information

MDPI

Article

UGC Knowledge Features and Their Influences on the Stock Market: An Empirical Study Based on Topic Modeling

Ning Li [1,*], Kefu Chen [2] and Huixin He [3]

[1] School of Communication, Fujian Normal University, Fuzhou 350007, China
[2] Management School, Xiamen University, Xiamen 361000, China
[3] Computer Science and Engineering School, Huaqiao University, Xiamen 361021, China
* Correspondence: lining@fjnu.edu.cn

Abstract: According to the natural language perspective, UGC has been significantly used for the screening of key nodes in knowledge discovery and strategic investment. This article presents a new research framework that is proposed for the decomposition of UGC knowledge feature extraction into topic recognition and language analysis, mainly. For visual analysis of associated topics, the LDAvis approach is utilized. Then, risk features of UGC knowledge are assigned according to language attribution. Based on previous studies, the risk attribute lexicon is further updated by judging semantic distance through word vectors. This research uses platform data and individual stock data as samples for subject recognition and knowledge feature extraction. A regression model is constructed based on the panel data after natural language processing to verify the feedback effect of the market at strategic risk measurement. It can be found from the conclusion that the change in market behavior is regular and correlates with the change in the UGC risk degree of individual stocks. The purpose of this paper is to examine the value of UGC in investment decision-making from the perspective of knowledge discovery. The research content can provide a reference for data mining, fintech, strategic risk monitoring, and other related works.

Keywords: UGC; natural language analysis; LDA topic model; knowledge features; market impact

check for updates

Citation: Li, N.; Chen, K.; He, H. UGC Knowledge Features and Their Influences on the Stock Market: An Empirical Study Based on Topic Modeling. *Information* **2022**, *13*, 454. https://doi.org/10.3390/info13100454

Academic Editor: Andrea Giovanni Nuzzolese

Received: 16 July 2022
Accepted: 20 September 2022
Published: 27 September 2022

1. Introduction

The development of the Internet has changed the way people learn knowledge. The way news is produced has also become more inclusive [1,2]. A massive amount of User-Generated Content (UGC) is the main way of content production for many social media platforms [3,4]. The application of UGC in industrial practice involves many fields such as economics, brand marketing, government management, and so on [5,6]. Under the strategic investment background, to make decisions, investors look to public company news and user-generated opinions. Therefore, discovering knowledgeable information from social media user-generated content has been important to make effective strategic investments.

In the study of social learning and decision-making behavior, how to sufficiently acquire knowledge and how to read the market's feedback on such knowledge are of great significance to the screening of nodes in technological exploration. With the development of social media, the investor social network has rapidly become a channel for the dissemination and exchange of investment knowledge. More targeted professional service platforms have emerged. Among them, the emergence of Investor Based Social Network (IBSN) not only enables independent investors to express their own investment opinions but also enables financial researchers and analysts to collect a large number of investors' ideas [7]. IBSN users include senior investors, financial experts, consultants, and other professionals, as well as company employees or common shareholders. These users' tacit knowledge includes keen judgment on market orientation and company changes. Therefore, the original analysis reports, reposts, and replies published by these users have an important research value.

However, UGC in finance is different from other areas. Firstly, UGC from the financial industry includes not only public word of mouth about the company's brand and products, but also includes decisions shared by experienced investors. At the level of strategic management, the requirements of UGC knowledge feature extraction need to be improved. In addition to extracting effective features, more research should be done to evaluate whether the contents contain high-risk or low-risk information. This requires an additional identification of the risk attributes contained in the information content. Secondly, from the UGC communication effect, some UGC content in investors' social networks has a strong communication effect, especially for listed companies with higher risks. Investors are more inclined to pay attention to such content [8]. Therefore, this paper puts forward the following research question:

Question 1: In investor social networks, what kinds of strategic references does UGC provide for listed companies?

Question 2: How can we evaluate the market feedback on UGC knowledge features?

Based on previous studies, this paper puts forward a new research framework. The framework is a systematic process of topic identification, risk assessment for knowledge features, and market response to UGC. The main contribution of this paper is to transform the unstructured content of UGC into structured content with a risk assessment label and to test the effective feedback from the market by empirical study. The empirical conclusion verifies the feedback effect of the market on the knowledge characteristics of UGC. This also indicates that the market will be more affected by UGC in IBSN within the medium (10 days) and short-term (5 days), depending on the types and the structures of UGC. This research can provide a reference for data mining, fintech, securities investment, and other related works.

2. Related Studies

2.1. User-Generated Content, UGC

With the development of text mining and natural language processing technology [9], there are more advanced processing technologies for document language analysis. For example, text clustering and topic model [10], etc. With the penetration of social media in all walks of life, text analysis based on user-generated content has gradually attracted the attention of scholars. Among them, the topic model method can better explain the dependent variables, which is helpful to identify user portraits, user behaviors, and brand characteristics. Therefore, the research on the modeling algorithm of user-generated content has been widely used in both management science and social science [11,12]. The existing research is based on media news, Weibo blogs, and online comments, which are processed by natural language, and then combined with relevant theories, the user-generated content is tagged and analyzed, to better provide decision-making support services to the government and users [8,13].

In order to highlight the research focus, the users involved in this article are mainly investors and financial enthusiasts. The generated content of platform users is mainly evaluation postings and comments from listed companies and financial markets. UGC involved in this paper includes two types: long text UGC and short text UGC. Among them, long texts are mainly investment analysis postings, while short texts are mainly stock talks. Investment analysis postings, because of their long length (usually more than two pages) and regular format (including title, abstract, objective statement, and accurate prediction results) are also called original articles on the platform [14]. IBSN has high requirements for investment analysis postings, and it must meet the rigor and readability standards before it can be released. Such postings have to be reviewed by platform editors to determine their professionalism and novelty. Stock talks are defined in this paper as comments made by users on original content on social platforms, or personal opinions and posts published in BBS forums of listed companies. Registered users can post Stock talks without editing and auditing.

2.2. Topic Modeling Analysis

Latent Dirichlet Allocation (LDA) was proposed by Blei in 2003 [10], which added a layer of Dirichlet conjugate prior distribution based on the PLSI model [15]. In recent years, some evolutionary models have emerged. For example, the LDAvis model has been recognized in the improvement research of explaining the relationship between the main topic and the sub-topic. The LDAvis improves the representation function of LDA and uses phrases instead of words to represent a topic. Moreover, LDAvis also maps the topic recognition results to a two-dimensional space based on a multidimensional scaling algorithm and then reveals the relationship between topics and words [16]. In recent years, the LDAvis model has shown high-quality thematic analysis results in the empirical research of data mining [17]. So far, scholars have put forward rich usage scenarios and corresponding optimization methods for LDA, which can be used for topic recognition of user-generated content in different scenarios. For example, scholars have improved and optimized the classical LDA topic model by using additional information such as corpus optimization, time window sliding mode, and users' social network as prior variables or observation variables [18–20]. Combined with the embedding of potential new variables, such as identity, emotion, psychology, and perception [13], the model is conditioned [21]. Generally speaking, the topic model is widely used in the research of unstructured data in various fields. Therefore, based on the LDAvis and the optimized LDA model, this paper analyzes the language of user-generated content in social networks based on investors. Furthermore, it verifies the reliability of the algorithm results by combining the analysis of empirical models and actual cases.

2.3. Knowledge Discovery and Knowledge Feature Measurement

Knowledge Discovery Databases (KDD) have become an important subject in modern management science [22]. KDD is a complex process of identifying effective, potentially useful and understandable patterns from data sets. The general KDD system framework includes four steps: data preparation, data cleaning, data mining, and data verification. At present, there are many kinds of KDD research, based on investors' social networks, studying user portraits, recommendation systems, cold start problems, and sales forecasts [7,13,23]. However, there are still many research areas to be explored in investor-based social networks. For example, in the processing and application of unstructured texts, more advanced technologies and novel research perspectives need to be used in investment risk identification, portfolio recommendation, market public opinion early warning, and enterprise strategic planning. From the usefulness of investment decisions, these studies are the best way to further explore [8]. In particular, language analysis [12], which pays more attention to the actual effect of natural language processing in the financial field, is increasingly applied to the research of decision management and strategy execution. The research of knowledge discovery in IBSN study is of great significance to strategic decisions and investment decisions [24,25]. In listed companies, to avoid possible litigation risks in the IPO process, the directors of the company will choose to disclose the risks and reorganize information in narrative materials. IBSN users' tacit knowledge includes keen judgment on market orientation and company changes. Therefore, the knowledge discovery study of the UGC of IBSN platform is more helpful to discover this tacit knowledge.

Prollochs et al. identified the topics of news and financial reports issued by listed companies, and are further manually screened at 20 × 20 high-frequency words, aiming at summarizing the topic names and combing the stems of risk attributes [12]. To ensure the consistency and integrity of information processing, our research uses Prollochs and other research results for reference, processes each document with natural language, further assigns risk attributes to the processing results, and finally, automatic clusters feature vectors through the topic model. The language analysis of UGC of investors' social platform is based on the acquisition of big data and natural language processing, and then deduces the membership results of documents from the word attributes. Combined with the judgment of the semantic distance of the topic model and word vectors, the UGC content and the stock price of listed companies are made into panel data; thus, realizing the empirical analysis after the correlation between the risk attributes of listed companies implied by UGC and the actual changes of stock prices.

3. Research Methods

One of the research purposes of this article is to verify that UGC contains certain wisdom, which can get feedback from financial markets. Different from the operation process of the classic KDD, this study framework pays more attention to the feedback effect of the fluctuation of the excess return of the listed company's stock price on the fluctuation of the strategic risk evaluated in UGC content in data verification. As shown in Figure 1, this framework includes three processes: UGC data acquisition and natural language processing, LDA topic recognition and visualization, and language analysis and empirical verification.

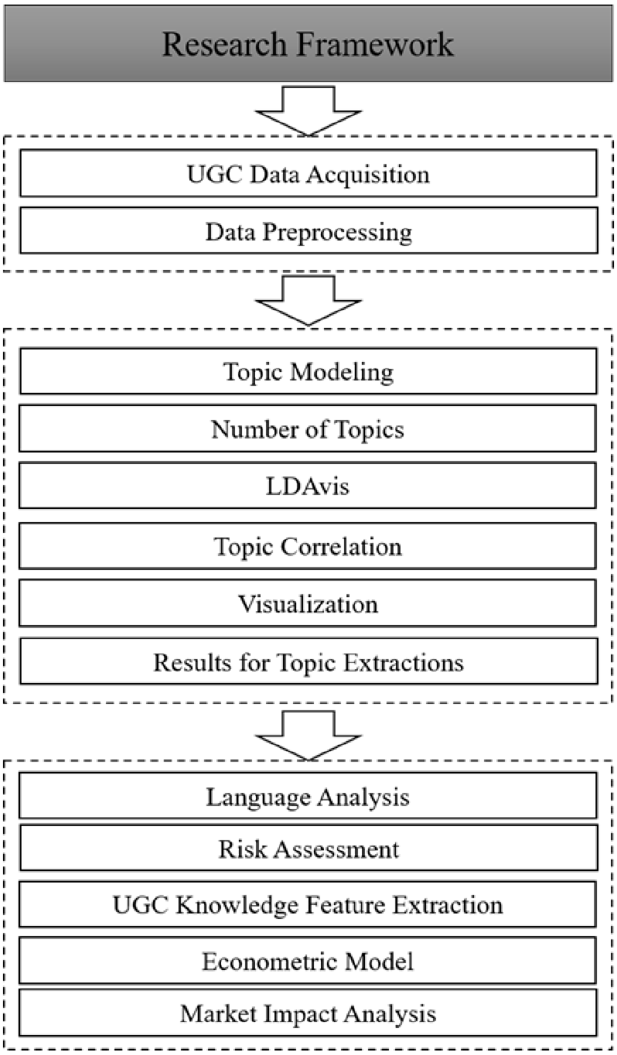


Figure 1. Research frame diagram.

This research framework has some advantages: first, it provides an overall analysis of the strategic position of the market; secondly, it provides a quantifiable measure of the strategic risk features contained in the topic; and finally, the experimental results are suggestive and can detect some subtle viewpoints and perceptions.

3.1. UGC Data Collection and Natural Language Processing

In this process, we will extract key information and store it locally for future research. See Section 4.1 The Data Source and Data Preprocessing for a detailed explanation of data acquisition and data processing. Conventional processing of natural language includes six steps:

Step 1: Exclude the following samples: articles, reviews, and posts from companies not listed on the US stock market; samples with missing data on relevant variables; a sample of non-English expressions.

Step 2: To the ambiguity of unstructured data, at the preprocessing stage the sample was spell-checked to eliminate the content of spelling errors. A stopword list is also added, the list contains numerical content, expression characters, and other characters. Anything in the sample that is consistent with the list of stopwords will be eliminated.

Step 3: The reduction in some words omits no deeper meaning of the stop words.

Step 4: The quantification of lexical semantic relatedness has many applications in NLP, and many different measures have been proposed [26]. All of them use WordNet as their central resource [27]. A knowledge base in the form of WordNet's lexical relations is used to automatically locate training examples in a general text corpus [28]. In this article, the title and abstract fields in the original analytical articles of UGC are extracted, and WordNet is used as the semantic-oriented English Dictionary.

Step 5: Dictionary operation: Through Python natural language processing, "gensim" is called to convert the document into vector mode according to the LDA model. The document set in "gensim" is expressed in the form of corpora, which is essentially a two-dimensional matrix format. In the actual operation, the number of words is very large (tens of thousands or even 100,000), and the number of words in a document is limited, so it will cause a great waste of memory using a traditional dense matrix. After a document is partitioned into words, a dictionary is generated using "dictionary = corpora. Dictionary (texts)". The "save" function can then be used to persist the dictionary.

Step 6: Use the pickle tool to train the corpus and save the generated lexicon in the document. Spacy is an industrial-level Python natural language processing tool [29]. Spacy makes extensive use of Cython to improve the performance of related modules, so it has practical application value in the industry [30]. This article uses the functions of word tokenize in spacy, including sentence breaking, stem extractions, and part-of-speech tagging. The main goal is to restore English word forms so that they can be better used in machine learning.

3.2. LDA Topic Recognition and Visualization

LDA describes a three-tier Bayesian network [31]. φ_k represents the polynomial distribution of the kth topic on each word. ϑ_m represents the topic distribution of document m. According to the topic $z_{m,n}$, the vocabulary $w_{m,n}$ is obtained from its corresponding distribution sampling. For document m, we can get the joint distribution of its whole generation process:

$$\begin{aligned} &p(w_m, z_m, \vartheta_m, \varphi|\alpha, \beta) \\ &= \prod_{n=1}^{N_m} p(w_{m,n}|\varphi_{z_{m,n}})p(z_{m,n}|\vartheta_m)p(\vartheta_m|\alpha)p(\varphi|\beta) \end{aligned} \quad (1)$$

The ultimate goal of the algorithm is to estimate the parameters, to obtain the topic distribution of each document and the distribution between each topic and vocabulary. To estimate φ and ϑ, we use Gibbs Sampling, that is to sample one component of joint distribution at a time, and keep the other components unchanged.

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t} \quad (2)$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k} \tag{3}$$

The number of potential topics is an important decision in topic modeling. In the aspect of evaluating the effect of the topic model, perplexity and coherence are the two explanatory evaluation indexes [32]. The coherence index is calculated by “model. coherence” function in the Python NPL system. According to literature, the coherence index has a better evaluation effect than the perplexity index [33]. The basic idea of the perplexity calculation is to estimate the fitting degree of the test set corpus according to the distribution between the topic and the words. The calculation formula is as follows:

$$P(\widetilde{W}|M) = exp(-\frac{\sum_{m=1}^{M} lgp(\overrightarrow{\widetilde{w}}_{\widetilde{m}}|M)}{\sum_{m=1}^{M} N_m}) \tag{4}$$

In Formula (4),

$$\mathrm{lg}p(\overrightarrow{\widetilde{w}}_{\widetilde{m}}|M) = \sum_{t=1}^{V} n_{\widetilde{m}}^{(t)} lg\left(\sum_{k=1}^{K} \varphi_{k,t} \cdot \vartheta_{\widetilde{m},k}\right) \tag{5}$$

The lower the value of the perplexity degree, the better the effect of LDA, and the better the topic distribution model fit the training set data. However, the perplexity index cannot visually compare the results of multiple topics. To solve this problem, evaluation methods such as LDAvis, color visualization, and topic content distribution map can provide a visual comparison of topics [34]. In order to show the topic distribution and word distribution, this section chooses LDAvis as the tool for topic visualization.

A reliable and readable topic model should distinguish different topics as clearly as possible. The LDAvis is a visualization tool for the dynamic clustering relationship of LDA model results [34]. The main parameters involved in this method include topic distance, topic meaning, and topic number. Topic distance refers to the semantic distance between topics on the topic model. This definition can effectively distinguish different topics and represents the description of the “document-topic” relationship with the original data. Secondly, thematic meaning refers to the validity of each thematic meaning in the thematic model, and it is a description of the relationship between “subject words”. Generally speaking, nouns can better describe the thematic meaning than verb words, and specific words can better describe the thematic meaning than general words. Therefore, the differences in vocabulary related to each topic in the model represent the overall meaning of subtopics. Generally speaking, the advantage of the visualization method is that you can see the amount of data contained in each topic, the degree of correlation between topics, and the keywords of each topic. Therefore, the visualization results can explain an observation phenomenon more intuitively. However, visualization tools also have shortcomings, that is, they can’t evaluate the topic classification and extract features from the numerical point of view. Therefore, based on the second stage of research, this study will further develop the third stage of the financial risk language evaluation model, solve the quantitative problem of subject feature extraction, and provide a measurement basis for the follow-up empirical test.

3.3. Language Analysis and Empirical Verification

The third step is language analysis and the empirical effect test. Language analysis measures the usage of terms with different risk dimensions, and mainly adopts a rule-based method to calculate the frequency of corresponding terms according to a predetermined word list [35]. Thus, obtaining a method with high computational efficiency, less supervision, and strong explanatory power. This method is reliable and robust, so it is usually used to mine text data from accounting data.

$$R_d = \frac{number\ of\ risk\ words}{total\ number\ of\ words} \tag{6}$$

R_d represents the risk polarity of a document. In this research, we measure the risk characteristics of each document. In addition to the Formula (6), R_d can also be calculated by weight. In this paper, the two results are compared with the experimental results. The results show that the results calculated by the Formula (6) are more relevant to the market and can be reflected by the stock market. In addition, only a few words are marked as risk-related expressions for some documents, leading R_d closer to zero. Therefore, for this kind of document, follow the principle of standardization, and standardize the corresponding scores to zero mean and standard deviation of 1, to make a better comparison.

Based on previous research results, 176 root words representing risks were obtained. For sample-based machine learning, this study tries to find out more words representing risks based on these 176 root words by the machine learning method of word vector distance [12]. The method comprises three specific steps: step one, extracting the roots of new words output after machine learning, and sorting them out as new investors' social platform users to generate a content risk attribute root dictionary. The second step is to sort out these roots, check them manually and analyze them by experts, eliminate spoken language, slang, month, numbers, non-risk words, and so on, and then re-output the results. The third step, as shown in Appendix A Figure A1, is to get the final risk root dictionary, which is suitable for this study, after removing the repeated values within and between groups. To test the practical significance of this research, in the third step, we will also observe the influence of feature extraction results on stock price changes in different time windows through empirical analysis. See Section 5 Stock Market Impact for the regression model and the empirical results.

4. Empirical Research

4.1. The Data Source and Data Preprocessing

Based on the research objective, this paper takes Seeking Alpha (hereinafter referred to as SA, http://seekingalpha.com, accessed on 12 February 2020) as the data source. Web crawler is used to obtain reliable data from the SA website, and the data from July 2018 to December 2018 are selected as samples. SA is an investment research website based on freelance writers discussing financial markets' movements and sharing investment experiences. There are three main rentals for investors to express their opinions. First, users can submit investment analysis postings to SA, as shown in Appendix A Figure A2a. In the following paragraph, we use the UGC Article as the measurement variable of user-generated investment analysis postings. The contents of the investment analysis postings mainly include title, company, summary, analysis results, and user information. In the following paragraph, we use the UGC Article Title as the measurement variable of user-generated investment analysis postings titles, and use the UGC Article Summary as the measurement variable of user-generated investment analysis postings summaries. Anyone can post on the SA, but it takes one day to review the user-generated investment analysis postings. In the IBSN platform, registered users can comment on the other users' postings or discuss freely in the Stock Talk Forum, as shown in Appendix A Figure A2b. Stock talks have the advantages of timeliness, but their disadvantages are loud noise and insufficient rigor [36]. Due to the different posting habits of users, some incomplete expressions often appear in talks. However, IBSNs usually automatically tag related company names and user investment behaviors to stock talks. In the following paragraph, we use UGC Stock Talk as the measurement variable of user-generated stock talks. Matching tags and stock talks can be helpful to classify incomplete expressions and improving the readability and credibility of NLP results.

Financial data comes from Wharton Research Data Services (WRDS) platform, which integrates Compustat, CRSP, TFN (THOMSON), TAQ and other famous database products. WRDS platform is the leading business intelligence, data analytics, and research platform to global institutions. This article downloads the stock index and the relevant financial data of listed companies from the WRDS-CRSP database [1]. The research sample covers 2, 996 listed companies, 10, 386 UGC Articles, and 125, 247 UGC Stock Talks before data cleaning and data standardization.

According to the process of step 1, step 2 and step 3 in Section 3.1 UGC data collection and natural language processing, the result is 22, 192 observations. The corpus used covers 2, 996 different companies operating in a wide range of industries. For original articles, the average length per article has 1, 897 words. To eliminate the effect of extreme values, Winsorize adjustments are made for all continuous variables at the 1% and 99% quantiles.

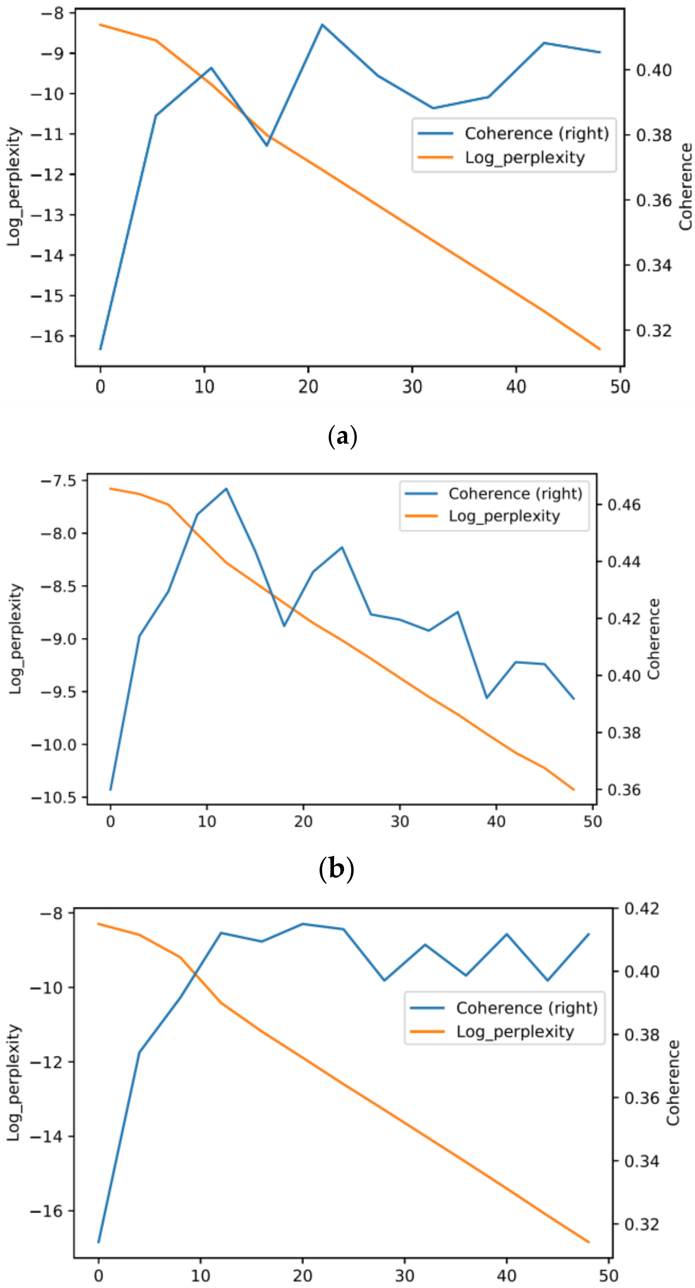
4.2. Topic Identification

Selecting the number of topics to be identified is the most important task in the number of topics output algorithms. According to the content described in Section 3.2 Formulas (4) and (5), perplexity and coherence are used to select the number of topics. Coherence was used to calculate the coherence index.

According to the results in Figure 2, it can be seen that: (1) As the number of topics increases, the degree of confusion of title summaries and stock talks gradually decreases. The confounding degree of the summary is the highest, while the confounding degree of the UGC Article Title and the UGC Stock Talk is lower than that of the UGC Stock Summary, and the confounding degree of the latter two is similar. (2) The topic coherence of the UGC Stock Talk gradually becomes stable after reaching the optimal level, while the topic coherence of the UGC Article Title gradually becomes stable after reaching the optimal level. However, the coherence of the UGC Stock Summary consistently decreases after reaching the optimal level. Therefore, from the perspective of coherence, the coherence of the UGC Stock Summary is relatively low. (3) The UGC Article Title content reached the maximum from K = 20 to K = 25 and showed a declining trend and gradually stabilized. The content of the UGC Stock Talk reached the maximum from K = 10 to K = 15 and showed a declining trend and gradually stabilized. Combined with Figure 2 and the explanation of the perplexity and coherence index in Section 3.2, the results support the establishment of the optimal number of topics extracted for the UGC Article Title and the UGC Stock Talk. In the following sections, we will focus on the empirical results of these two types of content.

The above results can also be verified through a topic visualization tool. Set the number of topics K to 20, β to 0.01, then run LDAvis and get the results in Figure 3. The size of each circle represents the number of corpuses related to the topic in all corpus. If the topic nodes are distributed intensively and there are many overlapping parts, it means that the topic similarity is high and the output quantity needs to be adjusted. In the contrast between Figure 3a,b, there is a more uniform distribution of the topic of the UGC Article Title, and the topic of the overlap is less. That is because the UGC Article Title is a refined expression of the key information of the document, with a short length and strong referent. In the contrast between Figure 3a,c, there are large areas of overlap in Figure 3c between topics 14 and 20, topics 8 and 10, topics 3 and 16, topics 5 and 11. The number of stock talks topics should be adjusted to reduce the output quantity. Appendix A Figure A3 shows the visualization result of the UGC Stock Talk when the number of topics is set to 13. At this time, the distribution of topic nodes is uniform and the overlap part is less.

Based on the above evaluation results of the number of topics, in the following sections, we will focus on the empirical results of the UGC Article Title and the UGC Stock Talk. To ensure the reliability and rigor of the results, we check the output topics and keywords in the form of a manual review. The review process is divided into two steps. First, two graduate students majoring in finance are invited to check the consistency between keywords and subject content. Secondly, five financial experts are invited to name and grade the subject.



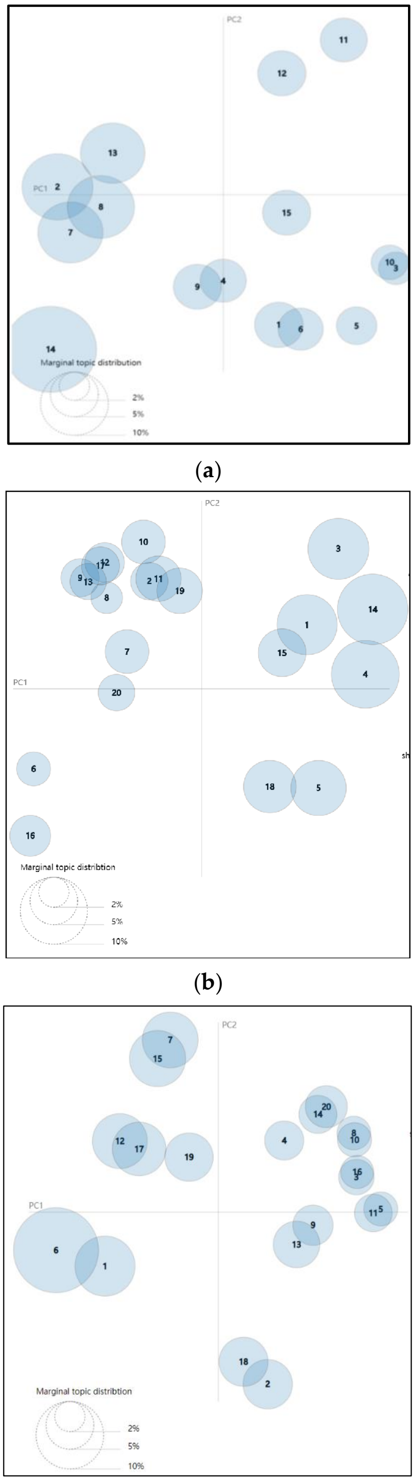
(**a**)

(**b**)

(**c**)

Figure 2. Perplexity and coherence of UGC Topic Number K. (**a**) Perplexity and coherence for UGC Article Title. (**b**) Perplexity and coherence for UGC Article Summary. (**c**) Perplexity and coherence for UGC Stock Talk.

From the results, we choose the one with a higher score as the name of the topic. So far, topic extraction results include 20 topics from the UGC Article Title and 13 topics from the UGC Stock Talk.



(**a**)

(**b**)

(**c**)

Figure 3. Visualization distribution of LDAvis themes (K = 20). (**a**) The UGC Article Title. (**b**) The UGC Article Summary. (**c**) The UGC Stock Talk.

4.3. Topic Extraction Results

Tables 1 and 2, respectively, display the topic names, keywords, and company names with a high frequency under each topic. Due to the limitation of table length, the keywords listed in Table 1 are used to reflect the top six words with the highest weight distribution of the words, and the keywords listed in Table 2 represent the top ten words with the highest weight distribution of the stock talk words.

Table 1. Topic Extractions for UGC Article Title.

	Topic Name	Topic%	Keyword Stems	Risk Assessment	High Frequency Company
1	Financial attractive	5.8	finance, attract, cheap, mine, exposure, bioscience	0.004890	Microsoft, Tencent
2	Therapeutics report	4.5	therapeutic, Boeing, drive, podcast, airline, Airbus	0.002189	Boeing
3	Earnings	5	earn, analyze, Amazon, promos, approve, valuat	0.000044	Amazon, McMoran
4	Undervalued company	5.1	undervalue, busy, develop, revenue, growth, lithium	0.041475	AstraZeneca
5	Tesla model	4.7	Tesla, China, electro, central, expect, model	0.000049	Tesla
6	Share price	5.5	price, share, target, catalyst, selloff, solar	0.000660	Starbucks
7	Trade	5.3	group, trade, Micron, solid, quarter, wrong	0.065879	Micron
8	Biotech-Pharma	5	pharma, posit, dividend, ready, Abbvie, biotech	0.000046	Abbvie, Netflix
9	Health industry	4.4	industry, pharmaceut, global, better, start, health	0.008343	Intel
10	China and America	5.1	update, future, money, resources, America, Chinese	0.010947	Progenics
11	Technology growth	5.3	potent, growth, technology, history, partner, remain	0.019210	Merck
12	Holding capital	4.8	upside, ahead, hold, copied, return, term	0.005370	Vertex
13	Investment opportunity	6.8	opportune, value, buy, offer, trade, point	0.001507	
14	Energy growth in China	4	energy, review, growth, reward, China, rais	0.000059	
15	International healthcare sales	4.8	invest, reason, sale, intern, Alibaba, science	0.003980	Alibaba, Gilead
16	Strong brand	6	strong, Facebook, result, brand, growth, Chespeake	0.000035	Facebook, Chespeake
17	Portfolio	4.8	portfolio, posit, growth, acquit, medic, become	0.006085	Pinduoduo
18	Forecast	4.7	continue, short, look, higher, product, pipeline	0.000230	
19	Biotech-Tech	4.8	Apple, biotech, great, thing, investor, grow	0.000094	Apple
20	Legal issue	3.6	momentum, Nvidia, profit, Sonos, weak, expands	0.006453	Nvidia, Sonos

Table 2. Statistical Table of Topic Information of UGC Stock Talk.

	Topic Name	Topic%	Keyword Stems	Risk Assessment
1	Cancer immunotherapy	4.0	immune, cancer, kill, milestone, buyout, index, Canadian, ship, mutual, agreement	0.012008
2	Bitcoin investment	9.5	right, value, investor, shareholder, bitcoin, board, portfolio, market, current, crypto	0.004822
3	Video technology	4.7	video, high, technic, signal, Youtube, close, target, report, California, potenti	0.001796
4	Earnings & Dividends	7.5	earn, dividend, yield, growth, buy, estimate, expect, equity, bottom, winner	0.007154
5	Stock price movement	15.5	share, price, market, go, posit, short, trade, increase, rais, higher	0.017160
6	Technology announcement	13.1	busy, system, Saudi, technology, industry, problem, base, announces, potency, project	0.019387
7	Stock consultant analysis	5.3	analysis, stock consult, watches, support, breakout, bound, bullish, strong, rang, stat	0.000011
8	Oil stocks under natural environment	6.3	expect, price, bullish, crude, forecast, analyst, storm, product, bearish, hurricane	0.019863
9	Brand contribution	3.1	income, swingstocktrad, figure, update, user, really, brand, risk, Germany, selloff	0.010922
10	Financial derivatives	5.5	trade, shareplann, report, swing, strategi, call, option, video, finance, good	0.002698
11	Social media market opportunity	5.5	revenue, video, Facebook, margin, study, Twitter, anyone, sale, friend, guidance	0.000025
12	China-U.S. trade tariffs	12.2	China, heisenbergreport, Trump, market, trade, tariff, economy, interest, rate, break	0.022065
13	The algorithm-based stock trading pattern	7.9	level, trade, market, chart, profit, start, move, earn, growth, pattern	0.009371

According to Table 1, the distribution of 20 UGC Article Title topics is relatively uniform, and the proportion of each topic is relatively close. Due to the rigorous and professional requirements of the UGC investment analysis postings, most topics involve at least one listed company name with high frequency. Company names usually appear in the form of company abbreviations, which makes it possible to identify companies' nodes in natural language processing. For example, AstraZeneca is the company name with the highest frequency under the fourth topic (Undervalued Company). The contents listed in the last column of the table are the top two company names with the highest frequency in each topic.

According to Table 2, the distribution of 13 UGC Stock Talk topics shows a high and low trend based on their proportion distributions. For example, top three topics with the highest distribution are #5 Stock price movement topic, #6 Technology announcement topic, and #12 China-U.S. trade tariffs topic. Different from the UGC Article Title, the frequency rate of company names in stock talks is lower than other keywords, so the names of listed companies are not shown in Table 2.

By comparing Tables 1 and 2, three aspects can be discovered as follows.

(1) Topic proportion distribution difference. Generally, investment analysis postings have length requirements and the posting format is strictly required. Such postings look more like analytical articles, too long or too short will be rejected. The postings should not be too oral and should meet a certain professional writing level and financial analysis ability. Users must agree to the disclosure standards and the editing services. Platform editors will not revise users' ideas, but will polish the titles, abstracts, and texts. Comparing the investment analysis postings, there are no format requirements for the stock talks. Repeated words and expressions often appear in such talks.

(2) Topic issue difference. The topics of the UGC Stock Talk reflect the hot issues that ordinary users are concerned about. The stock talk contents are more specific and more direct. In contrast, the topics of the UGC Article Title reflect the strategic issues that the whole stock market is concerned about. The investment analysis posting contents are more macro and more comprehensive.

(3) Keyword stem difference. This article uses the functions of word tokenize in spacy, including stem extractions and part-of-speech tagging. Therefore, keywords in Tables 1 and 2 are presented in the form of stems. In Table 1, many high-frequency stems, such as "portfolio", "growth", and "invest", appear in multiple topics at the same time. These keyword stems show that the style of the article conforms to the professionalism in the field of strategic management and the concise expression of the UGC Article Title. In Table 2, there are fewer repeated keywords in the UGC Stock Talk topics. It is not difficult to find that there are more oral expression stems under the UGC Stock Talk keywords, such as bullish, bounce, bearish, stats, etc., which can reflect the "talk" attribute of the forum.

4.4. Language Analysis Results

As for the risk assessment of each topic, the risk rating range is [0, 1]. The closer the rating result is to 1, the higher the risk knowledge expressed by this topic; the closer the rating result is to 0, the lower the risk knowledge expressed by this topic, or the higher the knowledge level of market optimism.

Based on the output results of the topic model, the risk weights of keywords in each topic are calculated. The calculated results are used as the risk assessment values for each topic. For example, in Table 1, #4 Undervalued Company topic has a high risk among the 20 UGC Article Title topics. In its thematic sense, Undervalued Company expresses that the share prices of above-market companies are below their intrinsic value, that is, they are undervalued. Therefore, from the perspective of strategy management, the problem of stock price deviation from value belongs to a high-risk feature.

As mentioned in Chapter 1 Introduction, UGC in finance is different from other areas. UGC from the financial industry can also reflect the public word of mouth about the brand and products. For example, the risk rating of the 16th topic (Strong brand) in Table 1 is shown as the low-risk topic among the 20 UGC Article Title topics. In the thematic sense, this topic expresses the strength and advantages of the brand. In general, brand strength advantage directly represents consumers' trust in the brand or company, and investors choose an optimistic attitude towards the listed companies with strong brands. Therefore, brand strength advantage belongs to a low-risk feature.

After transforming the unstructured content of UGC into structured content, language analysis assigns risk assessment value to each topic. Next, to answer Question 2—(How can we evaluate the market feedback on UGC knowledge features?), further regressions will be needed for empirical tests on the basis of panel data.

5. Stock Market Impact

5.1. Market Feedback on Risk Attributes in UGC Knowledge Features

We use this section to discover the market feedback on UGC knowledge. Based on the results of Tables 2 and 3, we transform the language features of UGC unstructured data into structured data and merge them with the companies' financial data to form panel data. The basic idea of the regression model is to compare the change rule of the fluctuation of stock returns before and after the release date of UGC. The change rule will be helpful to discover the strategic risks of listed companies and to provide a reference for investors' decision-making.

Table 3. Impact of the UGC Article Title knowledge features on Stock Price.

	Dependent Variable									
	($ARet_{i,t+10}$)		**($ARet_{i,t+5}$)**		**($ARet_{i,t+3}$)**		**($ARet_{i,t+2}$)**		**($ARet_{i,t+1}$)**	
	Coeff.	**t**	**Coeff.**	**t**	**Coeff.**	**t**	**Coeff.**	**t**	**Coeff.**	**t**
Constant	−0.005 **	−3.31	−0.005 ***	−2.61	0.002	0.88	0.001	0.36	−0.005 **	−2.34
$UGCArticle\_k_{i,t}$	−0.217 **	−2.37	−0.109	−0.43	−0.165	−0.66	−0.294	−1.08	−0.258	−0.95
MONTH	Yes		Yes		Yes		Yes		Yes	
Observations	4376		4376		4376		4376		4376	
R^2	0.003		0.003		0.0002		0.0003		0.003	
Adjusted R^2	0.002		0.002		−0.0003		−0.0002		0.003	
Res. Std. Error (df = 21,961)	0.048		0.056		0.055		0.059		0.059	
F statistic (df = 8; 21,961)	6.167 ***		5.665 ***		0.356		0.582		6.946 ***	

Note: ** $p < 0.05$; *** $p < 0.01$.

According to the literature, Abnormal Return is used to measure stock return [37,38]. In this article, the short-term and medium-term stock market influence within ten days is considered, so the window period is 1 to 10 trading days after the document is released.

By establishing a simple regression model, this paper tests the response of the stock market to subject knowledge. $ARet_{i,t}$ indicates the stock price excess return of company i on the day of the event. $ARet_{i,t+j}$ indicates the stock price excess return after the event j trading days $(j = 1, 2, 3, 5, 10)$. $UGCArticle\_k_{i,t}$ represents the mean value of risk knowledge characteristics expressed by all article titles of company i on tth day. $UGCTalk\_k_{i,t}$ represents the mean value of risk knowledge characteristics expressed by all forum discussion posts of company i on tth day. The feature values are calculated according to the product of the weight of the topic to which each document belongs and the risk value of the topic. $Month$ is the virtual variable and $\varepsilon_{i,t}$ is the residual term. The regression model is shown below, and the regression results are shown in Tables 3 and 4.

$$ARet_{i,t+j} = \alpha_0 + \alpha_1 UGCArticle\_k_{i,t} + \alpha_2 Month + \varepsilon_{i,t} \tag{7}$$

$$ARet_{i,t+j} = \beta_0 + \beta_1 UGCTalk\_k_{i,t} + \beta_2 Month + \varepsilon_{i,t} \tag{8}$$

Table 4. Impact of the UGC Stock Talk knowledge features on Stock Price.

	Dependent Variable									
	($ARet_{i,t+10}$)		**($ARet_{i,t+5}$)**		**($ARet_{i,t+3}$)**		**($ARet_{i,t+2}$)**		**($ARet_{i,t+1}$)**	
	Coeff.	**t**	**Coeff.**	**t**	**Coeff.**	**t**	**Coeff.**	**t**	**Coeff.**	**t**
Constant	0.002	0.94	0.002	0.72	0.004	1.59	0.0004	0.16	0.002	0.63
$UGCTalk\_k_{i,t}$	−0.810 ***	−3.28	−0.781 ***	−3.25	−0.660 ***	−2.78	−0.336	−1.42	−0.323	−1.28
MONTH	Yes		Yes		Yes		Yes		Yes	
Observations	4881		4881		4881		4881		4881	
R^2	0.002		0.002		0.002		0.001		0.0004	
Adjusted R^2	0.002		0.002		0.001		0.0001		−0.00004	
Res. Std. Error (df = 21,961)	0.071		0.069		0.068		0.068		0.072	
F statistic (df = 8; 21,961)	5.377 ***		5.899 ***		3.963 ***		1.327		0.909	

Note: *** $p < 0.01$.

The results in Table 3 show that the knowledge characteristics of the UGC Article Title have an impact on the stock market products on the 10th day after the publication of the article. The t-value is −2.37, which has a significant impact on the results, and the correlation coefficient is −0.217, indicating that the risk characteristic strength of the UGC Article Title will restrict the stock excess return. The analysis shows that the main reason is that risk knowledge represents investment risk, and investors should be careful to choose the company's stock investment in the future. Therefore, from the perspective of the impact of investment risk on stock price, the characteristics of risk knowledge will lead to the reduction in excess return or even negative value. Table 4 shows that the knowledge characteristics of the UGC Stock Talk began to have an impact on the stock market on the third day after the post was published. On the 10th day after the publication, the effect was significant, the t value was −3.28, and the correlation coefficient was −0.8910, indicating that the risk characteristics of the UGC Stock Talk had a stronger restriction on the stock excess return 10 days after the posting. The main reason for the analysis is that the content released by the stock talk forum does not need to be reviewed, so it has strong timeliness and can have an impact in a short time. With the increase in the number of comments and the heat of discussion, the degree of influence will be strengthened. By comparing the influence of the UGC Article Title and knowledge characteristics of the UGC Stock Talk on stock price, there are several findings:

UGC is not only a leading indicator of stock market performance [39]. UGC knowledge features can have a significant impact on the stock market [40]. It further proves the effectiveness of extracting UGC knowledge features from the perspective of investment risk. The risk knowledge characteristics of UGC have a negative impact on the stock price excess return, that is, the stronger the investment risk knowledge expressed by UGC, the lower the stock price excess return. The UGC risk knowledge spreads rapidly, which has an impact on the stock market within 10 trading days after the event is released. Among them, the knowledge characteristics of the UGC Stock Talk have a faster impact on stock prices than the knowledge characteristics of the UGC Article Title and have a significant impact three days after publication. This is consistent with the timeliness and freedom of posting in forums.

5.2. Enlightenment

From the empirical results, it is not difficult to find the most concerned listed companies in IBSN platforms and the related content. From the perspective of scientific research, the research framework of this paper can help management researchers, investors and financial enthusiasts to quickly discover the public opinion voices and hot spots of investors. Secondly, from the perspective of strategic management and investment management, this research finds out the discussion contents closely related to the strategic planning, product release, global sales, market forecast, and international situation of listed companies. These contents have a certain causal relationship with the stock price changes of listed companies in the short-term window through quantitative indicators of risk characteristics. Therefore, the research results can be used to predict the short-term changes of stock prices in the market, and can also be used to understand the short-term strategy implementation of listed companies. The research significance of this paper is to provide a reference for company managers, investment decision makers, and platform operators. Especially for listed companies with high-risk features, it is suggested that the enterprise management should actively respond to the problems in the market, adjust the strategic deployment in time, and respond positively with the help of social media to protect the image of investors and companies.

6. Conclusions

The main contribution of this paper is to transform the unstructured content of UGC into structured content with a risk assessment label and to test the effective feedback from the market by empirical study. In this paper, a research framework based on the LDA model and knowledge characteristics is presented. LDAvis is used to visualize analyses of related topics, knowledge is identified according to the risk characteristics of empowering topics in documents, and an empirical analysis based on social networks is made using the stock market as an example.

The main conclusions of the article include the following four points:

(1) Topics commented by UGC mostly reflect the hot issues concerned by ordinary users.

(2) The UGC investment analysis postings (or the UGC Article Title) are more professional, and the UGC stock talks in the forum are more colloquial.

(3) Different UGC types lead to differences in topic recognition, language styles, and knowledge characteristics. Therefore, subject identification and language analysis should be conducted separately.

(4) The effect of market feedback on UGC risk assessment knowledge features, which are measured by language analysis, is significantly different. The stronger the expressed risk characteristics, the lower the excess return of stock price. The empirical results show that due to the timeliness of stock talks, the market responds faster to the UGC Stock Talk.

However, there are still some shortcomings in the research process: first, there is a lack of retrospective thinking about the purpose of analysis reports in natural language processing. Second, the algorithm of the topic model lacks in-depth optimization. Therefore, the algorithm optimization of the theme model based on UGC generated by investors' social platforms will become a continuous research direction. At the same time, we will also explore more subject knowledge systems and application scenarios, and expand the principles of scientific management and enterprise intelligence.

Author Contributions: Conceptualization, N.L., K.C. and H.H.; methodology, N.L.; software, H.H.; validation, N.L.; resources, H.H.; data curation, H.H.; writing—original draft preparation, K.C. and H.H.; writing—review and editing, N.L. and K.C.; visualization, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This paper takes Seeking Alpha (hereinafter referred to as SA, http://seekingalpha.com, accessed on 12 February 2020) as the data source of the research sample. Financial data comes from Wharton Research Data Services (WRDS) platform. The data presented in this study are available on request from the corresponding author.

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

Appendix A

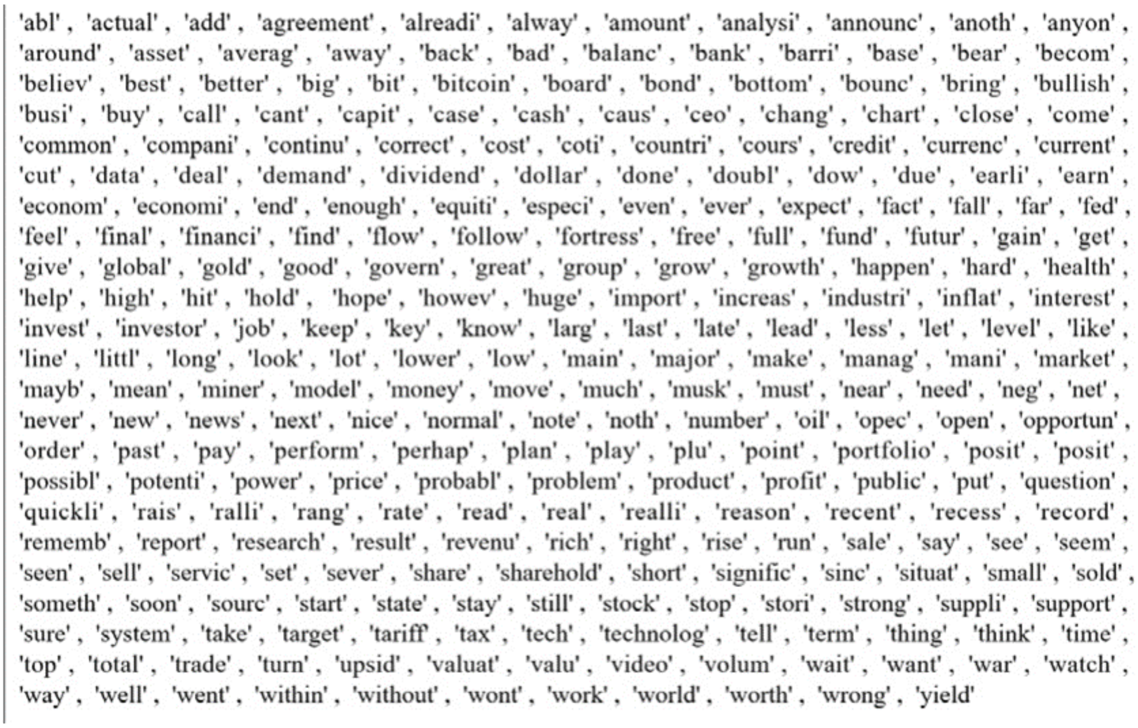
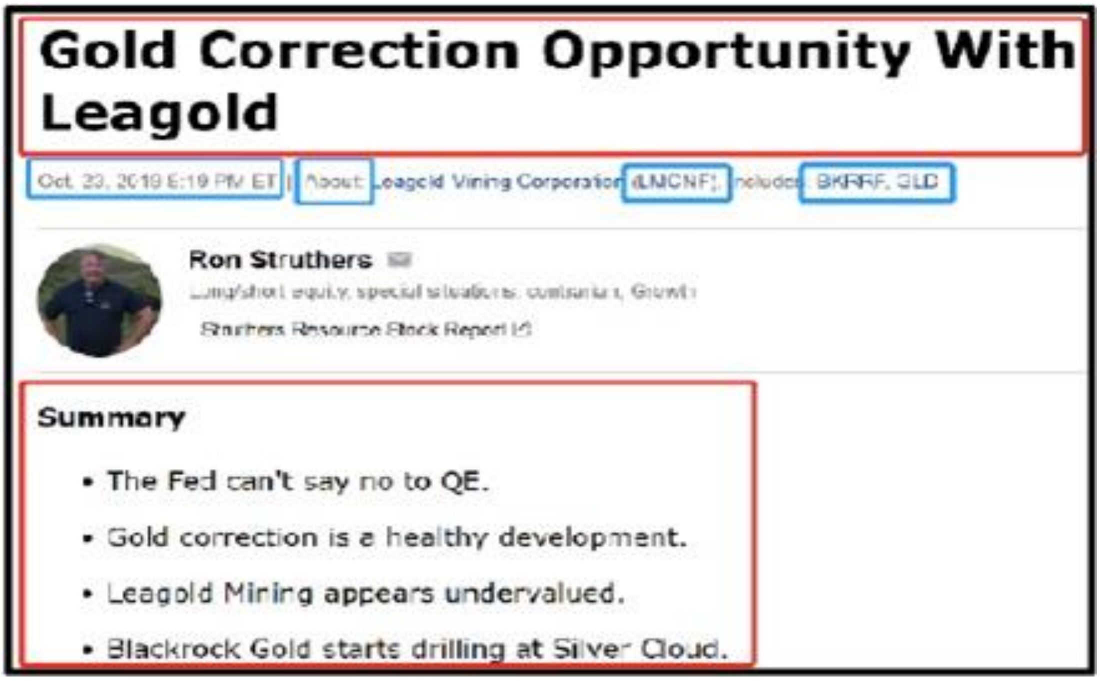

'abl' , 'actual' , 'add' , 'agreement' , 'alreadi' , 'alway' , 'amount' , 'analysi' , 'announc' , 'anoth' , 'anyon' ,
'around' , 'asset' , 'averag' , 'away' , 'back' , 'bad' , 'balanc' , 'bank' , 'barri' , 'base' , 'bear' , 'becom' ,
'believ' , 'best' , 'better' , 'big' , 'bit' , 'bitcoin' , 'board' , 'bond' , 'bottom' , 'bounc' , 'bring' , 'bullish' ,
'busi' , 'buy' , 'call' , 'cant' , 'capit' , 'case' , 'cash' , 'caus' , 'ceo' , 'chang' , 'chart' , 'close' , 'come' ,
'common' , 'compani' , 'continu' , 'correct' , 'cost' , 'coti' , 'countri' , 'cours' , 'credit' , 'currenc' , 'current' ,
'cut' , 'data' , 'deal' , 'demand' , 'dividend' , 'dollar' , 'done' , 'doubl' , 'dow' , 'due' , 'earli' , 'earn' ,
'econom' , 'economi' , 'end' , 'enough' , 'equiti' , 'especi' , 'even' , 'ever' , 'expect' , 'fact' , 'fall' , 'far' , 'fed' ,
'feel' , 'final' , 'financi' , 'find' , 'flow' , 'follow' , 'fortress' , 'free' , 'full' , 'fund' , 'futur' , 'gain' , 'get' ,
'give' , 'global' , 'gold' , 'good' , 'govern' , 'great' , 'group' , 'grow' , 'growth' , 'happen' , 'hard' , 'health' ,
'help' , 'high' , 'hit' , 'hold' , 'hope' , 'howev' , 'huge' , 'import' , 'increas' , 'industri' , 'inflat' , 'interest' ,
'invest' , 'investor' , 'job' , 'keep' , 'key' , 'know' , 'larg' , 'last' , 'late' , 'lead' , 'less' , 'let' , 'level' , 'like' ,
'line' , 'littl' , 'long' , 'look' , 'lot' , 'lower' , 'low' , 'main' , 'major' , 'make' , 'manag' , 'mani' , 'market' ,
'mayb' , 'mean' , 'miner' , 'model' , 'money' , 'move' , 'much' , 'musk' , 'must' , 'near' , 'need' , 'neg' , 'net' ,
'never' , 'new' , 'news' , 'next' , 'nice' , 'normal' , 'note' , 'noth' , 'number' , 'oil' , 'opec' , 'open' , 'opportun' ,
'order' , 'past' , 'pay' , 'perform' , 'perhap' , 'plan' , 'play' , 'plu' , 'point' , 'portfolio' , 'posit' , 'posit' ,
'possibl' , 'potenti' , 'power' , 'price' , 'probabl' , 'problem' , 'product' , 'profit' , 'public' , 'put' , 'question' ,
'quickli' , 'rais' , 'ralli' , 'rang' , 'rate' , 'read' , 'real' , 'realli' , 'reason' , 'recent' , 'recess' , 'record' ,
'rememb' , 'report' , 'research' , 'result' , 'revenu' , 'rich' , 'right' , 'rise' , 'run' , 'sale' , 'say' , 'see' , 'seem' ,
'seen' , 'sell' , 'servic' , 'set' , 'sever' , 'share' , 'sharehold' , 'short' , 'signific' , 'sinc' , 'situat' , 'small' , 'sold' ,
'someth' , 'soon' , 'sourc' , 'start' , 'state' , 'stay' , 'still' , 'stock' , 'stop' , 'stori' , 'strong' , 'suppli' , 'support' ,
'sure' , 'system' , 'take' , 'target' , 'tariff' , 'tax' , 'tech' , 'technolog' , 'tell' , 'term' , 'thing' , 'think' , 'time' ,
'top' , 'total' , 'trade' , 'turn' , 'upsid' , 'valuat' , 'valu' , 'video' , 'volum' , 'wait' , 'want' , 'war' , 'watch' ,
'way' , 'well' , 'went' , 'within' , 'without' , 'wont' , 'work' , 'world' , 'worth' , 'wrong' , 'yield'

Figure A1. Risk Characteristics Roots Based on Forum Posts (284 roots).


Gold Correction Opportunity With Leagold

Oct. 23, 2019 6:19 PM ET | About: Leagold Mining Corporation (LMCNF), Includes: BKRRF, GLD

Ron Struthers

Long/short equity, special situations, contrarian, Growth

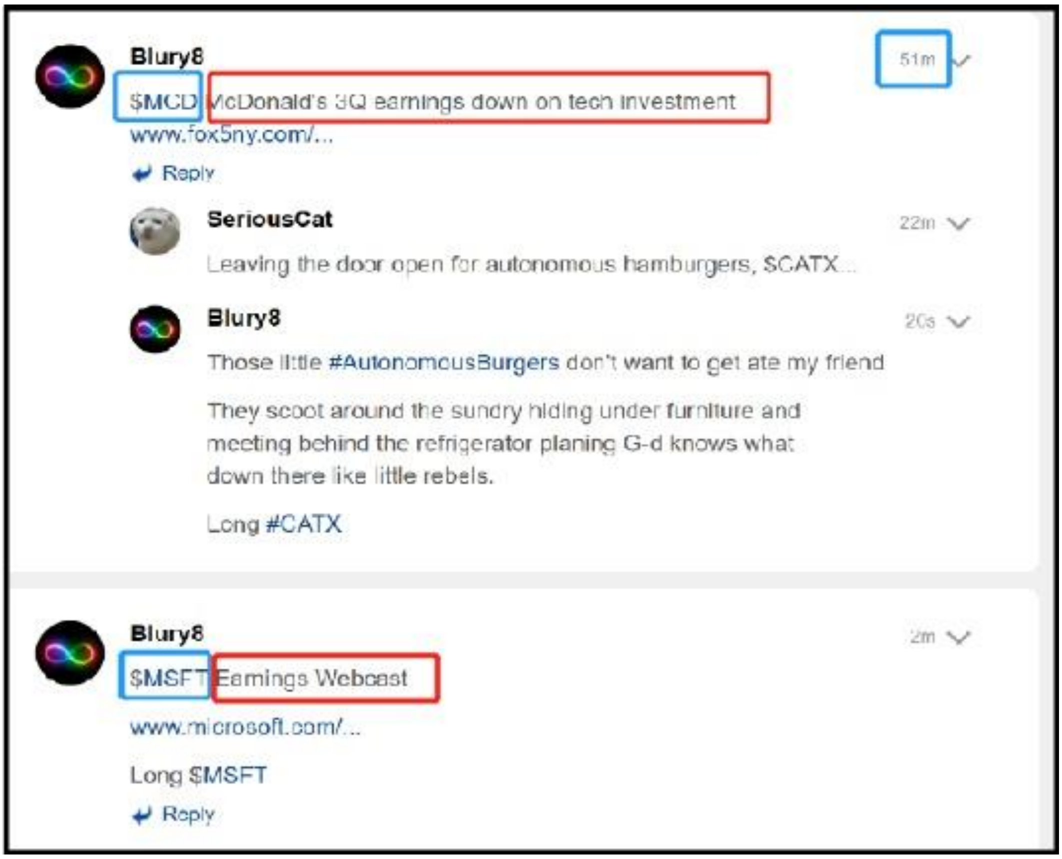
Struthers Resource Stock Report

Summary

- The Fed can't say no to QE.
- Gold correction is a healthy development.
- Leagold Mining appears undervalued.
- Blackrock Gold starts drilling at Silver Cloud.

(**a**)

Figure A2. *Cont.*



Blury8 51m

$MCD McDonald's 3Q earnings down on tech investment

www.fox5ny.com/...

Reply

SeriousCat 22m

Leaving the door open for autonomous hamburgers, $CATX...

Blury8 20s

Those little #AutonomousBurgers don't want to get ate my friend

They scoot around the sundry hiding under furniture and meeting behind the refrigerator planing G-d knows what down there like little rebels.

Long #CATX

Blury8 2m

$MSFT Earnings Webcast

www.microsoft.com/...

Long $MSFT

Reply

(**b**)

Figure A2. Screenshots of sample published articles and forum discussions by Seeking Alpha (Source: 24 October 2019, Seeking Alpha screenshot at http://seekingalpha.com, accessed on 12 February 2020). (**a**) Title and abstract of the SA article. (**b**) SA Stock Talk forum discussion.

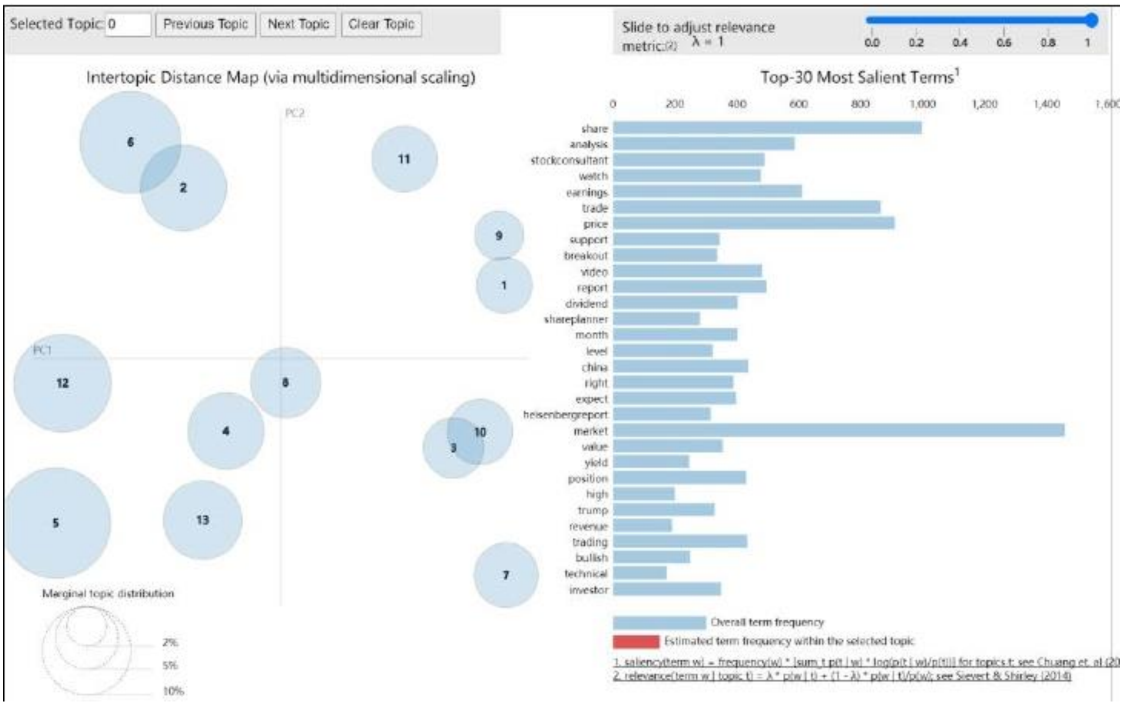


Figure A3. Visual distribution of topics discussed in Discussion Forum (K = 13).

References

1. Mannens, E.; Verwaest, M.; Van de Walle, R. Production and multi-channel distribution of news. *Multimed. Syst.* **2008**, *14*, 359–368. [CrossRef]
2. Domingo, D.; Masip, P.; Meijer, I.C. Tracing digital news networks towards an integrated framework of the dynamics of news production, circulation and use. *Digit. J.* **2015**, *3*, 53–67.
3. dos Santos, M.L.B. The "so-called" UGC: An updated definition of user-generated content in the age of social media. *Online Inf. Rev.* **2022**, *46*, 95–113. [CrossRef]
4. Sun, R.; Hong, X.-J. Social Presence and User-Generated Content of Social Media in China. *Int. J. Semant. Web Inf. Syst.* **2019**, *15*, 35–47. [CrossRef]
5. Wang, X.; Chen, X. The Impact of Graphic and Text Matching on Consumer Perceived Usefulness of User Generated Content. *Manag. Sci.* **2018**, *31*, 101–115.
6. Hou, L.; Li, J.; Li, X.-L.; Tang, J.; Guo, X. Learning to Align Comments to News Topics. *ACM Trans. Inf. Syst.* **2017**, *36*, 1–30. [CrossRef]
7. Tu, W.T.; Yang, M.; Cheung, D.W.; Mamoulis, N. Investment recommendation by discovering high-quality opinions in investor based social networks. *Inf. Syst.* **2018**, *78*, 189–198. [CrossRef]
8. Wang, L.; Li, S.W.; Chen, T.Q. Investor behavior, information disclosure strategy and counterparty credit risk contagion. *Chaos Solitons Fractals* **2019**, *119*, 37–49. [CrossRef]
9. Singh, R.; Srivastava, S. Stock prediction using deep learning. *Multimed. Tools Appl.* **2017**, *76*, 18569–18584. [CrossRef]
10. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
11. Zhang, Y.; Wei, H.; Ran, Y.; Deng, Y.; Liu, D. Drawing openness to experience from user generated contents: The An interpretable data—Driven topic modeling approach. *J. Expert Syst. Appl.* **2020**, *144*, 113073. [CrossRef]
12. Prollochs, N.; Feuerriegel, S. Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Inf. Manag.* **2020**, *57*, 103070. [CrossRef]
13. Nam, H.; Joshi, Y.V.; Kannan, P. Harvesting brand information from social tags. *J. Mark.* **2017**, *81*, 88–108. [CrossRef]
14. Krishnamurthy, S.; Dou, W. Note from special issue editors. *J. Interact. Advert.* **2008**, *8*, 1–4. [CrossRef]
15. Hofmann, T. Probabilistic Latent Semantic Indexing. In Proceedings of the Sigir'99: Proceedings of 22nd International Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999.
16. Peng, G.; Yuefen, W.; Zhu, F. Analysis of Topic Extraction Effect of Scientific Literature Based on LDA Topic Model in Different Corpus. *Libr. Inf. Serv.* **2016**, *60*, 112–121.
17. Liu, Z.; Xu, H.; Yue, L.; Fang, S. Research on Core Technology Theme Recognition Method Based on Chunk-LDAVIS. *Libr. Inf. Sci.* **2019**, *63*, 73–84.
18. Li, C.; Feng, S.; Zeng, Q.; Ni, W.; Zhao, H.; Duan, H. Mining dynamics of research topics based on the combined LDA and Wordnet. *IEEE Access* **2019**, *7*, 6386–6399. [CrossRef]
19. Xu, Y.; Li, Y.; Liang, Y.; Cai, L. Topic-sentiment evolution over time: A manifold learning-based model for online news. *J. Intell. Inf. Syst.* **2020**, *55*, 27–49. [CrossRef]
20. Rosen-Zvi, M.; Chemudugunta, C.; Griffiths, T.; Smyth, P.; Steyvers, M. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.* **2010**, *28*, 73–86. [CrossRef]
21. Wang, H.; Wu, F.; Lu, W.; Yang, Y.; Li, X.; Li, X.; Zhuang, Y. Identifying objective and subjective words via topic modeling. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 718–730. [CrossRef]
22. Fayyad, U.N. From Data Mining to Knowledge Discovery: On Overview. *Adv. Knowl. Discov. Data Min.* **1996**, *1*, 12.
23. Walter, J.; Lechner, C.; Kellermanns, F.W. Knowledge transfer between and within alliance partners: Private versus collective benefits of social capital. *J. Bus. Res.* **2007**, *60*, 698–710. [CrossRef]
24. Rennolls, K.; Society, I.C. An intelligent framework (O-SS-E) for data mining, knowledge discovery and business intelligence. In Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05), Copenhagen, Denmark, 22–26 August 2005; pp. 715–719.
25. Cazzella, S.; Dragone, L. The Role of Domain Knowledge in KDD-Based Strategic Marketing Applications. In Proceedings of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics, Orlando, FL, USA, 18–21 July 2004; pp. 381–386.
26. Budanitsky, A.; Hirst, G. Evaluating WordNet-based measures of lexical semantic relatedness. *Comput. Lin-Guistics* **2006**, *32*, 13–47. [CrossRef]
27. Miller, G.A. Wordnet—A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
28. Leacock, C.; Miller, G.A.; Chodorow, M. Using corpus statistics and WordNet relations for sense identification. *Comput.-Tional Linguist.* **1998**, *24*, 147–165.
29. Schmitt, X.; Kubler, S.; Robert, J.; Papadakis, M.; LeTraon, Y. A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019; pp. 338–343.
30. Omran, F.A.; Treude, C. Choosing an NLP library for analyzing software documentation: A systematic literature review and a series of experiments. In Proceedings of the 14th International Conference on Mining Software Repositories, Buenos Aires, Argentina, 20–21 May 2017.

31. Zhang, X.; Wen, Y.; Xu, H.; Liu, Z. Evolution of Prophet Prediction-Correction Topic Strength Model—An Empirical Study in Stem Cell Field. *Libr. Inf. Serv.* **2020**, *64*, 78–92.
32. Arun, R.; Suresh, V.; Madhavan, C.V.; Narasimha Murthy, M.N. On finding the natural number of topics with latent Dirichlet allocation: Some observations. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Hyderabad, India, 21–24 June 2010; Volume 1, pp. 391–402.
33. Mimno, D.; Wallach, H.M.; Talley, E.; McCallum, A. Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Scotland, UK, 27–31 July 2011; pp. 262–272.
34. Zhang, H. *Research on Technology Prediction Method from the Perspective of Data Fusion*; Jilin University: Changchun, China, 2019.
35. Hedlund, D.; Ahlund, A. Language has a home: How case officers make use of language analysis in asylum decisions. *J. Ethn. Migr. Stud.* **2020**, *47*, 1578–1595. [CrossRef]
36. Zou, R.; Yu, J. Social network analysis of informal academic communication in digital age: A case study of the small wood vermin life science forum. *Inf. Sci.* **2015**, *33*, 81–86.
37. Luss, R.; D'Aspremont, A. Predicting abnormal returns from news using text classification. *Quant. Financ.* **2015**, *15*, 999–1012. [CrossRef]
38. Kauffman, R.J.; Spaulding, T.J.; Wood, C.A. Are online auction markets efficient? An empirical study of market liquidity and abnormal returns. *Decis. Support Syst.* **2009**, *48*, 3–13. [CrossRef]
39. Ramirez, E.; Gau, R.; Hadjimarcou, J.; Xu, Z. User-generated content as word-of-mouth. *J. Mark. Theory Pract.* **2018**, *26*, 90–98. [CrossRef]
40. Tirunillai, S.; Tellis, G.J. Does chatter really matter? Dynamics of user-generated content and stock performance. *Mark. Sci.* **2012**, *31*, 198–215. [CrossRef]