sustainability

MDPI

Article

Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining

Jose Ramon Saura [1,*], Pedro Palos-Sanchez [2] and Antonio Grilo [3]

[1] Department of Business Economics, Faculty of Social Sciences and Law, Rey Juan Carlos University, Paseo Artilleros s/n, 28032 Madrid, Spain
[2] Department of Business Administration and Marketing, University of Seville, 41018 Seville, Spain; ppalos@ues.es
[3] UNIDEMI, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa; acbg@fct.unl.pt
* Correspondence: joseramon.saura@urjc.es

Received: 11 January 2019; Accepted: 4 February 2019; Published: 11 February 2019

Abstract: The main aim of this study is to identify the key factors in User Generated Content (UGC) on the Twitter social network for the creation of successful startups, as well as to identify factors for sustainable startups and business models. New technologies were used in the proposed research methodology to identify the key factors for the success of startup projects. First, a Latent Dirichlet Allocation (LDA) model was used, which is a state-of-the-art thematic modeling tool that works in Python and determines the database topic by analyzing tweets for the #Startups hashtag on Twitter (n = 35.401 tweets). Secondly, a Sentiment Analysis was performed with a Supervised Vector Machine (SVM) algorithm that works with Machine Learning in Python. This was applied to the LDA results to divide the identified startup topics into negative, positive, and neutral sentiments. Thirdly, a Textual Analysis was carried out on the topics in each sentiment with Text Data Mining techniques using Nvivo software. This research has detected that the topics with positive feelings for the identification of key factors for the startup business success are startup tools, technology-based startup, the attitude of the founders, and the startup methodology development. The negative topics are the frameworks and programming languages, type of job offers, and the business angels' requirements. The identified neutral topics are the development of the business plan, the type of startup project, and the incubator's and startup's geolocation. The limitations of the investigation are the number of tweets in the analyzed sample and the limited time horizon. Future lines of research could improve the methodology used to determine key factors for the creation of successful startups and could also study sustainable issues.

Keywords: startups business; technology management; sustainable startups; sentiment analysis; text data mining

1. Introduction

In recent years, advances in new technologies have meant that companies have adopted new business models that incorporate globalization and using the Internet as a promotion tool for products and services [1]. With the evolution of technologies since the first decade of the 21st century, these business models have been adapting to include new processes and social changes, as well as the new demands of consumers, who are increasingly supported in this new digital era where the use of new technologies has become a habit in both professional and personal worlds [2,3].

In this new Digital age, companies adopt business models that are scalable by using technologies that can help them to understand what and how their users and clients think [4]. Users' thinking is expressed on digital platforms and environments and is called User Generated Content (UGC) [5,6].

Over the last decade, UGC has been used in research to determine the key factors for any chosen topic [7]. The term 'startup' was coined for business models using technology. A startup is a technology-based company that offers a new product or service using the added value of the incorporated technology. This is defined as 'innovation through technology' [8].

Startups use scalable business models, that is, startups make investments for the improvement of the technology on which they base their project, and once the technology has been improved, the product or service is created [3]. The product or service is launched when the product is ready, and many startups create successful products and services for consumers whose consumption habits are based on the digital age. Examples of services and products that were created as startups are WhatsApp, Facebook, Instagram, and the technological giant Alphabet (Google) [8].

Startups are small companies that start from an innovative idea using technology and with time and experience become a solid and solvent technological and innovative company that is sustainable over time. In a global ecosystem where new technologies and processes are produced on a daily basis, it is important to know the key factors that can make a startup successful, as well as identifying the technologies that will determine what humans will do in the coming years. [9,10].

Currently there are technological processes that generate data and information in real time. New technologies such as big data, data mining, and artificial intelligence or business intelligence are the results of analyzing this data, and all provide important value for companies [11]. Startups base their business models on innovation. Innovation is the process of searching for a value that improves a current product or service or satisfies a demand that has not been covered until now [12,13]. Technological innovation is what startups do with new products and services by working with an emerging technology and applying it to a new or existing product [14].

In this sense, it is interesting that academic researchers can study what the success factors for startups are, and also compare how the findings of these research studies fit with the startup industry. The startups industry needs to know what the success factors for their business are, since they want to develop successful and profitable business models over time. Consequently, this study aimed to identify the key factors that make a startup successful by analyzing the comments made in UGC on the Twitter social network. In addition, new technologies were used to carry out the methodology of this study. First, Latent Dirichlet Allocation (LDA), which is a state-of-the-art topic modeling tool that works in Python, was used to find the topics of a database made up of extracted tweets with the hashtag #Startups on Twitter (n = 35.401). Sentiment Analysis was performed with a Supervised Vector Machine algorithm (SVM) that works with Machine Learning in Python. This algorithm was applied to the results of the LDA model to divide the identified topics into negative, positive, and neutral, for the key factors that make a startup successful. Finally, a Textual Analysis was performed with the qualitative analysis software Nvivo in order to identify the key factors for the success and creation of startups using the results found for the users' sentiments about the topics identified in the Twitter UGC.

2. Literature Review

2.1. UGC Analysis

The UGC analysis was performed on samples generated as a result of on-line comments, user generated content, and reviews made on online platforms. An online review or comment is a piece of text in a public profile on the Internet that describes a user's experience with a product, service, or topic [15]. By studying this type of UGC on the Internet, a solid, causal relationship can be built that has a powerful meaning and can be useful for effective research [16,17]. At the same time, advances in technology that give rise to new research models help to improve Text Data Mining techniques, which help, among other things, to automatically find information in large databases or to recognize topics in database data generated by comments, reviews, and User Generated Content on social networks [18]. For example, the Latent Dirichlet Allocation (LDA) is a modeling tool that is able to identify topics from a database of qualitative reviews and comments and quantify and count the

number of comments made about any topic [6,19]. Table 1 below shows the characteristics of LDA models when analyzing UGC-type content in other studies.

Table 1. Characteristics of User Generated Content (UGC) analysis in research studies. LDA = Latent Dirichlet Allocation.

Characteristics	[18]	[5]	[20]	[21]	[22]	[23]	[3]	[24]	This Research
Online Rating	√	√	-	√	√	-	-	-	-
Comments	-	√	√	√	√	√	√	√	√
LDA	-	-	-	√	√	√	√	√	√
Social Interactions	-	-	-	√	√	√	√	√	√
Topic Frequency	-	-	√	√	-	√	√	√	√

Source: Adapted from Jia [6]

A benefit of analyzing Twitter interactions and UGC is that user comments about other companies are included. This allows the amount of engagement to be measured. In this way, researchers are able to analyze a user's motivation due to a UGC comment. Jia [6] and Wang and Zhai [25] discovered that the most important types of motivations were knowledge and sense of belonging from the content generated by chat groups on the Internet. These were found by analyzing the chat messages without directly asking the users who wrote the comments.

The research by Liang et al. [26] found users' motivations by studying users' textual expression on the Internet. In addition, finding correlations between users and the number of ratings can also be a way to quantify these methodologies in order to obtain metrics for the motivation and satisfaction of Internet users who generate content. For example, Saura et al. [3] analyzed the reviews of hotel users with the UGC on the TripAdvisor social network. Companies can see if their consumers, or users, are happy or satisfied with them by using a UGC analysis approach and can also find out the reasons for the users2019 feelings.

2.2. Sentiment Analysis with Social Network Analysis

Sentiment Analysis is a research methodology that analyzes the feelings of a given sample, which normally comes from digital environments such as online platforms or social networks, to find the different opinions with different methodological approaches. It has been confirmed that Sentiment Analysis can identify the feelings and therefore the opinions of product users in order to understand how these feelings and opinions affect the users' decision making [3,8]. There are different options and approaches for this technique. Approximations can be made using special software for applying machine learning, artificial intelligence techniques, and hybrid models. Other options are available, such as algorithm training with Data Mining techniques, which are processes used to improve the probability of success of an algorithm with machine learning based on the accuracy of the results [27,28].

Several studies have been carried out with machine learning models to analyze social networks, users' opinions, and to identify the key factors that influence different cases. Supervised methods using the classification and categorization of key factors, such as Maximum Entropy (MaxEnt) and Support Vector Machines (SVMs), have been used to perform social network analysis with machine learning using technological research methods to identify the important factors in different areas of research [14,29]. However, there are other types of approaches based on sentiment analysis with UGC, such as Naïve Bayes, Linear Regression, or Deep Learning [3]. These studies used keywords; ratings of feelings about a topic; semantic meaning; concepts and semantic theory; feelings about topics such as hashtags, retweets, or points on social networks; and valuation identifiers for products and services on the Internet [30].

In the research by Liang [26], a semi-supervised dual recurrent neural network was used to perform Sentiment Analysis. This is similar to a traditional neural network and can be used to evaluate a data set over a long period of time [6,27]. This technique allows an effective and efficient Sentiment Analysis to be carried out [6]. In the research by Reyes-Menendez et al. [17], Sentiment Analysis was

performed on the UGC for a hashtag (#WorldEnvironmentDay) and in the research by Saura et al. [3] and [14], the authors used a SVM algorithm that identified the sentiments of a sample collected from social networks and divided them into negative, positive, and neutral sentiments. In Hasan et al. [31], a recursive neural network was used to understand the meaning of different comments.

It can be seen that the feelings expressed in UGC can be analyzed with Neural Connection Analysis of groups of interacting users; Textual Analysis, which is the analysis of words and their sentiments to determine key factors; Time, which is the analysis of the time over which this UGT takes place, and the analysis of patterns of sentiments in social networks; Hashtags, URLs, or Mentions, which is the analysis of labels of user groups; Topic, which is the analysis of sentiments of content categories with the same topic; and finally, Classification of Information, which is a feature of Sentiment Analysis that uses information keywords with the sample [6,32,33]. Table 2 shows the main characteristics of Sentiment Analysis.

Table 2. Summary of the main research using Sentiment Analysis.

Characteristics	[34]	[35]	[36]	[17]	[37]	[3]	This Research
Neural Connection Analysis	√	-	√	-	√	-	-
Textual Analysis	-	√	√	√	√	√	√
Time	-	√	-	-	-	-	√
Hashtags, URLs, or Mentions	-	√	-	√	-	√	√
Topic	-	√	√	√	-	√	√
Classification of Information	√	-	√	√	-	√	√

Source: Adapted from Saura et al. [3].

2.3. Textual Analysis

Textual analysis is a text mining analysis approach that determines key factors by analyzing large amounts of data. It is a qualitative approach that uses the weight and repetition of text in a given sample to determine the keywords that express the sentiments shown about the subject under study [38]. In the research by Vázquez and Escamilla [39], a textual analysis process is undertaken with the Nvivo software, which aimed to identify attitudes towards the main factors for the health of the elderly. In the research by Saito et al. [40], textual analysis was used to predict re-tweets from the relevance of the UGC content on Twitter.

Likewise, Jiang et al. [41] analyzed the fundamental factors that affect a concept called "re-tweetability" for each tweet when using a predictive filter for the collaboration between users [42], the connections, and the repetition of keywords in tweets. Therefore, Textual Analysis can be used to determine and identify the keywords with the greatest weight in a given sample and study the influence of these on the content.

Table 3 shows a summary of the main characteristics of the approximations when using textual analysis to identify key factors in UGC analysis.

Table 3. Summary of the main characteristics of Textual Analysis approaches.

Characteristics	[43]	[44]	[14]	[40]	[41]	[45]	[37]	[6]	This Research
Classification into Nodes	√	√	-	√	√	√	√	-	-
Categorization	√	√	√	√	√	-	-	√	√
Word Count	√	√	√	-	-	√	√	√	√
Keywords	-	-	√	-	-	√	√	√	√

Source: Adapted from Saura et al. [3]

3. Research Questions

The following Research Questions (RQs) were proposed for this study using the above information because of the interest shown by startups in identifying technologies for their business models. Previous studies have shown that the most important topics can be found for different industries and areas by using UGC analysis and approximations [6,13]. In addition, LDA can be used to identify topics in the UGC on social networks [29]. This study used the following research questions to identify whether important business topics for startups can be found from the comments in UGC content on Twitter:

RQ 1: *Can important business topics for Startups be found in the UGC content on Twitter?*

Different studies have used methodological approaches with UGC content to find the feelings expressed by the comments and opinions of users on social networks such as Twitter, Google Maps, TripAdvisor, or Booking.com [14,15,17]. In this study, the Twitter social network was used for Sentiment Analysis of the topics commented on in Twitter users' UGC. These were divided into positive, negative, and neutral sentiments:

RQ 2: *What sentiments are expressed about the topics for startup business success in Twitter UGC?*

Key factors are important factors that influence the advance of a topic [14,25]. For startup businesses, key factors could be the leadership of the managers, the management of the team members, or the innovation technology chosen by the startup [36]. Other key indicators could be related to investors or new business models for sustainable approaches. The following research question was proposed for this research:

RQ 3: *Could key indicators for startup business success be found from the sentiment topics of Twitter UGC content, and can these results consequently determine negative, positive, and neutral factors for success?*

4. Methodology

The methodology was divided into three-phases. First, a Latent Dirichlet Allocation (LDA) model was used, which is a state-of-the-art thematic modeling tool that works in Python and determines the database topic by analyzing tweets for the #Startups hashtag on Twitter (n = 35.401 tweets). Secondly, a Sentiment Analysis was performed with a Supervised Vector Machine (SVM) algorithm that works with Machine Learning in Python to divide the identified topics into negative, positive, and neutral for the key factors that make a startup business successful. Thirdly, a Textual Analysis was performed on the results with Text Data Mining techniques using the Nvivo qualitative analysis software.

4.1. Data Sampling

The sample for this study was structured using information from previous studies that used the same methodology for a sample of 2000 Tweets and another of 10,000 Tweets [17,46]. Palomino et al. [46] extracted information from 6333 Tweets for the #getoutside hashtag and Reyes-Menendez et al. [17] used the #WorldEnvironmentDay hashtag with a sample of 5874 tweets. The public Twitter API (Application programming interface) was used to download a total of n = 35,401 tweets in order to extract data. Initially, the sample consisted of 44,101 tweets, but after the database cleaning process, the final sample was reduced to 35,401 tweets. This step was done using the MAC version of Python software 3.7.0. The tweets that contained the keywords: "startup", "start-up", "startups", and "start-ups" in English were used [17,38]. The database of downloaded tweets was cleaned to eliminate tweets that were repeated because they were news, duplicate content, or retweets. The images and multimedia files published next to the Tweet text were not analyzed. Using Saura et al. [3], the sample of tweets was validated with the following criteria:

- Active Twitter profile (profiles without activity in the three months prior to the use of #Startups were deleted)
- Twitter user profile had a profile photo and a cover picture
- No retweets. Retweets from the same tweet about #startup, "start-up", "startups" and "start-ups" were removed (i.e., considered as duplicate content)
- Public profiles. Only public profiles and tweets using #Startups in English were included
- Minimum 80 characters. Tweets had to be at least 80 characters long (including spaces) and use the #Startups hashtag. This means that tweets without the "#" or a wrong label like "# startups" were omitted.

4.2. Topic Identification Using LDA

The LDA model is based on a probabilistic assumption that assumes that content is generated in two steps [6,14,47]. The first step identifies words and separates each word into a different document. The next step randomly identifies the distribution of the topics in a sample, and then selects the main topics found in that sample [6,13–15]. In real situations, neither the distribution of topics in documents nor the distribution of words in topics is known a priori [6]. The importance of the hidden and observed variables is the joint distribution expressed mathematically in (1) below:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) x \prod_{d=1}^{D} p(\theta_d) x \sum_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \quad (1)$$

where βi is the distribution of a word in topic i, with total K topics; θd is the proportion of topics in document d, with total D documents; z_d is the topic assignment in document d; z_{dn} is the topic assignment for the nth word in document d, with total N words; w_d is the observed words for document d; and $w_{d,n}$ is the nth word for document d.

Finally, the topics and words were identified using Equation (2) below for Gibbs sampling [6]. The calculation was performed with Python LDA 1.0.5 software in this research.

$$p(\beta_{1:k}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \frac{p(\beta_{1:k}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \quad (2)$$

4.3. Sentiment Analysis

After the topics had been identified using the LDA model, the Python algorithm supplied by MonkeyLearn (MonkeyLearn, San Francisco, CA, USA) was used by connecting to its API [14,15,36]. This was done after training machine learning with the data-mining processes and subdividing the sample into positive, negative, and neutral sentiments about the technology startup industry. A total of 481 samples were processed with data-mining. The sample of Tweets was fed into the MonkeyLearn application and the interface was linked to the Sentiment Analysis algorithm until a probability percentage of >0.674 was reached [3,13–15]. Algorithm training was carried out after identifying the content that was exclusively related to the research topic, including the ironic and sarcastic comments [17]. Throughout the entire process, the content that was not related to #startups was discarded from the sample and training.

4.4. Textual Analysis

The databases were processed for sentiments using different stages of the Nvivo software in which the tweets were categorized into the following three nodes: Positive (N_1), Neutral (N_2), and Negative (N_3) [13,14,48]. The data entry process was manual for Nvivo although the databases were already divided into Sentiments. The researchers then created the node structure and filtered the database, eliminating the words identified as connectors, prepositions, articles, and plural forms [6,49]. The nodes were defined as data containers that were grouped according to their characteristics. It should be noted

that the design and development of nodes is a way to analyze pure data and to achieve the highest possible descriptive and research quality. An important indicator for the analysis using Nvivo is the weighted percentage [17,50], which shows the number of times the data in a node is repeated in the sample. To calculate the weighted percentage, the following formula was used:

$$K = \sum k_i / n_i = \{1, \ldots, n\}\ n = [1, 25] \quad (3)$$

In this formula, a query that allows the program to search the text is used to find K. The behavior of each of the words and each tweet can be seen, and the value of K was found for the #Startups hashtag. Using this process, the average value of K for all the tweets was calculated in order to obtain the global value [13,51]. Figure 1 shows the steps of the methodology used in this investigation.

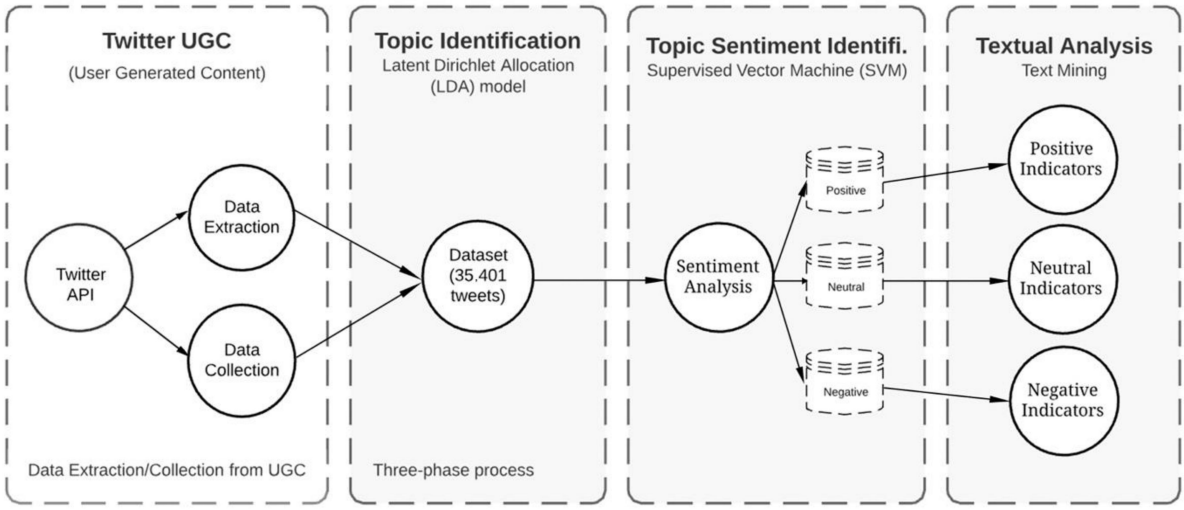

Figure 1. Summary of the three-phase methodology process; Source: the authors.

5. Results Analysis

5.1. Latent Dirichlet Allocation (LDA) Model

The topics identified with the LDA are shown in Table 4. In the LDA process, the words were automatically categorized into topics, and the researchers gave each topic a name after analyzing the group of words. Manually naming the topics is a standard procedure in LDA-based topic identification [21,23,24]). The name of a topic is usually selected by researchers by taking the top 10 ranking words in the topic classification and forming a meaningful name for the topic from these words [52].

Table 4. Identified topics for startups.

Topic Name	Topic Description
Business Angels	Relationship with investors or business angels to obtain financing for startups.
Business Plans	Information about how to prepare a business plan for startups, which is adapted to its ecosystem.
Startup Project	Information about the startup's foundation, creation, management, and team structure
Startup Methodology	Lean startup method for the development of successful startups. Guidelines to structure the projects.
Startup Incubators	Information about start-up incubators or accelerators that offer startup acceleration and promotion programs in their training programs.
Startup Jobs	Job profiles and job offers in startups. Specialist profiles for developers or digital marketing.
Startup Founders	Information for startup CEOs (Chief Executive Officer) and team leaders.
Technology-Based Startup	Startups that develop or improve the technologies on which their business model is based, seeking innovation and excellence in sustainable business processes and quality.
Startup Geo-Location	Location of startups and information about them. Main startup's location identified.
Startup Tools	Tools that startups use to organize team management and collaboration between the startups' team members.
Startup Frameworks and Programming Languages	Programming languages and frameworks that are usually used in startups to develop their projects.

5.2. Topic Sentiment Identification

Sentiment Analysis of the topics obtained with the LDA model identified the feelings expressed in these topics [6,53]. Sentiment Analysis was separately done on the tweets included in each topic, allowing the topics to be separated into different feelings that were later used in a textual analysis. The sentiment analysis algorithm probability of success is established by (i) the quality of the sample, after having been filtered and refined by the authors, as well as (ii) the number of times the algorithm is trained on the dataset. In this research study, we trained the algorithm with a total of 481 samples that were processed with data-mining techniques. The sample of Tweets was fed into the MonkeyLearn application and the interface was linked to the Sentiment Analysis algorithm until average probability percentages of >0.794 (positive sentiment), >0.802 (neutral sentiment), and >0.693 (negative sentiment) were reached [3,13–15]. The results of the sentiment analysis are shown in Figure 2.

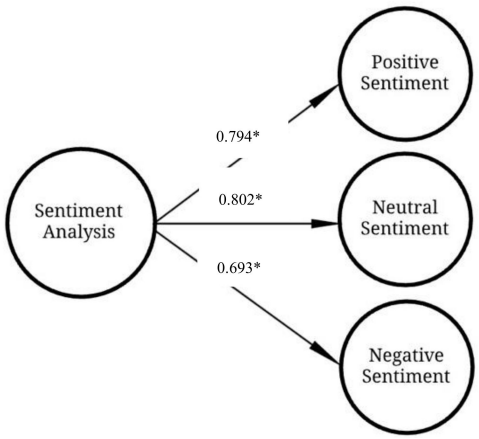

Figure 2. Results of sentiment analysis. * Accuracy: This is the probability of success obtained after the training of the algorithm.

The Sentiment Analysis identified the different sentiments shown about each topic. Table 5 shows the name and description of the topic and the identified sentiment (positive, negative, or neutral).

Table 5. Sentiment for each topic.

Topic Name *	Words in the Topic	Sentiment
Business Angels	invest in startups, funding for startups, investor for a startup, startup capital, money for a startup, raise capital for a startup, investing in startups, angel's startup, venture capital startups	Negative
Business Plans	market a startup, finance a startup, write a business plan, business idea, startups businesses, business plan startup, equity,	Neutral
Startup Project	build a startup, fund a startup, startup costs, startup business, tech startups, small business startup, early stage startups,	Neutral
Startup Methodology	lean startup, education startups, startups books, startups academics reports, sustainable startup, crowd-funding startups	Positive
Startup Incubators	startup institute, startups academy, startup hub,	Neutral
Startup Jobs	work for a startup, startup jobs, startup intern, startup hiring,	Negative
Startup Founders	leader, startup manager, entrepreneur, entrepreneurship, leadership, leader	Positive
Technology-Based Startup	Machine learning, Iot (Internet of Things), AI (Artificial Intelligence), Big Data, Social Network Analysis, Cryptocurrencies, Bitcoin, Digital Marketing, SEO (Search Engine Optimization), SEM (Search Engine Marketing), Social Media Optimization, Neuromarketing	Positive
Startup Geolocation	India, Berlin, Boston, USA, UK, Israel, Dallas, China, Silicon Valley	Neutral
Startup Tools	Slack, Trello, Google Analytics, Docker, GitHub	Positive
Startup Frameworks and Programming Languages	PHP, Python, Java, Go, JS, Node.js, Angular.js, Django, MySQL, PostgreSQL, HTML, CSS	Negative

5.3. Textual Analysis Results

The textual analysis identified factors about startups from the sentiments expressed. The factors that were identified as positive, negative, and neutral for the success of a startup are shown in Tables 6–8. The application of Textual Analysis with Nvivo software identified three nodes for the feelings shown in each topic. Text Data Mining was performed on each of these to find the most important factors from the weight of each in the selected themes. The words were grouped into different nodes according to the number of times the words were repeated in the dataset.

Table 6. Results for startups' positive indicators (N_1).

N_1	Key Factors	Weighted Percentage	Count
Startup Tools	Tools for management in startups are essential due to the large number of processes that are carried out at the same time. Organizational tools help startup teams to communicate better between different processes	2.18	452
Technology-Based Startup	Artificial Intelligence is the future for innovative startups Machine learning is one of the technologies that leads innovation in startups	1.98	245
Startup Founders	The leadership of the CEO in a startup is considered key to its success Teamwork and highly specialized profiles are key to success in a startup	0.97	212
Startup Methodology	The creation and development process in a startup is unique. It must include special procedures for the particular business model used by the startup.	0.98	190

Table 7. Results for startup's neutral indicators (N_2).

N_2	Key factors	Weighted Percentage	Count
Business Plans	Business plans in startups define the viability of the products or services offered. It is of critical importance before receiving an investment.	2.06	310
Startup Project	Startup projects should be sustainable, exponential, and innovative. In addition, they should be based on technological breakthroughs.	1.43	259
Startup Incubators	Startup incubators are an opportunity to start projects with the help of mentors and funding. Startup accelerators are important for small startups that need help to develop their ideas and business plans.	1.31	237
Startup Geolocation	The location of a startup can help its success.The ecosystems and locations where there are many startups can help the projects be successful because of the surrounding community.	0.97	179

Table 8. Results for startup's negative indicators (N_3).

N_3	Key Factors	Weighted Percentage	Count
Startup Frameworks and Programming Languages	Although they are important for the development of startups, there may be problems when trying to find adequately specialized professionals in programming.	2.11	382
Startup Jobs	These are usually low quality with low salaries, although there is a dynamic work environment.	1.36	275
Business Angels	High financial charges when there is a need for investment. The return demanded is too high.	1.22	2.54

Once similar words were grouped into independent nodes, a qualitative approximation was carried out to find the factors of each indicator. N_1 was analyzed for the factors of positive sentiment, N_2 was analyzed for the neutral factors, and N_3 was analyzed for the negative factors.

6. Discussion

This study identified the main topics for the development of successful startups. A large number of social network users' opinions were analyzed in order to identify the relevant factors for this study. The information collected from the UGC on the Twitter social media has given us interesting results in this study.

The positive factors for startups which were stated by users' sentiments in their UGC were identified. The UGC topics identified were related to the management tools used by startups to improve their internal processes; artificial intelligence and machine learning technologies; the attitude of the startups' management and the team leaders; as well as the correct progression of the startup business model that should be based on sustainability and innovation. Negative sentiments were also identified for the key factor about the framework and programming languages that startups use because of the difficulty to find relevant expertise in these areas.

Likewise, the high returns charged by business angels for investment in startup-type projects was also identified as a negative key factor. The neutral sentiment factors were those related to the progression of the business models, the type of projects, the startup incubators, and the location of the startups.

This research study identified the main topics for the success of a startup and also the main factors by analyzing the feelings detected in the UGC on Twitter. This study used a three-phase methodological approach for the analysis of UGC on Twitter. This approach is valid when using an LDA model with defined topics, on which text mining techniques are applied with a machine learning approach. This methodological text mining approach is valid for the analysis of content on social networks to identify important factors in defined research areas.

As has been observed in the results of this research study, the positive factors for a successful startup are characterized by the type of tools it uses, the technology it develops, the leadership and empathy of CEOs, and their methodologies for the project development. It can be said that artificial intelligence, machine learning processes, and the attitude of startup managers are key factors for startups to succeed.

Other factors that obtained a neutral result are the standards for success including the development of business plans, the type of project, and the support of startup incubators, and the geolocation of startups. As for the negative factors, we should highlight that they are factors that can harm the success of a startup if they are not well employed, such as the type of programming languages used, the quality of the job offers, and finally, the treatment from and negotiations with the business angels.

7. Conclusions

This research used a three-phase methodological process to extract the main topics about startups that appeared in Twitter users' UGC. The sentiments of these comments were identified and the key indicators for startup business success were found from the Twitter users' comments.

Important topics were identified in the startup's ecosystem, such as the importance of business plans; the startups' projects; sustainable business models; employee profiles in startups; theoretical and educational support; development or programs of institutions such as startup incubators or accelerators; and attitudes to investors and business angels. In addition, the technologies, applications, tools, and programming languages that startups use were also identified as important topics to consider.

These topics were grouped according to the sentiment that users show about them. These sentiments were negative, positive, and neutral. Key factors for startups in each topic were then identified using the comments in these groups. The key factors found allow us to understand the user's sentiment and attitude to these key issues and factors. This information is important for the creation and development of a successful startup project. Topics are composed by the main points of the startup development process and can be used by practitioners to improve their strategies or rethink their tactics.

RQ1 was answered in this study since the main topics on Twitter have been identified from the large volume of data obtained from the Twitter UGC.

RQ2 was also verified as the sentiments shown for each topic were found and rated by the importance given to the topic in the opinions of the social network users.

RQ3 was answered positively and the route for successful startup business in the digital era has been shown.

The indicators for this route were found from the sentiments shown in users' opinions. Both academics and professionals can use these indicators to create and follow successful startup business models using the results found in this study.

7.1. Theoretical Implications

The theoretical implications of this study about the comments made on social networks, especially on Twitter, for startup business success are for researchers. Data Text Mining allows meaning to be given to large amounts of data that have been grouped by topics. Innovative methods and methodological approaches were used for the analysis of the data in this study, and patterns and indicators were identified that were not found before.

Researchers can use the methodological approach proposed in this research to increase the literature available about research into startups or use these methods to improve and consolidate future studies.

7.2. Practical Implications

This study gives a wide range of practical results that professionals can use in the startup industry. CEOs and startup leaders can take advantage of the key indicators identified in this study to improve their projects by ensuring that the key factors for the success of a startup business according to UGC on Twitter are included in the business plan and project. This study can be a used as a guide to the issues found in the startup ecosystem from the large amount of Twitter data that was analyzed. Entrepreneurs who are considering a startup project can use this research to understand the structure of the startup ecosystem. CEOs and startup leaders can use the topics and key indicators identified in this study to develop and improve their projects.

The limitations of this study are due to the size of the sample, the topic chosen for the study, and the methodological approach taken to reach the conclusions and implications presented. Future lines of research could improve the methodological process of text mining and increase the sample size to try to find new indicators for startups.

Author Contributions: J.R.S., P.P.-S., and A.G. conceived and designed the review; J.R.S. performed the methodology; P.P.-S. and A.G. analyzed the results; J.R.S., P.P.-S., and A.G. wrote the paper.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Zutshi, A.; Grilo, A.; Jardim-Gonçalves, R. A dynamic agent-based modeling framework for digital business models: Applications to Facebook and a popular Portuguese online classifieds website. In *Digital Enterprise Design & Management*; Springer: Cham, Switzerland, 2014; pp. 105–117.
2. Baum, J.A.; Silverman, B.S. Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *J. Bus. Ventur.* **2004**, *19*, 411–436. [CrossRef]
3. Saura, J.R.; Reyes-Menendez, A.; Alvarez-Alonso, C. Do online comments affect environmental management? Identifying factors related to environmental management and sustainability of hotels. *Sustainability* **2018**, *10*, 3016. [CrossRef]
4. Baum, J.A.; Calabrese, T.; Silverman, B.S. Don't go it alone: Alliance network composition and startups' performance in Canadian biotechnology. *Strateg. Manag. J.* **2000**, *21*, 267–294. [CrossRef]
5. Anderson, M.; Magruder, J. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *Econ. J.* **2012**, *122*, 957–989. [CrossRef]
6. Jia, S. Leisure Motivation and Satisfaction: A Text Mining of Yoga Centres, Yoga Consumers, and Their Interactions. *Sustainability* **2018**, *10*, 4458. [CrossRef]
7. Islam, M.; Fremeth, A.; Marcus, A. Signaling by early stage startups: US government research grants and venture capital funding. *J. Bus. Ventur.* **2018**, *33*, 35–51. [CrossRef]
8. Kopera, S.; Wszendybył-Skulska, E.; Cebulak, J.; Grabowski, S. Interdisciplinarity in Tech Startups Development–Case Study of 'Unistartapp'Project. *Found. Manag.* **2018**, *10*, 1–10. [CrossRef]
9. Hagen, C.; Bergh, N.S.; Christensen, S. Startups Seeking Business Angel Financing-From the Entrepreneur's Perspective. Master's Thesis, NTNU, Trondheim, Norway, 2018.
10. Taylor, B.D.; McNair, D.E. Virtual School Startups: Founder Processes in American K-12 Public Virtual Schools. *Int. Rev. Res. Open Distrib. Learn.* **2018**, 19. [CrossRef]
11. Wouters, M.; Anderson, J.C.; Kirchberger, M. New-Technology Startups Seeking Pilot Customers: Crafting a Pair of Value Propositions. *Calif. Manag. Rev.* **2018**, 19. [CrossRef]
12. Herráez, B.; Bustamante, D.; Saura, J.R. Information classification on social networks. Content analysis of e-commerce companies on Twitter. *Rev. Espac.* **2017**, *38*, 16.
13. Saura, J.R.; Palos-Sanchez, P.R.; Correia, M.B. Digital Marketing Strategies Based on the E-Business Model: Literature Review and Future Directions. In *Organizational Transformation and Managing Innovation in the Fourth Industrial Revolution*; IGI Global: Hershey, PA, USA, 2019; pp. 86–103.
14. Saura, J.R.; Palos-Sanchez, P.R.; Rios Martin, M.A. Attitudes to environmental factors in the tourism sector expressed in online comments: An exploratory study. *Int. J. Environ. Res. Public Health* **2018**, *15*, 553. [CrossRef] [PubMed]
15. Saura, J.R.; Palos-Sanchez, P.; Reyes-Menendez, A. Marketing a través de Aplicaciones Móviles de Turismo (M-Tourism). Un estudio exploratorio. *Int. J. World Tourism* **2017**, *4*, 8. [CrossRef]
16. Fukugawa, N. Is the impact of incubator's ability on incubation performance contingent on technologies and life cycle stages of startups? evidence from Japan. *Int. Entrep. Manag. J.* **2018**, *14*, 457–478. [CrossRef]
17. Reyes-Menendez, A.; Saura, J.R.; Alvarez-Alonso, C. Understanding #WorldEnvironmentDay User Opinions in Twitter: A Topic-Based Sentiment Analysis Approach. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2537. [CrossRef]
18. Ye, Q.; Law, R.; Gu, B.; Chen, W. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Comput. Hum. Behav.* **2011**, *27*, 634–639. [CrossRef]
19. Palos-Sanchez, P.; Saura, J.R.; Martin-Velicia, F. A study of the effects of Programmatic Advertising on users' Concerns about Privacy overtime. *J. Bus. Res.* **2019**, *96*, 61–72. [CrossRef]
20. Lee, T.Y.; Bradlow, E.T. Automated marketing research using online customer reviews. *J. Mark. Res.* **2011**, *48*, 881–894. [CrossRef]
21. Büschken, J.; Allenby, G.M. Sentence-based text analysis for customer reviews. *Mark. Sci.* **2016**, *35*, 953–975. [CrossRef]

22. Hao, H.; Zhang, K.; Wang, W.; Gao, G. A tale of two countries: International comparison of online doctor reviews between China and the United States. *Int. J. Med. Inform.* **2017**, *99*, 37–44. [CrossRef]
23. Miller, M.; Banerjee, T.; Muppalla, R.; Romine, W.; Sheth, A. What are people tweeting about Zika? An 561 exploratory study concerning symptoms, treatment, transmission, and prevention. *JMIR Public Health Surveil.* **2017**, *3*, e38. [CrossRef]
24. Liu, X.; Burns, A.C.; Hou, Y. An investigation of brand-related user-generated content on Twitter. *J Advert.* **2017**, *46*, 236–247. [CrossRef]
25. Wang, F.; Zhai, Y. Social structure and evolvement of WeChat groups: A case study based on text mining. *J. China Soc. Sci. Technol. Inform.* **2016**, *35*, 617–629.
26. Liang, Y.; Liu, Y.; Chen, C.; Jiang, Z.G. Extracting topic-sensitive content from textual documents: A hybrid topic model approach. *Eng. Appl. Artif. Intell.* **2018**, *70*, 81–91. [CrossRef]
27. Arora, A.; Fosfuri, A.; Rønde, T. *Waiting for the Payday? The Market for Startups and the Timing of Entrepreneurial Exit (No. w24350)*; National Bureau of Economic Research: Cambridge, MA, USA, 2018.
28. Bennett, D.; Yábar, D.P.B.; Saura, J.R. University Incubators May Be Socially Valuable, but How Effective Are They? A Case Study on Business Incubators at Universities. In *Entrepreneurial Universities*; Innovation, Technology, and Knowledge Management; Peris-Ortiz, M., Gómez, J., Merigó-Lindahl, J., Rueda-Armengot, C., Eds.; Springer: Cham, Switzerland, 2017.
29. Palos-Sanchez, P.; Martin-Velicia, F.; Saura, J.R. Complexity in the Acceptance of Sustainable Search Engines on the Internet: An Analysis of Unobserved Heterogeneity with FIMIX-PLS. *Complexity* **2018**, 1–19. [CrossRef]
30. 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]
31. Hasan, A.; Moin, S.; Karim, A.; Shamshirband, S. Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Math. Comput. Appl.* **2018**, *23*, 11. [CrossRef]
32. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
33. Garbuio, M.; Lin, N. Artificial Intelligence as a Growth Engine for Health Care Startups: Emerging Business Models. *Calif. Manag. Rev.* **2018**. [CrossRef]
34. Pak, A.; Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the LREC, Valletta, Malta, 17–23 May 2010.
35. Kuo, T.-T.; Hung, S.-C.; Lin, W.-S.; Peng, N.; Lin, S.-D.; Lin, W.-F. Exploiting latent information to predict diffusions of novel topics on social networks. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, Jeju Island, Korea, 8–14 July 2012; pp. 344–348.
36. Honeycutt, C.; Herring, S.C. Beyond microblogging: Conversation and collaboration via Twitter. In Proceedings of the 42nd Hawaii International Conference on System Sciences, Hawaii, HI, USA, 5–8 January 2009; pp. 1–10.
37. Boyd, D.; Golder, S.; Lotan, G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In Proceedings of the IEEE 43rd Hawaii International Conference on Social Systems (HICSS), Kauai, HI, USA, 5–8 January 2010.
38. Bologna, G.; Hayashi, Y. A Rule Extraction Study from SVM on Sentiment Analysis. *Big Data Cognit. Comput.* **2018**, *2*, 6. [CrossRef]
39. Vásquez, G.A.; Escamilla, E.M. Best Practice in the Use of Social Networks Marketing Strategy as in SMEs. *Procedia Soc. Behav. Sci.* **2014**, *148*, 533–542. [CrossRef]
40. Saito, K.; Nakano, R.; Kimura, M. Prediction of information diffusion probabilities for independent cascade model. In *Knowledge-Based Intelligent Information and Engineering Systems*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 67–75.
41. Jiang, B.; Liang, J.; Sha, Y.; Li, R.; Liu, W.; Ma, H.; Wang, L. Retweeting behavior prediction based on one-class collaborative filtering in social networks. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Tuscany, Italy, 17–21 July 2016; ACM: New York, NY, USA, 2016; pp. 977–980.
42. Reyes-Menendez, A.; Palos-Sanchez, P.R.; Saura, J.R.; Martin-Velicia, F. Understanding the Influence of Wireless Communications and Wi-Fi Access on Customer Loyalty: A Behavioral Model System. *Wirel. Commun. Mob. Comput.* **2018**. [CrossRef]

43. Kwon, S. *Gerontechnology: Research, Practice, and Principles in the Field of Technology and Aging*; Springer Publishing Company, LLC: New York, NY, USA, 2017.
44. Ramirez-Andreotta, M.; Brody, J.; Lothrop, N.; Loh, M.; Beamer, P.; Brown, P. Improving Environmental Health Literacy and Justice through Environmental Exposure Results Communication. *Int. J. Environ. Res. Public Health* **2016**, *13*, 690. [CrossRef]
45. Rosa, H.; Carvalho, J.P.; Astudillo, R.; Batista, F. Detecting user influence in twitter: Pagerank vs. katz, a case study. In Proceedings of the Seventh European Symposium on Computational Intelligence and Mathematics, Cádiz, Spain, 7–10 October 2015.
46. Palomino, M.; Taylor, T.; Göker, A.; Isaacs, J.; Warber, S. The Online Dissemination of Nature–Health Concepts: Lessons from Sentiment Analysis of Social Media Relating to "Nature-Deficit Disorder". *Int. J. Environ. Res. Public Health* **2016**, *13*, 142. [CrossRef] [PubMed]
47. Palos-Sanchez, P.R. Drivers and Barriers of the Cloud Computing in SMEs: The Position of the European Union. *Harv. Deusto Bus. Res.* **2017**, *6*, 116–132. [CrossRef]
48. Reyes-Menendez, A.; Saura, J.R.; Palos-Sanchez, P.; Alvarez-Garcia, J. Understanding User Behavioral Intention to adopt a Search Engine that promotes Sustainable Water Management. *Symmetry* **2018**, *10*, 584. [CrossRef]
49. Palos-Sanchez, Saura, Jr.; Reyes-Menendez, A.; Esquivel, I.V. Users Acceptance of Location-Based Marketing Apps in Tourism Sector: An Exploratory Analysis. *J. Spat. Organ. Dyn.* **2018**, *6*, 258–270.
50. Gosh, D.; Guha, R. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and geographic information system. *Cartogr. Geogr. Inform. Sci.* **2013**, *40*, 90–102. [CrossRef] [PubMed]
51. Saura, J.R.; Reyes-Menendez, A.; Palos-Sanchez, P. Un Análisis de Sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de# BlackFriday. *Revista Espacios* **2018**, *39*, 16.
52. Palos-Sánchez, P.R.; Arenas-Márquez, F.J.; Aguayo-Camacho, M. Determinants of Adoption of Cloud Computing Services by Small, Medium and Large Companies. *J. Theor. Appl. Inf. Technol.* **2017**, 95.
53. Palos Sánchez, P.R. Estudio organizacional del cloud computing en empresas emprendedoras. *Rev. 3c Tecnol.* **2017**, 6. [CrossRef]