topics and research directions. They grouped articles into several relevant topics, followed by dictionary analysis to identify relations between terms and topics of grouping articles. This research showed that credit banking was the main trend, with topics of risk, fraud detection, credit approval, and bankruptcy. By this approach, the probability of each document to belong to a certain topic could be estimated. In this way, it is possible to identify topics capturing more attention.

Data from social media can be used to find discussed topics at a certain time. Previous research indicates that these data "can be a good source of entity-oriented topics that have low coverage in traditional media news" (Zhao et al., 2011, p. 46) [65]. Input in the model should be a matrix of document-terms format with TF-IDF frequencies as values or binary representation (0 or 1). A common approach for topic extraction is unsupervised machine learning approach.

New approaches also take deep learning techniques for topic extraction. Popular word embedding, in this case, is lda2vec, which is a modification of word2vec presented in sentiment analysis (Moody, 2016) [67]. Lda2vec uses word2vec principles and expands this to word, document and topic vectors. Topic extraction helps us answer the question "WHAT" is talked about for example the institution or competitors. Usually, topics are represented as word clouds, but they can be visualized by some more complex graphical representation (LDAvis—Intertopic distance map is visualized with PCA). Lee and So Young (2017) [66] proposed a framework to identify raise and fall of emerging topics in financial industry using abstracts of financial business model patents, in order to discover topics from documents, aiming to enable understanding of the changing trends of financial business models over time.

5.6. Social Network Analysis

Social network analysis is the process that is based on graph theory and used for a better understanding of social structures. When it comes to SNA, structure refers to nodes and edges. For example, in the case of Twitter, each node would be one Titter user, and each edge is a relationship between two users (the user is connected to another by the user using a retweet). Table 7 presents selected papers dealing with the "social network analysis" technique.

Table 7. Selected papers dealing with the "social network analysis" technique.

Authors	Research
Ediger et al. (2010) [68]	Metrics for social network analysis: centrality measures, node degrees (used to find users who are highly connected), closeness (goal is to find users who can spread information to others), clustering coefficient, PageRank
L'Huillier et al. (2011) [69]	Integration of social network analysis and topic detection
Mao, Jin, and Zhu (2015) [70]	Social network analysis to explore the way that bank customers impact each other

Usual metrics calculated with SNA techniques are (Ediger et al., 2010) [68]: centrality measures, node degrees (used to find users who are highly connected), closeness (goal is to find users who can spread information to others), clustering coefficient, PageRank. Social network analysis is a different type of analysis in comparison to text analysis, but it is used here to show how text analysis and its result can be integrated with this analysis (L'Huillier et al., 2011) [69]. For example, when identifying users who can easily spread a message to a network of interests (with SNA techniques), textual information from the followers of that user can be used to discover common interests. This information can be used for marketing campaigns to generate the best keywords. Mao, Jin, and Zhu (2015) [70] used SNA to explore the way that bank customers impact each other in order to detect the most influential customers.

6. Data Sources Used and Typical Applications of Text Mining in Finance

RQ3: *Which data sources are the most often used for text mining in the financial sector, and for which purposes?*

Financial and banking institutions, being in a competitive environment, seek new ways to reach customers. Presented papers indicate that text mining represents the hidden door for discovering information in a pile of unstructured data collected from various sources. Text analytics aims to discover key points that could lead to new decisions, such as "who," "where," "when," "why," and "how" which could bring new decisions. Some examples of use include customer analytics possibly derived from data acquired from social media and informal conversations, aiming to detect customers,

enhance their engagement, or offer specific services, or to develop new business models based on detection of a preferred way of communication. Annual reports, e-mails, external data coming from sensors, transactions or free-form text can be used to enhance services or for risk and fraud detection. The findings are summarized in relation to the type of source used for text mining. As discussed, financial institutions used two primary data sources for text mining: external and internal data sources.

Table 8 presents the most important text mining techniques according to the type of data source: internal or external. In addition, example sources are outlined together with the example applications.

Table 8. The most often used data sources for text mining in the financial sector.

Text-Mining Technique	Internal Data	Example of Internal Data	External Data	Example of External Data	Example Application
Keyword extraction			✓	News, social media feeds, patents	Fraud detection and Stock market prediction
Named entity recognition			✓	Publicly available non-financial data, e.g., social media	Customer relationship management
Gender prediction			✓	Publicly available non-financial data, e.g., social media	Customer relationship management
Sentiment analysis			✓	News; social media feeds; Financial statements	Fraud detection and Stock price prediction
Topic extraction	✓	Legal documents	✓	News, social media feeds	Summarization

It can be noted that in most of the studies examined, the authors used external data. The reasons for this can be twofold. First, authors would prefer to use the external data since they are public and free to use, while the internal data are the ownership of the company and numerous restrictions can apply in using them as a data source for text mining. For example, the nature of financial data used for fraud detection is very sensitive. Financial organizations are reluctant in sharing that information. Papers dealing with fraud detection are usually based on small datasets of just a few hundred samples which is not enough for text mining to extract information that is useful in the real-world situation. Still, these small datasets provide us glimpse into the world of financial fraud and can help us derive a way to text mining usage. Second, the financial sector may be more prone to use external data, and the use of internal data for various purposes is still rare in practice.

Authors use in most of the cases news, social media feeds, patents, and financial statements, as the external sources for text mining analysis. Only for one of the text mining techniques, internal sources are used, specifically legal documents.

The following example applications emerged: (i) customer relationship management; (ii) fraud detection; (iii) stock price prediction; and (iv) summarization. Two of these applications were also the research focuses on the most influential papers in the field, which gathered the largest number of citations: stock market prediction (Schumaker and Chen, 2009 [32]; Tetlock, 2007 [33]; Bollen et al., 2011 [35]; Hagenau et al., 2013 [37]), and fraud (Loughran and McDonald, 2010 [34]; Humpherys et al., 2011 [28]).

Fraud detection has become a significant concern for financial organizations. Several text mining approaches have been developed mostly for large amounts of financial statements. Fraudulent activity can take many forms like money-laundering, insurance fraud, piracy (software), identity theft, and embezzlement and so on. Usually, fraud detection is conducted in the financial sector using quantitative data. For example, different important features can be found in those text files that can help fraud detection (Chye Koh and Kee Low, 2004, p. 463) [71] like "quick assets to current

liabilities, market value of equity to total assets, total liabilities to total assets, interest payments to earnings before interest and tax, net income to total assets, and retained earnings to total assets". Challenges of fraud detection in the financial industry are: typical classification problems like feature selection, model optimization and problem domain, the imbalance between fraud types and detection method studied, privacy issues, computational issues like the computational performance of models in real-time systems, fraudster new and innovative ways of making fraud that need to be yet studied. Loughran and McDonald (2010) [34] focused to fraud and unexpected earnings, and Humpherys et al. (2011, p. 585) [28] conducted the "identification of fraudulent financial statements using linguistic credibility analysis". Glancy et al. (2011) [72] present the process of detecting fraud using management statements. The financial statements and their text are like any textual data, unstructured and the goal of text-mining is to give the structure to that data set in order to extract information and knowledge. When it comes to fraud detection, first and one of the most important tasks is to create a larger dataset for training, if possible. This dataset needs to have both fraudulent and non-fraudulent statements from various organizations of different sizes. Quality of this initial step is affecting every other step in the text mining process. Next step is to clean textual data and perform pre-processing steps similar to sentiment analysis case. Standardization of text and structure creation is key in this step. After that, text mining can be used in order to extract characteristics that can help the detection of fraudulent behavior. Tree algorithm and SVM are popular in this detection. After model evaluation and model selection, implementation of a model into the system or "into the wild" to detect new fraudulent statements is needed. Fraud detection is among the most difficult text mining techniques (some fraud types have higher success like credit card transaction fraud). Systemic Functional Linguistics theory can also be used for fraud detection (Dong et al., 2016) [73]. This approach is all about to feature creation and text classification. Authors proposed feature set that can be used as they stated to achieve above baseline accuracy. Their example is a presentation of new information to investors so they can bring better decisions, to auditors to recognize fraud risk, to regulators to investigate only suspicious behavior and firms. Some of the features generated under this approach are the ratio of positive and negative words and total number of words, LDA topics, the total number of the first person singular pronouns, and the ratio of words number and sentences number, TF-IDF weights and other. All features can be divided into the next categories: Ideational (topics, opinions, emotions), Interpersonal (modality, personal pronoun), and Textual (writing style, genre).

Customer relationship management has traditionally been based on the internal databases of customers (Zekić Sušac et al., 2015) [74], and various data mining approaches have been used in order to improve it (Furner et al., 2012) [75]. Named entity recognition has recently become a rich source of information relevant for customer relationship management, since it allows financial institutions, to for the extraction of client names, bank account numbers, IBAN from their internal databases and link them to external sources, such as social media. There are dictionaries with predefined named entities that every organization can use for quick start and result. For better results, solutions that are more complex are needed. Since tweets and comments from social media and websites usually lack context and are noisy, there are more complex solutions like supervised approach for named entity recognition (Ritter et al., 2011) [50]. Gender recognition is also relevant to customer relationship management. While Charness and Gneezy (2012) [76] investigated gender differences in different countries in risk-taking, Lotto (2018) [54] used various determinants, among which gender prediction, to predict financial inclusion, and compared it with traditional banking services. Therefore, gender recognition of customers using text mining can be of high significance. For example, Phuong and Phuong (2014) [52] performed research on predicting users' gender based on browsing history, important for marketing and personalization. Kucukyilmaz et al. (2006) [53] performed an analysis of gender prediction in computer-mediated-communication/ chatbots.

Stock price prediction aims to determine the future value of an organization and their stocks. This information can bring more profit to the information owner and hence the great interest in these analyses. The hypothesis is that company stock prices can be predicted and this is part where it

gets complex. Similar research has been demonstrated using news and macroeconomic indicators (Elshendy et al., 2017) [77]. It is not just about currently available data like history change of stock prices but also textual documents, which can bring new insights into these hybrid approaches. Previous methods in this field did not yield impressive results, and despite low accuracy, changes have been made and model accuracies raised. Schumaker and Chen (2009) [6] focusing to text-mining of financial news articles analysis using words and phrases representations for the purpose of stock market prediction, while Tetlock (2007) [33], Bollen et al. (2011) [35], Hagenau et al. (2013) [37] focused to the relationship between the news, blogs, and social media and the stock market. Combining time series data for stocks and their prices with information gathered from text mining is a key part. This is one of the popular examples of text mining in the financial industry. Time series datasets contain data about a stock event over time, and they lack context, which is tried to fill with text mining techniques. Textual information enriches our base time series dataset by extracting news articles related to stocks of interest and thanks to big data technologies; this complex task could be done in real-time. Textual data have rich information and hypothesis is that company's report or breaking news can affect the stock price. Forums where special topics about the financial world should be covered and where financial experts meet can be a good source of textual data. Sentiment or topic models from previous tasks can be integrated together with time series in the hybrid model. Combination of time series and textual data show improvements in net profit in comparison of just using one of those parts alone (Zhai at al., 2007) [78].

Summarization of textual documents is of great importance for business of any financial institution. Understanding of textual documents, as well as an easy search of those documents, can be achieved with the summarization. Text summarization techniques summarize legal documents in four structures: intro, context, juridical analysis, and conclusion (Farzindar and Lapalme, 2004) [79]. To achieve this pre-processing step are used to split the text into chunks (sentence or token) and then annotated with POS tagging step (structures like those that intro, conclusion and other previously mentioned are detected here). Final steps are filtering (removal of unnecessary steps) and selection (high-score units are found) where text mining is done in the last step. Text summarization has two methods: extractive and abstractive where extractive method uses tokens from original documents and creates the summary, while abstractive methods generate completely new tokens to better capture the meaning of the original document. For the purpose of text summarization clustering (Wagh, 2013) [80] can be used for a better search. The pre-processing steps are also needed, such as stemming, and clustering algorithms are used for grouping keywords, phrases or documents in homogenous groups.

7. Conclusions

By reviewing 123 papers, this paper aims to provide answers to the three research questions, and for that purpose, a qualitative analysis of literature has been conducted using a systematic literature review, citation, and co-citation investigation.

The first research question was answered using the bibliometric analysis. The most important studies with the highest number of citations in the field have been identified, and a brief overview of the themes is given. In addition, papers that are the source of the field have been presented prior to the critical connection with recent studies identified. Based on this, the paper contributed to the existing literature through an overview of the most significant studies published in the Web of Science databases. Research trends have been identified as well. After reviewing the papers, it is possible to conclude that the research focus is on stocks price prediction, financial fraud detection and market forecast utilizing online text mining. The research results reveal that the current research trends of text mining are related to the need to analyze large amounts of data on websites and pages on social media, and to identify and test various text-mining techniques.

The second research question was answered by providing the analysis of techniques for text mining in the financial sector. Analysis of big amounts of data represents the transition to analytic-driven business, conducted by big companies, small enterprises or research teams, in order to

identify significant information and transform it into new knowledge. Text analytics or text mining of big data, conducted by various techniques (keyword extraction, named entity recognition, gender prediction, sentiment analysis, topic extraction, and social network analysis) has moved from research centers to real-world institutions, such as financial and banking institutions.

The third research question was answered by the analysis of data sources used for text mining techniques. Results revealed that most of the research focuses on external data sources, such as news and online media posts for the purpose of stock market predictions, and fraud detections. The number of research studies using internal data sources is low. Therefore, the utilization of internal data sources will be a rich source of future research with both theoretical and practical contributions. Various research using internal text sources, such as emails, corporate wikis, financial statements, and project reports could be useful for various purposes, such as human resource management, internal audit, and customer relationship management. In addition, various multimedia files could also be the high-value additional component of text mining analysis (Pouli et al., 2015 [81]; Stai et al., 2018 [82]; Ma et al., 2011 [83]).

The main limitation of our work is the usage of bibliometric approaches to the literature analysis, which has certain limitations. By selecting the database for studies search (Web of Knowledge), specific studies remain invisible to this analysis (Batistič et al., 2017 [12]).

Research results also generate several paths for future research directions. First, more up-to-date outlook to the usage of text mining in finance could be attained with the use of so-called "grey" literature sources, such as case studies, corporate reports, and text-mining software projects (Adams et al., 2017 [84]). Second, usage of text mining in finance should be reviewed according to different decisions that are made based on its results (e.g., tactical, operational and strategic decisions). Taxonomy of various decisions based on text mining in finance could be developed in order to support decision making in a more effective manner, following the work of Gray et al. (2014) [18]. Third, characteristics of organizations that have implemented text mining in their business processes should be investigated, with the goal of identifying best-practice approaches, but also obstacles that stand on the way to the successful implementation of text mining in finance. Finally, more in-depth analysis of data sources used for text mining in finance should be conducted, focusing more on the internal documents as the domain of the analysis.

Author Contributions: All authors contributed equally to this paper. M.P.B. and Ž.K. conceptualized the research; M.P.B. and L.T. developed the methdoology; L.T. applied software for citation analysis, S.S. and Ž.K. wrote the original draft, while M.P.B. and L.T. conduct the review and editing of final version of the paper.

Funding: This research received no external funding.

Acknowledgments: Authors would like to acknowledge the support and advice provided by the special issue editor, as well as reviewers, which greatly improved the quality of the paper with their comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. List of Papers Included in the Literature Review

1. Al Nasseri, A.; Tucker, A.; de Cesare, S. Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. Expert Systems with Applications, 42(23), 2015, pp. 9192–9210.
2. Alostad, H.; Davulcu, H. Directional prediction of stock prices using breaking news on Twitter. In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015, pp. 523–530.
3. AL-Rubaiee, H.; Qiu, R.; Li, D. Visualising Arabic Sentiments and Association Rules in Financial Text. International journal of advanced computer science and applications, **2017**, 8(2), pp. 1–7.
4. Alvarado, J.C.S.; Verspoor, K.; Baldwing, T. Domain Adaptation of Named Entity Recognition to Support Credit Risk Assessment. In Proceedings of Australasian Language Technology Association Workshop, University of Western Sydney, Australia, 8–9th December, 2015, pp. 84–90.

5. Ammann, M.; Frey, R.; Verhofen, M. Do newspaper articles predict aggregate stock returns? Journal of behavioral finance, **2014**, 15(3), pp. 195–213.
6. Bai, X.; Dong, Y.; Hu, N. Financial report readability and stock return synchronicity. Applied Economics, **2019**, 51(4), pp. 346–363.
7. Balakrishnan, R.; Qiu, X. Y.; Srinivasan, P. On the predictive ability of narrative disclosures in annual reports. European Journal of Operational Research, 202(3), **2010**, pp. 789–801.
8. Boskou, G.; Kirkos, E.; Spathis, C. Assessing Internal Audit with Text Mining. Journal of Information Knowledge Management, 17(2), **2018**, 1850020.
9. Bowers, A. J.; Chen, J. Ask and ye shall receive? Automated text mining of Michigan capital facility finance bond election proposals to identify which topics are associated with bond passage and voter turnout. Journal of Education Finance, **2015**, 164–196.
10. Cao, M., Chychyla, R., & Stewart, T. Big Data analytics in financial statement audits. Accounting Horizons, 29(2), **2015**, pp. 423–429.
11. Chen, C. L.; Liu, C. L.; Chang, Y. C.; Tsai, H. P. Opinion mining for relating subjective expressions and annual earnings in US financial statements. Journal of information science and engineering, 29, **2014**, 743–764.
12. Chen, K.; Li, X.; Xu, B.; Yan, J.; Wang, H. Intelligent agents for adaptive security market surveillance. Enterprise Information Systems, **2017**, 11(5), pp. 652–671.
13. Chen, K.; Yin, J.; Pang, S. A design for a common-sense knowledge-enhanced decision-support system: Integration of high-frequency market data and real-time news. Expert Systems, **2017**, 34(3), e12209.
14. Chen, W.; Lai, K.; Cai, Y. Topic generation for Chinese stocks: a cognitively motivated topic modeling method using social media data. Quantitative finance and economics, **2018**, 2(2), pp. 279–293.
15. Chung, W. BizPro: Extracting and categorizing business intelligence factors from textual news articles. International Journal of Information Management, 34(2), **2014**, pp. 272–284.
16. Chye Koh, H.; Kee Low, C. Going concern prediction using data mining techniques. Managerial Auditing Journal, **2004**, 19(3), pp. 462–476.
17. Coussement, K.; Van den Poel, D. Integrating the voice of customers through call center emails into a decision support system for churn prediction. Information Management, 45(3), **2008**, pp. 164–174.
18. Dong, W.; Liao, S.; Liang, L. Financial Statement Fraud Detection using Text Mining: a Systemic Functional Linguistics Theory Perspective. In Proceedings of the Pacific Asia Conference on Information Systems (PACIS), Chiayi, Taiwan, 26th June, **2016**, AISeL, p. 188.
19. Duan, Z.; He, Y.; Zhong, Y. Corporate social responsibility information disclosure objective or not: An empirical research of Chinese listed companies based on text mining. Nankai Business Review International, 9(4), **2018**, pp. 519–539.
20. Ediger, D.; Jiang, K.; Riedy, J.; Bader, D.A.; Corley, C. Massive social network analysis: Mining twitter for social good. In Proceedings of 39th International Conference on Parallel Processing, San Diego, CA, USA, 13–16th September, 2010, IEEE, pp. 583–593.
21. Eler, D.M.; Grosa, D.; Pola, I.; Garcia, R.; Correia, R.; Teixeira, J. Analysis of Document Pre-Processing Effects in Text and Opinion Mining. Information, 2018, 9(4), p. 100.
22. Feuerriegel, S.; Gordon, J. Long-term stock index forecasting based on text mining of regulatory disclosures. Decision Support Systems, 112, **2018**, 88–97.
23. Feuerriegel, S.; Gordon, J. News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. European Journal of Operational Research, 272(1), **2019**, pp. 162–175.
24. Feuerriegel, S.; Prendinger, H. News-based trading strategies. Decision Support Systems, 90, **2016**, 65–74.

25. Fisher, I. E.; Garnsey, M. R.; Hughes, M. E. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Intelligent Systems in Accounting, Finance and Management, **2016**, 23(3), pp. 157–214.
26. Ghailan, O.; Mokhtar, H. M.; Hegazy, O. (2016). Improving Credit Scorecard Modeling Through Applying Text Analysis. institutions, International Journal of Advanced Computer Science and Applications, 7(4), **2016**, 512–517.
27. Groth, S. S.; Muntermann, J. An intraday market risk management approach based on textual analysis. Decision Support Systems, 50(4), **2011**, pp. 680–691.
28. Groth, S. S.; Siering, M.; Gomber, P. How to enable automated trading engines to cope with news-related liquidity shocks? Extracting signals from unstructured data. Decision Support Systems, **2014**, 62, pp. 32–42.
29. Gül, S.; Kabak, Ö.; Topcu, I. A multiple criteria credit rating approach utilizing social media data. Data Knowledge Engineering, **2018**, 116, pp. 80–99.
30. Gunduz, H.; Cataltepe, Z. Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. Expert Systems with Applications, 42(22), **2015**, pp. 9001–9011.
31. Guo, P.; Shen, Y. The impact of Internet finance on commercial banks' risk taking: evidence from China. China Finance and Economic Review, 4(1), **2016**, p. 16.
32. Hagenau, M.; Liebmann, M.; Neumann, D. Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems, **2013**, 55(3), pp. 685–697.
33. Hájek, P. Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. Neural Computing and Applications, 29(7), **2018**, pp. 343–358.
34. Hajek, P.; Henriques, R. Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods. Knowledge-Based Systems, **2017**, 128, 139–152.
35. Han, W.; Fang, Z.; Yang, L. T.; Pan, G.; Wu, Z. Collaborative policy administration. IEEE Transactions on Parallel and Distributed Systems, 25(2), **2014**, pp. 498–507.
36. Hasan, K.S.; Ng, V. Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27th June, 2014, 1, pp. 1262–1273.
37. Hasan, K.S.; Ng, V. Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27th June, 2014, 1, pp. 1262–1273.
38. Holton, C. Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multi-billion dollar problem, Decision Support Systems, **2009**, 46(4), pp. 853–864.
39. Hong, W.; Wang, W.; Weng, Y.; Luo, S.; Hu, P.; Zheng, X.; Qi, J. STOCK PRICE MOVEMENTS PREDICTION WITH TEXTUAL INFORMATION. Mechatronic Systems and Control, 46(3), **2018**, pp. 141–149.
40. Hsu, M. F.; Yeh, C. C.; Lin, S. J. Integrating dynamic Malmquist DEA and social network computing for advanced management decisions. Journal of Intelligent Fuzzy Systems, 35(1), **2018**, pp. 1–11.
41. Huang, C. J.; Liao, J. J.; Yang, D. X.; Chang, T. Y.; Luo, Y. C. Realization of a news dissemination agent based on weighted association rules and text mining techniques. Expert Systems with Applications, 37(9), **2010**, pp. 6409–6413.
42. Huang, H.; Li, Y.; Zhang, Y. Investors' attention and overpricing of IPO: an empirical study on China's growth enterprise market. Information Systems and e-Business Management, 16(4), **2018**, pp. 761–774.

43. Humpherys, S.L.; Moffitt, K.C.; Burns, M.B.; Burgoon, J.K.; Felix, W.F. Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems, **2011**, 50(3), pp. 585–594.
44. Kamaruddin, S. S.; Bakar, A. A.; Hamdan, A. R.; Nor, F. M.; Nazri, M. Z. A.; Othman, Z. A.; Hussein, G. S. A text mining system for deviation detection in financial documents. Intelligent Data Analysis, 19(s1), **2015**, pp. S19-S44.
45. Kamaruddin, S. S.; Hamdan, A. R.; Bakar, A. A.; Mat Nor, F. Deviation detection in text using conceptual graph interchange format and error tolerance dissimilarity function. Intelligent Data Analysis, **2012**, 16(3), pp. 487–511.
46. Kraus, M.; Feuerriegel, S. Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems, 104, **2017**, 38–48.
47. Krishnamoorthy, S. Sentiment analysis of financial news articles using performance indicators. Knowledge and Information Systems, 56(2), **2018**, pp. 373–394.
48. Kucukyilmaz, T.; Cambazoglu, B.B.; Aykanat, C.; Can, F. Chat Mining for Gender Prediction. In Proceedings of the 4th International Conference in Advances in Information Systems (ADVIS), Izmir, Tukrey, 18–20th October, 2006, Springer: Berlin, Heidelberg, pp. 274–283.
49. Kumar, B. S.; Ravi, V. A survey of the applications of text mining in financial domain. Knowledge-Based Systems, 114, **2016**, 128–147.
50. Lee, P.; Owda, M.; Crockett, K. Novel methods for resolving false positives during the detection of fraudulent activities on stock market financial discussion boards. International Journal of Advanced Computer Science and Applications (IJACSA), 2018, 9(1), pp. 1–10.
51. Lee, W.S.; So Young, S. Identifying Emerging Trends of Financial Business Method Patents. Sustainability, **2017**, 9(9).
52. L'Huillier, G.; Alvarez, H.; Ríos, S.A.; Aguilera, F. Topic-based social network analysis for virtual communities of interests in the dark web. ACM SIGKDD Explorations Newsletter, **2011**, 12(2), pp. 66–73.
53. Li, G.; Dai, J. S.; Park, E. M.; Park, S. T. A study on the service and trend of Fintech security based on text-mining: focused on the data of Korean online news. Journal of Computer Virology and Hacking Techniques, 13(4), **2017**, pp. 249–255.
54. Li, Q.; Chen, Y.; Wang, J.; Chen, Y.; Chen, H. Web media and stock markets: A survey and future directions from a big data perspective. IEEE Transactions on Knowledge and Data Engineering, 30(2), **2018**, pp. 381–399.
55. Li, W.; Chen, H.; Nunamaker Jr, J. F. Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. Journal of Management Information Systems, 33(4), **2016**, pp. 1059–1086.
56. Li, X.; Chen, K.; Sun, S. X.; Fung, T.; Wang, H.; Zeng, D. D. A commonsense knowledge-enabled textual analysis approach for financial market surveillance. INFORMS Journal on Computing, 28(2), **2016**, pp. 278–294.
57. Li, X.; Sun, S. X.; Chen, K.; Fung, T.; Wang, H. Design theory for market surveillance systems. Journal of Management Information Systems, 32(2), **2015**, pp. 278–313.
58. Lin, S. J.; Hsu, M. F. Decision making by extracting soft information from CSR news report. Technological and Economic Development of Economy, 24(4), **2018**, pp. 1344–1361.
59. Linardos, E.; Kermanidis, K. L.; Maragoudakis, M. Using financial news articles with minimal linguistic resources to forecast stock behaviour. International Journal of Data Mining, Modelling and Management, 7(3), **2015**, pp. 185–212.
60. Lotto, J. Examination of the Status of Financial Inclusion and its Determinants in Tanzania. Sustainability, **2018**, 10(8), p. 2873.
61. Loughran, T.; McDonald, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, **2010**, 66(1), pp. 35–65.

62. Lu, H. M.; Chen, H.; Chen, T. J.; Hung, M. W.; Li, S. H. Financial text mining: Supporting decision making using web 2.0 content. IEEE Intelligent Systems, 25(2), **2010**, pp. 78–82.
63. Lu, Y. C.; Shen, C. H.; Wei, Y. C. Revisiting early warning signals of corporate credit default using linguistic analysis. Pacific-Basin Finance Journal, 24, **2013**, 1–21.
64. Lugmayr, A.; Grueblbauer, J. Review of information systems research for media industry–recent advances, challenges, and introduction of information systems research in the media industry. Electronic Markets, 27(1), **2017**, pp. 33–47.
65. Mai, F.; Shan, Z.; Bai, Q.; Wang, X.; Chiang, R. H. How does social media impact bitcoin value? A test of the silent majority hypothesis. Journal of Management Information Systems, 35(1), **2018**, pp. 19–52.
66. Mao, H.; Jin, X.; Zhu, L. Methods of Measuring Influence of Bank Customer Using Social Network Model. American Journal of Industrial and Business Management, **2015**, 5(4), pp. 155–160.
67. Maragoudakis, M.; Serpanos, D. Exploiting financial news and social media opinions for stock market analysis using mcmc bayesian inference. Computational Economics, 47(4), **2016**, pp. 589–622.
68. Masawi, B.; Bhattacharya, S.; Boulter, T. The power of words: A content analytical approach examining whether central bank speeches become financial news. Journal of information science, 40(2), **2014**, pp. 198–210.
69. Molnar, Z.; Strelka, J. Competitive Intelligence for small and middle enterprises. E M EKONOMIE A MANAGEMENT, 15(3), **2012**, pp. 156–170.
70. Moro, S.; Cortez, P.; Rita, P. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications, 42(3), **2015**, pp. 1314–1324.
71. Motokawa, K. Human capital disclosure, accounting numbers, and share price. Journal of Financial Reporting and Accounting, 13(2), **2015**, pp. 159–178.
72. Nakayama, M.; Wan, Y. Exploratory Study on Anchoring: Fake Vote Counts in Consumer Reviews Affect Judgments of Information Quality. Journal of theoretical and applied electronic commerce research, **2017**, 12(1), pp. 1–20.
73. Narayanan, V.; Arora, I.; Bhatia, A. Fast and accurate sentiment classification using an enhanced Naive Bayes model. In Proceedings of 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), Hefei, China, 20th-23rd October, 2013, Springer: Berlin, Heidelberg, pp. 194–201.
74. Nardo, M.; Petracco-Giudici, M.; Naltsidis, M. Walking down wall street with a tablet: A survey of stock market predictions using the web. Journal of Economic Surveys, 30(2), **2016**, pp. 356–369.
75. Nassirtoussi, A. K.; Aghabozorgi, S.; Wah, T. Y.; Ngo, D. C. L. Text mining for market prediction: A systematic review. Expert Systems with Applications, 41(16), **2014**, pp. 7653–7670.
76. Nassirtoussi, A. K.; Aghabozorgi, S.; Wah, T. Y.; Ngo, D. C. L. Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications, 42(1), **2015**, pp. 306–324.
77. Neumann, M.; Sartor, N. A semantic network analysis of laundering drug money. Journal of Tax Administration, 2(1), **2016**, pp. 73–94.
78. Nishanth, K. J.; Ravi, V.; Ankaiah, N.; Bose, I. Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. Expert Systems with Applications, 39(12), 2012, pp. 10583–10589.
79. Nizer, P. S. M.; Nievola, J. C. Predicting published news effect in the Brazilian stock market. Expert Systems with Applications, 39(12), **2012**, pp. 10674–10680.
80. Nopp, C.; Hanbury, A. Detecting Risks in the Banking System by Sentiment Analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17th-21st September, **2015**, pp. 591–600.

81. Novalija, I.; Mladenić, D.; Bradeško, L. OntoPlus: Text-driven ontology extension using ontology content, structure and co-occurrence information. Knowledge-Based Systems, 24(8), 2011, pp. 1261–1276.
82. Oliveira, N.; Cortez, P.; Areal, N. The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications, 73, 2017, 125–144.
83. Ong, T. H.; Chen, H.; Sung, W. K.; Zhu, B. Newsmap: a knowledge map for online news. Decision Support Systems, 39(4), 2005, pp. 583–597.
84. Pang, B.; Lee, L. Opinion mining and sentiment analysis. Foundations and Trends®in Information Retrieval, 2008, 2(1–2), pp. 1–135.
85. Park, S.; Lee, W.; Moon, I. C. Associative topic models with numerical time series. Information Processing Management, 51(5), 2015, pp. 737–755.
86. Patrick J. The Scamseek Project – Text Mining for Financial Scams on the Internet. In: Williams G.J.; Simoff S.J. (eds) Data Mining. Lecture Notes in Computer Science, 3755, **2006**, Springer, Berlin, Heidelberg.
87. Phuong, D.V.; Phuong, T.M. Gender Prediction Using Browsing History. In *Knowledge and Systems* Engineering; Huynh, V.; Denoeux, T.; Tran, D.; Le, A.; Pham, S.; Eds.; Advances in Intelligent Systems and Computing, Springer: Cham, 244, **2014**, pp. 271–283.
88. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: an experimental study. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, Scotland, UK, 27th–31st July, 2011, Association for Computational Linguistics, pp. 1524–1534.
89. Roh, T.; Jeong, Y.; Yoon, B. Developing a Methodology of Structuring and Layering Technological Information in Patent Documents through Natural Language Processing. Sustainability, 2017, 9(11).
90. Rönnqvist, S.; Sarlin, P. Bank distress in the news: Describing events through deep learning. Neurocomputing, 264, **2017**, 57–70.
91. Sakai, H.; Masuyama, S. Assigning polarity to causal information in financial articles on business performance of companies. IEICE transactions on information and systems, 92(12), **2009**, pp. 2341–2350.
92. Saleiro, P.; Rodrigues, E. M.; Soares, C.; Oliveira, E. Texrep: a text mining framework for online reputation monitoring. New Generation Computing, 35(4), **2017**, pp. 365–389.
93. Santos, C. L.; Rita, P.; Guerreiro, J. Improving international attractiveness of higher education institutions based on text mining and sentiment analysis. International Journal of Educational Management, 32(3), **2018**, pp. 431–447.
94. Schniederjans, D.; Cao, E. S.; Schniederjans, M. Enhancing financial performance with social media: An impression management perspective. Decision Support Systems, 55(4), **2013**, pp. 911–918.
95. Schumaker, R.P.; Chen, H. Textual analysis of stock market prediction using breaking financial news. ACM Transactions on Information Systems, **2009**, 27(2), pp. 1–19.
96. Schumaker, R.P.; Zhang, Y.; Huang, C.N.; Chen, H. Evaluating sentiment in financial news articles. Decision Support Systems, **2012**, 53(3), pp. 458–464.
97. See-To, E. W.; Yang, Y. Market sentiment dispersion and its effects on stock return and volatility. Electronic Markets, 27(3), **2017**, pp. 283–296.
98. Seki, K.; Shibamoto, M. Construction and application of sentiment lexicons in Finance. International Journal of Multimedia Data Engineering and Management (IJMDEM), 9(1), **2018**, pp. 22–35.
99. Seo, J. H.; Park, E. M. A study on financing security for smartphones using text mining. Wireless Personal Communications, 98(4), **2018**, pp. 3109–3127.

100. Shynkevich, Y.; McGinnity, T. M.; Coleman, S. A.; Belatreche, A. Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. **2016**, Decision Support Systems, 85, 74–83.
101. Staines, J.; Barber, D. Topic factor models: Uncovering thematic structure in equity market data. Intelligent Data Analysis, 19(s1), **2015**, pp. S69-S85.
102. Takahashi, S.; Takahashi, M.; Takahashi, H.; Tsuda, K. Analysis of stock price return using textual data and numerical data through text mining. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, **2006**, pp. 310–316.
103. Tetlock, P. C. Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, **2007**, 62(3), pp. 1139–1168.
104. Tobback, E.; Naudts, H.; Daelemans, W.; de Fortuny, E. J.; Martens, D. Belgian economic policy uncertainty index: Improvement through text mining. International journal of forecasting, **2018**, 34(2), pp. 355–365.
105. Tsai, M. F.; Wang, C. J. On the risk prediction and analysis of soft information in finance reports. European Journal of Operational Research, 257(1), **2017**, pp. 243–250.
106. Tsai, M. F.; Wang, C. J.; Chien, P. C. Discovering finance keywords via continuous-space language models. ACM Transactions on Management Information Systems (TMIS), **2016**, 7(3), p. 7.
107. Tsukioka, Y.; Yanagi, J.; Takada, T. Investor sentiment extracted from internet stock message boards and IPO puzzles. International Review of Economics Finance, 56, **2018**, 205–217.
108. Tumarkin, R.; Whitelaw, R.F. News or noise? Internet Postings Stock Prices. Financial Analysts Journal, **2001**, 57(3), pp. 41–51.
109. Wagh, R.S. Knowledge discovery from legal documents dataset using text mining techniques. International Journal of Computer Applications, **2013**, 66(23), pp. 32–34.
110. Wang, B.; Huang, H.; Wang, X. A novel text mining approach to financial time series forecasting. Neurocomputing, **2012**, 83, 136–145.
111. Wang, H.; Wu, J.; Yuan, S.; Chen, J. On characterizing scale effect of Chinese mutual funds via text mining. Signal Processing, 124, **2016**, 266–278.
112. Wang, T.; Kannan, K. N.; Ulmer, J. R. The association between the disclosure and the realization of information security risk factors. Information Systems Research, 24(2), **2013**, pp. 201–218.
113. Xia, Y.; Su, W.; Lau, R. Y.; Liu, Y. Discovering latent commercial networks from online financial news articles. Enterprise Information Systems, 7(3), **2013**, pp. 303–331.
114. Xie, Y.; Jiang, H. Stock Market Forecasting Based on Text Mining Technology: A Support Vector Machine Method. JCP, 12(6), **2017**, pp. 500–510.
115. Xing, F. Z.; Cambria, E.; Welsch, R. E. Intelligent asset allocation via market sentiment views. ieee ComputatioNal iNtelligeNCe magaziNe, 13(4), **2018**, pp. 25–34.
116. Xing, F. Z.; Cambria, E.; Welsch, R. E. Natural language based financial forecasting: a survey. Artificial Intelligence Review, 50(1), **2018**, pp. 49–73.
117. Yan, J.; Wang, K.; Liu, Y.; Xu, K.; Kang, L.; Chen, X.; Zhu, H. Mining social lending motivations for loan project recommendations. Expert Systems with Applications, 111, **2018**, 100–106.
118. Yang, R.; Yu, Y.; Liu, M.; Wu, K. Corporate risk disclosure and audit fee: a text mining approach. European Accounting Review, 27(3), 2018, pp. 583–594.
119. Yao, C. Z.; Sun, B. Y.; Lin, J. N. A study of correlation between investor sentiment and stock market based on Copula model. Kybernetes, 46(3), 2017, pp. 550–571.
120. Yong, S. H. I.; Tang, Y. R.; Cui, L. X.; Wen, L. O. N. G. A text mining based study of investor sentiment and its influence on stock returns. Economic Computation Economic Cybernetics Studies Research, 52(1), 2017, pp. 183–199.
121. Yu, M.; Guo, C. Using news to predict Chinese medicinal material price index movements. Industrial Management Data Systems, 118(5), **2018**, pp. 998–1017.

122. Yuan, X.; Chang, W.; Zhou, S.; Cheng, Y. Sequential Pattern Mining Algorithm Based on Text Data: Taking the Fault Text Records as an Example. Sustainability, 10(11), **2018**, p. 4330.
123. Zhao, D. Frontiers of big data business analytics: patterns and cases in online marketing. In Big data and business analytics; Leibowitz, J.; Eds.; CRC Press: Boca Raton, **2013**, pp. 46–68.

References

1. Abrahamson, E.; Rosenkopf, L. Social Network Effects on the Extent of Innovation Diffusion: A Computer Simulation. *Organ. Sci.* **1997**, *8*, 289–309. [CrossRef]
2. Adams, R.J.; Smart, P.; Huff, A.S. Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *Int. J. Manag. Rev.* **2017**, *19*, 432–454. [CrossRef]
3. Alvarado, J.C.S.; Verspoor, K.; Baldwing, T. Domain Adaptation of Named Entity Recognition to Support Credit Risk Assessment. In Proceedings of the Australasian Language Technology Association Workshop, Parramatta, Australia, 8–9 December 2015; pp. 84–90.
4. Arner, D.W.; Barberis, J.; Buckley, R.P. The evolution of Fintech: A new post-crisis paradigm. *Georget. J. Int. Law.* **2015**, *47*, 1271. [CrossRef]
5. Bannan-Ritland, B. The role of design in research: The integrative learning design framework. *Educ. Res.* **2003**, *32*, 21–24. [CrossRef]
6. Batistič, S.; Černe, M.; Vogel, B. Just how multi-level is leadership research? A document co-citation analysis 1980–2013 on leadership constructs and outcomes. *Leadersh. Q.* **2017**, *28*, 86–103. [CrossRef]
7. Best, A.; Terpstra, J.L.; Moor, G.; Riley, B.; Norman, C.D.; Glasgow, R.E. Building knowledge integration systems for evidence-informed decisions. *J. Health Organ. Manag.* **2009**, *23*, 627–641. [CrossRef] [PubMed]
8. Bharti, S.K.; Babu, K.S. Automatic Keyword Extraction for Text Summarization: A Survey. 2017. Available online: https://arxiv.org/ftp/arxiv/papers/1704/1704.03242.pdf (accessed on 12 August 2018).
9. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
10. Charness, G.; Gneezy, U. Strong evidence for gender differences in risk taking. *J. Econ. Behav. Organ.* **2012**, *83*, 50–58. [CrossRef]
11. Chye Koh, H.; Kee Low, C. Going concern prediction using data mining techniques. *Manag. Audit. J.* **2004**, *19*, 462–476. [CrossRef]
12. Coussement, K.; Van den Poel, D. Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Inf. Manag.* **2008**, *45*, 164–174. [CrossRef]
13. Dong, W.; Liao, S.; Liang, L. Financial Statement Fraud Detection using Text Mining: A Systemic Functional Linguistics Theory Perspective. In Proceedings of the Pacific Asia Conference on Information Systems (PACIS), Chiayi, Taiwan, 26 June 2016; p. 188.
14. Ediger, D.; Jiang, K.; Riedy, J.; Bader, D.A.; Corley, C. Massive social network analysis: Mining twitter for social good. In Proceedings of the 39th International Conference on Parallel Processing, San Diego, CA, USA, 13–16 September 2010; pp. 583–593.
15. Eler, D.M.; Grosa, D.; Pola, I.; Garcia, R.; Correia, R.; Teixeira, J. Analysis of Document Pre-Processing Effects in Text and Opinion Mining. *Information* **2018**, *9*, 100. [CrossRef]
16. Elshendy, M.; Fronzetti Colladon, A. Big data analysis of economic news: Hints to forecast macroeconomic indicators. *Int. J. Eng. Bus. Manag.* **2017**, *9*, 1847979017720040. [CrossRef]
17. Fan, W.; Wallace, L.; Rich, S.; Zhang, Z. Tapping the power of text mining. *Commun. ACM* **2006**, *49*, 77–82. [CrossRef]
18. Farzindar, A.; Lapalme, G. Letsum, an automatic legal text summarizing system. In *Legal Knowledge and Information Systems: JURIX 2004: The Seventeenth Annual Conference*; Gordon, T., Ed.; IOS Press: Berlin, Germany, 2004; pp. 11–18.
19. Finacle Connect. Connecting the Banking World. Artificial Intelligence Powered Banking. 2018. Available online: https://active.ai/wp-content/uploads/2018/05/Finacle-Connect-2018-leading-ai-online.pdf (accessed on 12 August 2018).
20. Friedmann, E.; Lowengart, O. The Effect of Gender Differences on the Choice of Banking Services. *J. Serv. Sci. Manag.* **2016**, *9*, 361–377. [CrossRef]
21. Furner, C.P.; Zinko, R.; Zhu, Z. Examining the Role of Mobile Self-Efficacy in the Word-of-Mouth/Mobile Product Reviews Relationship. *Int. J. E-Serv. Mob. Appl. (IJESMA)* **2017**, *10*, 40–60. [CrossRef]

22. Galli, E.; Rossi, S.P.S. Bank Credit Access and Gender Discrimination: An Empirical Analysis. In *Contributions to Economics*; Springer: Berlin, Germany, 2014; pp. 111–123.
23. Glancy, F.H.; Yadav, S.B. A computational model for financial reporting fraud detection. *Decis. Support Syst.* **2011**, *50*, 595–601. [CrossRef]
24. Go, A.; Bhayani, R.; Huang, L. Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report, Stanford. 2009, Volume 1. Available online: https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf (accessed on 12 August 2018).
25. Gray, G.L.; Debreceny, R.S. A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. *Int. J. Account. Inf. Syst.* **2014**, *15*, 357–380. [CrossRef]
26. Grishman, R.; Sundheim, B. Message understanding conference-6: A brief history. In Proceedings of the COLING 1996: The 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5–9 August 1996; Volume 1, pp. 466–471.
27. Hagenau, M.; Liebmann, M.; Neumann, D. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decis. Support Syst.* **2013**, *55*, 685–697. [CrossRef]
28. Hajizadeh, E.; Ardakani, H.D.; Shahrabi, J. Application of data mining techniques in stock markets: A survey. *J. Econ. Int. Financ.* **2010**, *2*, 109–118.
29. Hasan, K.S.; Ng, V. Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; Volume 1, pp. 1262–1273.
30. Herráez, B.; Bustamante, D.; Saura, J.R. Information classification on social networks. Content analysis of e-commerce companies on Twitter. *Revista Espacios* **2017**, *38*, 16.
31. Holton, C. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decis. Support Syst.* **2009**, *46*, 853–864. [CrossRef]
32. Humpherys, S.L.; Moffitt, K.C.; Burns, M.B.; Burgoon, J.K.; Felix, W.F. Identification of fraudulent financial statements using linguistic credibility analysis. *Decis. Support Syst.* **2011**, *50*, 585–594. [CrossRef]
33. Hussin, M.F.; Kamel, M.S.; Nagi, M.H. An efficient two-level SOMART document clustering through dimensionality reduction. In Proceedings of the International Conference on Neural Information Processing, Calcutta, India, 22–25 November 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 158–165.
34. Jin, M.; Wang, Y.; Zeng, Y. Application of Data Mining Technology in Financial Risk Analysis. *Wirel. Pers. Commun.* **2018**, *102*, 3699–3713. [CrossRef]
35. Klopotan, I.; Zoroja, J.; Meško, M. Early warning system in business, finance, and economics: Bibliometric and topic analysis. *Int. J. Eng. Bus. Manag.* **2018**, *10*, 1847979018797013. [CrossRef]
36. Kucukyilmaz, T.; Cambazoglu, B.B.; Aykanat, C.; Can, F. Chat Mining for Gender Prediction. In Proceedings of the 4th International Conference in Advances in Information Systems (ADVIS), Izmir, Tukrey, 18–20 October 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 274–283.
37. Kumar, B.S.; Ravi, V. A survey of the applications of text mining in financial domain. *Knowl.-Based Syst.* **2016**, *114*, 128–147. [CrossRef]
38. Lee, W.S.; So Young, S. Identifying Emerging Trends of Financial Business Method Patents. *Sustainability* **2017**, *9*, 1670.
39. L'Huillier, G.; Alvarez, H.; Ríos, S.A.; Aguilera, F. Topic-based social network analysis for virtual communities of interests in the dark web. *ACM SIGKDD Explor. Newslett.* **2011**, *12*, 66–73. [CrossRef]
40. Lotto, J. Examination of the Status of Financial Inclusion and its Determinants in Tanzania. *Sustainability* **2018**, *10*, 2873. [CrossRef]
41. Loughran, T.; McDonald, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *J. Financ.* **2010**, *66*, 35–65. [CrossRef]
42. Ma, H.; Zhou, D.; Liu, C.; Lyu, M.R.; King, I. Recommender systems with social regularization. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, Hong Kong, China, 9–12 February 2011; pp. 287–296.
43. Mao, H.; Jin, X.; Zhu, L. Methods of Measuring Influence of Bank Customer Using Social Network Model. *Am. J. Ind. Bus. Manag.* **2015**, *5*, 155–160. [CrossRef]
44. Mathew, S. Financial Services Data Management: Big Data Technologies in Financial Services. Oracle White Paper. 2012. Available online: http://www.oracle.com/us/industries/financial-services/bigdata-in-fs-final-wp-1664665.pdf (accessed on 12 August 2018).

45. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* **2009**, *6*, e1000097. [CrossRef] [PubMed]
46. Moody, C.E. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. 2016. Available online: https://arxiv.org/abs/1605.02019 (accessed on 12 August 2018).
47. Moro, S.; Cortez, P.; Rita, P. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Syst. Appl.* **2015**, *42*, 1314–1324. [CrossRef]
48. Nakayama, M.; Wan, Y. Exploratory Study on Anchoring: Fake Vote Counts in Consumer Reviews Affect Judgments of Information Quality. *J. Theor. Appl. Electron. Commer. Res.* **2017**, *12*, 1–20. [CrossRef]
49. Narayanan, V.; Arora, I.; Bhatia, A. Fast and accurate sentiment classification using an enhanced Naive Bayes model. In Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), Hefei, China, 20–23 October 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 194–201.
50. Nassirtoussi, A.K.; Aghabozorgi, S.; Wah, T.Y.; Ngo, D.C.L. Text mining for market prediction: A systematic review. *Expert Syst. Appl.* **2014**, *41*, 7653–7670. [CrossRef]
51. Ngai, E.W.; Hu, Y.; Wong, Y.H.; Chen, Y.; Sun, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.* **2011**, *50*, 559–569. [CrossRef]
52. Niazi, M. Do systematic literature reviews outperform informal literature reviews in the software engineering domain? An initial case study. *Arab. J. Sci. Eng.* **2015**, *40*, 845–855. [CrossRef]
53. Nopp, C.; Hanbury, A. Detecting Risks in the Banking System by Sentiment Analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 591–600.
54. Ong, T.H.; Chen, H.; Sung, W.K.; Zhu, B. Newsmap: A knowledge map for online news. *Decis. Support Syst.* **2005**, *39*, 583–597. [CrossRef]
55. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends® Inf. Retr.* **2008**, *2*, 1–135. Available online: https://www.nowpublishers.com/article/Details/INR-011 (accessed on 24 February 2019). [CrossRef]
56. Pejic-Bach, M.; Pivar, J.; Krstić, Ž. Big Data for Prediction: Patent Analysis–Patenting Big Data for Prediction Analysis. In *Big Data Governance and Perspectives in Knowledge Management*; IGI Global: London, UK, 2019; pp. 218–240.
57. Phuong, D.V.; Phuong, T.M. Gender Prediction Using Browsing History. In *Knowledge and Systems Engineering*; Huynh, V., Denoeux, T., Tran, D., Le, A., Pham, S., Eds.; Advances in Intelligent Systems and Computing; Springer: Cham, Switzerland, 2014; Volume 244, pp. 271–283.
58. Pouli, V.; Kafetzoglou, S.; Tsiropoulou, E.E.; Dimitriou, A.; Papavassiliou, S. Personalized multimedia content retrieval through relevance feedback techniques for enhanced user experience. In Proceedings of the 2015 13th International Conference on Telecommunications (ConTEL), Graz, Austria, 13–15 July 2015; pp. 1–8.
59. Reyes-Menendez, A.; Saura, J.; Alvarez-Alonso, C. Understanding# WorldEnvironmentDay user opinions in Twitter: A topic-based sentiment analysis approach. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2537.
60. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: An experimental study. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, Scotland, UK, 27–31 July 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 1524–1534.
61. Roh, T.; Jeong, Y.; Yoon, B. Developing a Methodology of Structuring and Layering Technological Information in Patent Documents through Natural Language Processing. *Sustainability* **2017**, *9*, 2117. [CrossRef]
62. Saju, J.C.; Shaja, A.S. A Survey on Efficient Extraction of Named Entities from New Domains Using Big Data Analytics. In Proceedings of the 2nd International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), Tindivanam, India, 3–4 February 2017; pp. 170–175.
63. Saura, J.R.; Palos-Sánchez, P.; Cerdá Suárez, L.M. Understanding the digital marketing environment with KPIs and web analytics. *Future Internet* **2017**, *9*, 76. [CrossRef]
64. Schumaker, R.P.; Chen, H. Textual analysis of stock market prediction using breaking financial news. *ACM Trans. Inf. Syst.* **2009**, *27*, 1–19. [CrossRef]
65. Schumaker, R.P.; Zhang, Y.; Huang, C.N.; Chen, H. Evaluating sentiment in financial news articles. *Decis. Support Syst.* **2012**, *53*, 458–464. [CrossRef]

66. Srivastava, U.; Gopalkrishnan, S. Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks. *Procedia Comput. Sci.* **2015**, *50*, 643–652. [CrossRef]
67. Stai, E.; Kafetzoglou, S.; Tsiropoulou, E.E.; Papavassiliou, S. A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content. *Multimedia Tools Appl.* **2018**, *77*, 283–326.
68. Sun, S.; Luo, C.; Chen, J. A review of natural language processing techniques for opinion mining systems. *Inf. Fusion* **2017**, *36*, 10–25. [CrossRef]
69. Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. *J. Financ.* **2007**, *62*, 1139–1168. [CrossRef]
70. Tumarkin, R.; Whitelaw, R.F. News or noise? Internet Postings Stock Prices. *Financ. Anal. J.* **2001**, *57*, 41–51. [CrossRef]
71. Turner, D.; Schroeck, M.; Shockley, R. Analytics: The Real-World Use of Big Data in Financial Services. *J. Shanghai Jiaotong Univ. (Sci.)* **2012**, *21*, 210–214.
72. Vemuri, V.K. Mastering digital business: How powerful combinations of disruptive technologies are enabling the next wave of digital transformation, by Nicholas D. Evans. *J. Inf. Technol. Case Appl. Res.* **2017**, *19*, 128–130. [CrossRef]
73. Wagh, R.S. Knowledge discovery from legal documents dataset using text mining techniques. *Int. J. Comput. Appl.* **2013**, *66*, 32–34.
74. Wahono, R.S. A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks. *J. Softw. Eng.* **2015**, *1*, 1–16.
75. Wang, N.; Liang, H.; Jia, Y.; Ge, S.; Xue, Y.; Wang, Z. Cloud computing research in the IS discipline: A citation/co-citation analysis. *Decis. Support Syst.* **2016**, *86*, 35–47. [CrossRef]
76. Wuthrich, B.; Cho, V.; Leung, S.; Permunetilleke, D.; Sankaran, K.; Zhang, J. Daily stock market forecast from textual web data. In Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, 14 October 1998; pp. 2720–2725.
77. Yehia, A.M.; Ibrahim, L.F.; Abulkhair, M.F. Text Mining and Knowledge Discovery from Big Data: Challenges and Promise. *Int. J. Comput. Sci. Issues (IJCSI)* **2016**, *13*, 54–61.
78. Zekić-Sušac, M.; Has, A. Data Mining as Support to Knowledge Management in Marketing. *Bus. Syst. Res.* **2015**, *6*, 18–30. [CrossRef]
79. Zhai, C.; Velivelli, A.; Yu, B. A cross-collection mixture model for comparative text mining. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 743–748.
80. Zhai, Y.; Hsu, A.; Halgamuge, S.K. Combining news and technical indicators in daily stock price trends prediction. In Proceedings of the 4th International Symposium on Neural Networks (ISNN), Nanjing, China, 3–7 June 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 1087–1096.
81. Zhang, C.; Zhang, P. *Predicting Gender from Blog Posts*; University of Massachussetts: Amherst, MA, USA, 2010.
82. Zhang, D.; Zhou, L. Discovering golden nuggets: Data mining in financial application. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2004**, *34*, 513–522. [CrossRef]
83. Zhang, L.; Wang, S.; Liu, B. Deep Learning for Sentiment Analysis: A Survey. 2018. Available online: https://arxiv.org/abs/1801.07883/ (accessed on 12 August 2018).
84. Zhao, D. Frontiers of big data business analytics: Patterns and cases in online marketing. In *Big Data and Business Analytics*; Leibowitz, J., Ed.; CRC Press: Boca Raton, FL, USA, 2013; pp. 46–68.