UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science master thesis

Identify factors that determine the start-up ecosystem in Nigeria

First examiner:

Dennis Nguyen

Second examiner:

David Gauthier

Candidate:

Athanasios Souliotis

In cooperation with:

Zhiran Dai

Luiza Swierzawska

Liao Shoufan

July 7, 2023

Abstract

The examination of factors that determine the performance of a company has attracted substantial attention across the literature. However, the factors that define start-up ecosystems have not yet been studied in depth. According to the start-up ecosystem framework, the physical infrastructure, the existing challenge, and the financial opportunities are some of the factors that constitute this ecosystem.

The case of Nigeria is worth noticing. There are multiple arguments that confirm the establishment of Nigeria as the most important start-up hub in the African continent.

In this context, analyzing Twitter provides valuable insights into the aforementioned factors, and leveraging data science techniques allow for a deeper disentangling of these factors.

The most prominent methods to reveal those factors are topic modeling and sentiment analysis.

For this reason, the aim of the present study is to identify factors that determine the start-up ecosystem in Nigeria by implementing hashtag-based topic modeling and unsupervised sentiment analysis with BERT on the Twitter platform.

The main findings of the study revealed that utilizing hashtag-based topic modeling yielded more interpretable topics compared to the baseline model. This approach resulted in the identification of seven distinct factors that were relevant to the research analysis. All of them were sentimentally positive. The most popular factor was the one that referred to the financial issues of start-ups.

The results of this study will serve as a valuable resource for investors who want to fund a Nigerian start-up company, providing them with guidance into the dynamics of the start-up ecosystem.

Contents

1. Introduction

1.1 Content of the introduction

Nigeria is one of the countries with the most start-ups focused on technology worldwide and has played a leading role in the African continent (Disrupt Africa, 2022). More specifically, between 2015 - 2022, the Nigerian technological start-ups received a combined US$2,068,709,445 in funding a higher total than any other country (Disrupt Africa, 2022). According to the Nigerian start-up ecosystem report, the leading sub-sector is fintech with a total of 173 companies (36%) being present, followed by e-commerce and retail tech with 12.1 percent of technological start-ups (Disrupt Africa, 2022). Apart from that, there are plenty of motivations to focus on the case of Nigeria. First of all, despite the huge negative impact Covid-19 brought on the economy, African start-ups managed to flourish due to the opportunities generated by Covid-19 to find alternative ways to live such as making online payments (Fowler, 2021). Secondly, a key moment that created fertile ground for start-ups to grow was the Nigerian Startup Act in 2022, a joint initiative that ensures laws and regulations favor the attraction of investments (“Nigeria Start-up Act”, n.d.). Some benefits this act provides to start-ups are tax and fiscal incentives, access to government grants and loans, credit guarantee schemes, and the establishment of research institutions (“Nigeria Start-up Act”, n.d.). However, the technological start-up still faces challenges that will be addressed in the next section. Consequently, the objective of this study is to examine the factors that determine the tech start-up ecosystem in Nigeria by analyzing Twitter. Twitter is a popular platform to share and exchange ideas about business-related and specifically start-up issues. Therefore, the authors of this project will leverage data science techniques and especially Natural Language Processing (NLP): topic modeling and sentiment analysis. This study aims to offer a comprehensive overview of the factors that define the Nigerian start-up ecosystem and their general sentiment as discussed by Twitter users. As a result, it will provide valuable information to individuals interested in funding startup companies in Nigeria, informing them about key aspects they should be aware of.

In the following section, we will provide an explanation of the factors that constitute startup ecosystems. Additionally, we will present the existing factors in the Nigerian context and provide an overview of the most commonly used data science techniques for detecting these factors in social media platforms

1.2 Literature review

Factors of start-up ecosystem

Taking the initiative to invest in a company entails understanding the key factors an entrepreneurial team will face. The formulation of a new company is achieved based on some contextual drivers that make entrepreneurs invest. In general, examining the entrepreneurial ecosystem is crucial to provide investors with the necessary insights about the advantages and disadvantages of a specific field of investment. Before diving into the thorough exploration of the components an ecosystem consists of, and following E. Stam and Van De Ven (2019, p.812), we define an entrepreneurial ecosystem as a set of factors that are essential to enable entrepreneurship in a particular territory. According to the ecosystem framework as proposed by F. Stam and Spigel (2016, p. 4), the factors entrepreneurs should consider before launching an enterprise are the existing talent in the region, the market demand to satisfy a need of the society that other products have not yet fulfilled, the competition from other products and the available funding sources. Also, the potential regulatory as well as legal challenges for this territory should be taken into account. In addition, the presence of support organizations such as accelerators or incubators can motivate investors.

A study conducted by Mattes et al. (2022) proposed that the main factors determining start-up ecosystems are on the same level as the entrepreneurial ecosystem. Mattes et al. (2022) employed an updated start-up ecosystem framework, which was presented by E. Stam and Van De Ven (2019). In his study, he proposed 7 factors of the start-up ecosystem presented in Table 1.1. The first component is the physical infrastructure in a region which refers to the physical availability and quality of public facilities that allow start-ups to grow. Also, the available talent is related to the level of education and skills possessed by the human capital, especially for young people. Thirdly, knowledge is associated with the current state of progress and advancements within the fields of science and technology. The finance factor incorporates the availability of funding opportunities for start-up companies, including subsidies and grants. Furthermore, leadership implies the importance of start-up figures like CEOs. A fundamental factor for start-up companies is whether they incorporate innovation into their products or services. when starting a company, entrepreneurs possess an innovative idea in this specific region of interest and business sector (Ziakis et al., 2022, p. 20-21). Finally, coaching or mentoring provides guidance and limited funding to start-ups. An indicative example is the accelerator or incubator programs that have been instrumental in helping launch start-ups.

Table 1.1: Factors of start-up ecosystems

Factor
Physical infrastructure
Available talent
Knowledge
Finance
Leadership
Innovation
Coaching/mentoring

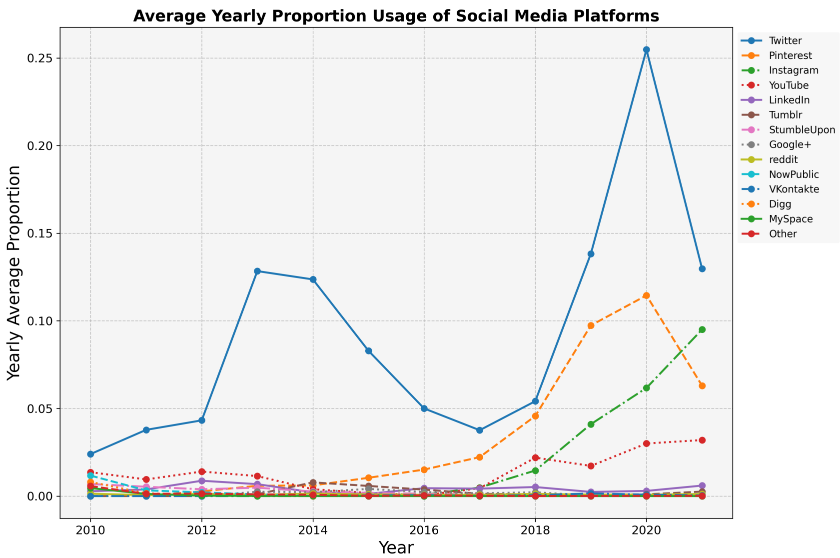
In the Nigerian context, the investors that provide the venture capital for the start-up companies should consider the factors that define the start-up ecosystem (Japan International Cooperation Agency (JICA), 2022). As described in (Japan International Cooperation Agency (JICA), 2022) and shown in Table 1.2, the positive factors outweigh the negative factors and Nigeria is considered a favorable place to start a business. Nigeria has a lot of experience compared to other African countries in launching start-ups, making it an attractive location for start-up investment. Also, while in their early stages, Nigerian start-ups tend to raise funds from local investors, in later stages seek international investors which may lead start-ups to be exposed to international markets support. In addition, most of the start-ups have participated in international accelerator programs. However, negative factors can be considered as limited access to funding, lack of technical skills, and inadequate technological infrastructure (Steamledge, 2023). Finally, about 88.4% of Nigerian start-ups are concentrated in Lagos, resulting in a regional imbalance in entrepreneurial opportunities (Disrupt Africa, 2022). Thus, examining the pros and cons of the Nigerian start-up ecosystem provide valuable information to potential foreign investors that are not aware of the ecosystem.

Table 1.2: Factors that determine the Nigerian start-up ecosystem

Positive Factors	Negative Factors
Longer-established start-ups ecosystem on the continent	Number of start-ups are not evenly distributed
Most of the start-ups have undergone some form of incubation or acceleration by an international program	Inadequate infrastructure
Raise the highest number of investors than any other country in the continent for 2020	Lack of government and corporate engagement and university support
Local investments are crucial in the early stages of a start-up	
Overseas investors are important in later stages	

For this reason, social media analytics are helpful to identify the factors that the entrepreneurial ecosystem consists of because users with diverse backgrounds such as investors or employees in these platforms express opinions and attitudes about them. Evidence shows that user-generated content from social media platforms can be translated into meaningful insights for businesses by leveraging data science techniques (Ruhi, 2014). According to Curran et al. (2011), Twitter has emerged as a popular microblogging platform where individuals actively engage in discussions related to business matters in the sense that it has become a go-to platform for people to exchange thoughts and ideas about various business-related topics. Particularly for Nigeria, Twitter has gained significant traction among users, positioning itself as a leading social media platform in the country. Based on the data shown in Figure 1.1, Twitter stands out as the dominant platform in Nigeria, experiencing exponential growth since 2018 [1]. Over the course of the ten-year period (2010-2020), Twitter has reached its highest peak, averaging a usage rate of 25% in 2020.

Figure 1.1: Usage rate of social media platforms in Nigeria. Source: data accessed and edited from StatCounter. (2023). Social Media Stats Nigeria. Retrieved from https://gs.statcounter.com



Identify factors using data science

According to the recent literature review, the most prominent and commonly used NLP techniques to identify key factors of start-ups that are expressive of positive and negative sentiment as well as success or failure factors of these companies involve topic modeling and sentiment analysis on tweets (Asgari et al., 2022; Saura et al., 2019; Saura et al., 2021). The same methods were used by Saura et al. (2018) to reveal significant factors in tweets that can shape the opinion of users who visit Spanish hotels. Generally, in these studies, a mix of these methods is combined hierarchically to

[1]Facebook was excluded from this plot because its usage rates far exceed those of the other platforms

detect key factors that define the start-up ecosystem in different countries.

Topic modeling

Topic modeling is implemented to reveal the main topics that social media users discuss start-up issues. The most commonly employed topic modeling method in the literature review for identifying indicators related to companies is the Latent Dirichlet Allocation (LDA) (Asgari et al., 2022; Saura et al., 2021; Pournemat & Weiss, 2022a; Singh et al., 2019). LDA will be further explained in the methods section. Consistent with the methods employed in the previous studies, LDA was utilized in conjunction with word cloud visualization to analyze word frequencies within each topic, revealing indicators that reflect the value priorities of social enterprises (Kao & Luarn, 2020). Similarly, to identify indicators of social identity in the #MeToo movement Reyes-Menendez et al. (2020) adopt discourse analysis to identify the frequency of terms as collocations or n-grams.

A comparative study by Albalawi et al. (2020) of various topic modeling techniques applied to short text data demonstrated that LDA and non-negative matrix factorization produced the most coherent topics. This study also emphasized that LDA can be implemented on tweets' hashtags to mitigate data sparsity, as mentioned by Albalawi et al. (2020). In addition, Mehrotra et al. (2013) discovered that hashtag-based topic modeling outperformed other aggregation methods, leading to the highest coherence scores. A study conducted by Culmer and Uhlmann (2021) showed that the lda2vec algorithm, an extension of LDA, combined with pooling techniques, generates more interpretable topics than traditional topic models. These studies shed light on how different variants of LDA can be utilized to generate topics that exhibit coherence.

A new approach to the literature review combines topic modeling with change discovery to find crucial events that affect a business initiative. The change discovery approach is a network analysis method that discovers "breaking events" from documents based on the co-occurrence of words in documents and clustered words (Pournemat & Weiss, 2022a). More specifically, change discovery has been utilized to find product opportunities from customer reviews (Ko et al., 2018). In another study, a hybrid topic model was created to extract topic-sensitive contents from a text corpus. This model considers not only the probability of a sentence being related to a topic but also how strongly connected the sentence is to the topical aspect of interest indicated by the user (Liang et al., 2018). Other studies use descriptive and content analysis techniques through various metrics calculations, including hashtags, word frequencies, and user metrics such as retweets, through word cloud visualizations (Singh et al., 2019, pp. 267-271).

Sentiment analysis

Secondly, sentiment analysis identifies if a topic related to start-up issues has positive or negative sentiment. The provided interpretations of what sentiment means can vary among the academic community, and it relies on the specific context in which sentiment analysis is conducted. Overall, sentiment analysis or opinion mining can be considered the process of analyzing peoples' opinions, attitudes, evaluations, and emotions towards entities such as organizations and products (Liu, 2015, p.1). Sentiment analysis aims to detect a wide spectrum of peoples' beliefs and emotions about reality. The main division of sentiment analysis methods is between polarity classification whether a text has a negative, positive, or even neutral tone, and emotion mining which refers to more complex emotion recognition tasks like finding the presence of specific emotions such as anger and surprise (Yadollahi et al., 2017). Also, sentiment analysis has been studied on different levels and units of analysis. Document-level sentiment analysis examines the sentiment of a whole text, sentence-level pays attention to the short text and phrase-level looks into the sentiment of specific phrases (Wankhade et al., 2022, pp. 5734-5735).

In the field of business, supervised machine learning has been used to identify key factors that contribute negatively or positively to start-up success and indicators that represent the entrepreneurial ecosystem. For example, Asgari et al. (2022) compare the results of three classifiers: Random Forest, Multilayer Perceptron (MLP), and Support Vector Machines (SVM) to predict if a factor of start-up success is negative or positive. MLP was proved as the most accurate method. In another study, Saura et al. (2019)

developed and trained an algorithm with adequate reliability of Krippendorff's alpha of sentiment analysis by using SVM to determine the sentiment of the topic of interest of investors for start-ups. A comparative study of different sentiment analysis methods on Twitter by Giachanou and Crestani (2016) emphasized that the most prevalent classifiers for sentiment analysis are SVM and Naïve Bayes. The unigram-based SVM is usually considered a baseline to develop classifiers with higher-level features. In order to obtain labeled data, several lexicons are employed such as SentiWordNet, MPQA, or VADER to find the sentiment of tweets (Giachanou & Crestani, 2016). One advantage of lexicon-based methods is that do not require annotated data (Giachanou & Crestani, 2016). However, the words that are in the lexicon should be updated frequently and this method ignores the context of the text.

To this end, deep learning techniques that learn multiple layers of features of the text can solve some of the limitations of supervised machine learning algorithms and unsupervised lexicon-based methods. Goularas and Kamis (2019) found that a combination of CNN and LSTM along with GloVe pre-trained word embedding results in higher accuracy in sentiment classification. One of the most famous deep learning techniques for sentiment classification is the Recurrent Neural Network (RNN). The RNNs layers structure allows feedback loops or short memory and for that reason, it is powerful to process sequential input values, like text (Alzubaidi et al., 2021) The output of the current input depends on the output of the previous time step. A more advanced technique deep learning technique is Transformer models which will be explained in the methods section.

1.3 Research questions

The main objective of the present study is to identify the main factors that determine the start-up ecosystem for technological start-ups in Nigeria by using NLP techniques. As previously stated, topic modeling has been extensively employed to reveal factors of start-ups. Therefore, in this study, it is expected that the mechanism of topic modeling will detect those factors of the Nigerian start-up ecosystem. To better understand how the generated topics will give us start-up-related factors, we will demonstrate an example[2]. Assuming we have two tweets: "Start-ups need venture funding to facilitate their growth", and "Investors think start-ups are promising companies for investments". Based on the LDA model, each of the topics is assigned with a probability of words. In our scenario, one topic has this probability distribution: 0.25*venture, 0.17*funding, 0.12*investments, 0.09*growth, 0.06*investors, etc.[3]. So, the words venture, funding, investments, and growth have a higher probability to be assigned on this topic and as result they will be grouped in this topic. Furthermore, this topic can be named "Finance". In addition, the general sentiment of these factors will be revealed by performing sentiment analysis. Finally, this study will contribute to the existing literature on start-ups and will provide meaningful insights to investors that are interested to support a tech start-up in Nigeria. The following research questions are formulated:

- Q1: Can the main topics of discussion for the start-up ecosystem be identified from Twitter users by using hashtag-based topic modeling? Can these topics represent factors of the Nigerian start-up ecosystem?
- Q2: Can the general sentiment (positive or negative) on each factor as represented by the topics and expressed by Twitter users be detected with the BERT model?

[2] Consider that this example does not completely explain LDA but just intuitionally aims to introduce the reader to the topic modeling mechanism. A detailed explanation of LDA will be provided in the methods section

[3] The asterisk (*) denotes the multiplication of the probability by the word

2. Data

2.1 Description of the data

The user-generated content was accessed from the Twitter platform. Relevant information is provided in Table 2.1. A total of 459,894 tweets related to technological issues in Nigeria were scraped using the Twitter API. The scraping of the data was performed by the supervisor of this project. Attributes such as author id, username, numerical information about the tweet itself and the user (i.e., reply count, like count, quote count, retweet count), as well as the text of the tweet, were retrieved.

Table 2.1: Tweets information

Type of tweet	Frequency
Scraped tweets	459894
Retrieved tweets	36166
Retweets	18048
Duplicates	581
Fuzzy duplicates	5287
Final tweets	12250

To focus on Nigerian startups, 36,166 tweets containing keywords such

Figure 2.1: Tweets before preprocessing

	id	tweet
7	8273821554	Lagos, Nigeria has been added on @StartupDiges...
208	14279110335	Why a Sudden Surge in Tech M&A? Startups Pay A...
210	14221652333	Have you read my Memeburn review of Nigeria's ...
211	14221497686	Have you read my Memeburn review of Nigeria's ...
220	13382724346	Great overview of the startup tech scene in Ni...

as "startups", "#startups", "startup" and "startups" were filtered. The data spans from January 27, 2010, to December 30, 2021. Figure 2.1 illustrates the relevant columns, namely the tweet id, the tweet's text, and the date which are essential for our analysis. We removed 18,048 retweets, 581 duplicates, and 5287 fuzzy duplicates that did not provide meaningful information to the dataset, resulting in 12250 tweets. The utilization of the RapidFuzz library [1] was helpful to identify and remove very similar tweets that would add unimportant details. Only the first occurrence of similar tweets was retained

2.2 Preparation of the data

Several preprocessing steps were applied to transform the data into a tidy format. Two custom functions were created for topic modeling and sentiment analysis. The key steps included:

- Conversion of all tweet text to lowercase for consistency.
- Elimination of whitespace characters, URLs, and non-alphabetic characters from the text.
- Retention of hashtags in the text, as they clearly express the topic of a tweet (Alash & Al-Sultany, 2020).
- Utilization of the SpaCy library[2] for part-of-speech tagging (POS tagging). Only proper nouns and nouns were included in the data frame during the feature selection process, as they produce more interpretable topics (Burscher et al., 2016). Emoticons, emojis, and punctuation marks as well as full stops, were maintained for sentiment analysis, as they are indicative of sentiment and BERT can handle them (Pano & Kashef, 2020)
- Modification of a stop words list to exclude words that do not add valuable information, such as prepositions or conjunctions.

[1] For more information, please check here: https://pypi.org/project/rapidfuzz/

[2] For more information, please check here: https://spacy.io/usage/linguistic-features

- Tokenization of the tweets into smaller units (words) to facilitate natural language processing and text mining tasks. Special care was taken to ensure the hashtag symbol ('#') is not separated from the following word during tokenization.
- Lemmatization using the NLTK (Natural Language Toolkit) library[3].Lemmatization groups together different inflected forms of the same word. Lemmatization is only performed for topic modeling and not for sentiment analysis because BERT understands various forms of a word.
- Extraction of hashtags into a new column.

Finally, the tweets were prepared in a cleaner and more structured format, ready for further analysis using natural language processing techniques. After the above preprocessing, the data frame consisted of these columns: id of each tweet, tweet, date, text for topic modeling, text for sentiment analysis, tokens for topic modeling, tokens for sentiment analysis, lemmas for topic modeling, unigrams, bigrams, and hashtags. Each column refers to a specific preprocessing task. Bigrams and unigrams were extracted in columns but we did not use them for further analysis.

[3]For more information, please check here: https://www.nltk.org/_modules/nltk/stem/wordnet.html

3. Method

3.1 Description of the method used

The methodology was divided into two stages: First, LDA was employed to find the main topics of discussion for the start-up ecosystem that are representative of factors expressed by Twitter users. Secondly, sentiment analysis was performed to find the sentiment polarity (negative, positive, or neutral) of these factors as generated by LDA. As previously stated in the literature review, topic modeling has been demonstrably used to reveal the underlying topics discussed by social media users related to business discourses. When implementing the topic modeling technique on a dataset appertaining to the Nigerian start-up ecosystem, it is anticipated that this method will reveal the hidden factors of the start-up ecosystem. The internal workings of topic modeling, when applied to a dataset focusing on start-ups in Nigeria, enable us to cluster tweets that pertain to similar aspects of start-ups into cohesive topics. For instance, if a tweet mentions "venture capital" and another tweet refers to "angel investors" which are both factors within the start-up ecosystem, they may be grouped together in a topic concerning financial issues of start-ups. An in-depth explanation of LDA will follow below.

LDA is the most commonly used method to detect various factors and indicators of business-related issues. LDA is a generative probabilistic method of a corpus of documents in which each document is represented as random mixtures over latent topics and each topic is characterized by a distribution over words (Blei et al., 2003, p.996). It calculates probability distributions of topics in a given document and words in it. At first, the algorithm randomly assigns a topic to every word which results in the formation of a distribution of the topic across the words. For example, topic 1 has the below distribution of words: topic1: 0.4*startups, 0.2*venture capital, 0.4*incubation. Secondly, it randomly assigns the topics to the set of documents. For example, document 1 has the below distribution of topics: document 1: 10% topic1, 40% topic2, 50% topic3. Then it generates a document by combining those two distributions together and randomly selecting the elements from each of the distributions. Finally, the distributions of the generated document are compared with the distributions of the actual document. A strong match between the distributions reflects the original document's topic composition and word distribution. The LDA algorithm iterates until it finds the best match between the actual and generated distributions. As stated by Blei et al. (2003), the underlying mechanism of how LDA works can be explained with the below mathematical formula:

$$P(\theta, \phi, z, w|\alpha, \beta) = \prod_{d=1}^{D} \left(P(\theta_d|\alpha) \prod_{n=1}^{N_d} P(z_{d,n}|\theta_d) P(w_{d,n}|\phi_{z_{d,n}}) \right) P(\phi|\beta)$$

(Blei et al., 2003)

where
θ represents the topic distribution,
ϕ represents the word distribution,
w represents the words in the corpus,
α represents the Dirichlet parameter which controls the topic-word distribution,
β represents the Dirichlet parameter which controls the document-word distribution,

However, LDA assumes multiple probability distributions for each document which might not be aligned with the fact that a tweet is considered short text and is more likely to be associated with one topic. For this reason, in the present study, hashtag-based LDA topic modeling was performed.

More specifically, hashtag-based topic modeling is a process, in which the topics are identified based on hashtags used on tweets. Therefore, the topic models are trained on the hashtags. In this study, it is expected that each tweet will have a higher probability to be assigned to only one topic. This method has been demonstrated to be efficient and has shown superior performance compared to traditional LDA when applied to Twitter data (Alash & Al-Sultany, 2020). The basic assumption is that sharing hashtags means referring to the same underlying topic. One significant drawback of this method is that it heavily relies on hashtags in tweets, which can lead to a loss of information because not all tweets include hashtags. In this study, the number of topics was selected based on the factors presented in the literature review and the coherent score, a measure that defines whether the topics are interpretable or not. (Pournemat & Weiss, 2022b). One of the most popular coherence metrics is called CV. A comparative study of different coherent measures conducted by Röder et al. (2015) showed that the CV metric exceeds the performance of existing measures. This metric uses a context vector which is a count of all the top words that occur in a particular topic. The mathematical formula is explained below:

$$\tilde{v}(W') = \left\{ \sum_{w_i \in W} \mathrm{NPMI}(w_i, w_j)^{\gamma} \right\}_{j=1,\ldots,|W|} \qquad \text{(Syed and Spruit, 2017)}$$

This formula depicts the calculation of the coherence score for a set of words by utilizing the NPMI(Normalized Pointwise Mutual Information). Higher values of CV (close to 1) indicate stronger semantic associations between the words, meaning more interpretable and coherent topics. In addition, the generated topics were interpreted as factors by examining the top keywords and the researchers named them accordingly. Those keywords are actually the hashtags of the tweets. The most representative tweets played a supportive role in the interpretation of the topics. To address ethical issues, all individuals mentioned in tweets, such as Twitter users, were anonymized by replacing their names with the placeholder [mention].

Hashtag-based topic modeling was compared with a baseline LDA model performed on the lemmatized tweets which include only nouns and proper nouns.

Secondly, an algorithm was developed to find the most dominant topic of each tweet which facilitates the sentiment analysis task. Among the various methods of sentiment analysis described in the literature review, this study employs polarity detection to assess the negativity or positivity of a factor. To annotate the unlabeled dataset with positive or negative labels, a deep learning approach and more specifically a pre-trained language model as a state-of-the-art model was utilized. For this reason, BERT, short for Bidirectional Encoder Representations from Transformers is employed. BERT is a pre-training technique that constructs comprehensive representations of text by analyzing both the left and right context on all layers by using a "masked language model". The latter model randomly masks at least 15% of the tokens/words from the input and BERT aims to guess or predict those masked words based on their context by reading the text unidirectionally (Devlin et al., 2019). The second unsupervised task the BERT is trained on is the next sentence prediction. In order to learn sentence relationships the main task works as follows: given a pair of sentences A and B, the task is to say whether sentence B is the actual sentence that comes after (50% of the time) or a random sentence from the corpus (50% of the time) (Devlin et al., 2019). In this study, we deployed a pre-trained model that has been fine-tuned for sentiment analysis in English [1] from the HuggingFace community which is a hub devoted to natural language processing and it includes the Transformers library [2]. To enhance computational efficiency, we chose the default model for sentiment analysis called DistilBERT which is a smaller, faster, and cheaper model trained by distilling BERT case with fewer parameters [3]. Before applying sentiment analysis to the tweets by utilizing

[1]Tweets text along with sentiment labels: negative or positive

[2]More information about the model can be found here: https://huggingface.co/learn/nlp-course/chapter1/3?fw=pt

[3]More information about the model can be found here: https://huggingface.co/docs/transformers/main/en/model_doc/distilbert#transformers.DistilBertForSequenceClassification

the pre-trained model, the sentiment labels (negative or positive) and the sentiment score ranging from -1 to 1 were accessed from the BERT model. A sentiment score above 0 indicates a positive sentiment and a sentiment score below 0 signifies a negative sentiment. A sentiment score close to -1 indicates strong negative sentiment and a score close to 1 indicates strong positive sentiment.

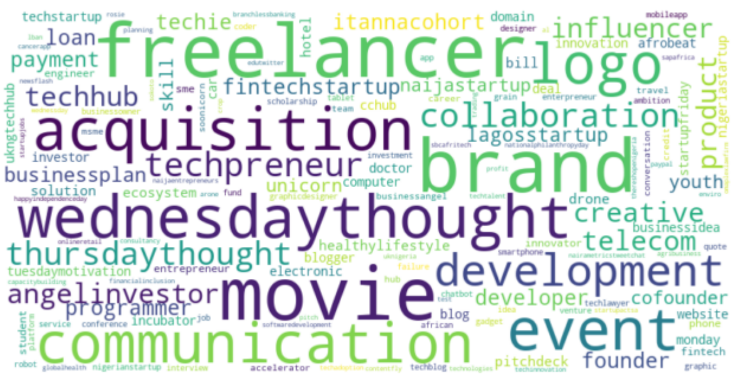
4. Results

4.1 Overview of the results

4.1.1 Exploratory data analysis

After preparing the tweets in a clean and suitable format, it is crucial to conduct exploratory data analysis to gain initial insights into the preprocessed data. This step ensures the successful completion of the data-cleaning process and introduces the reader to basic concepts related to the factors influencing the startup ecosystem in Nigeria. This analysis is conducted before performing hashtag-based topic modeling to uncover the various factors that affect this ecosystem. As a visual representation of the hashtag frequencies, a word cloud is created, as shown in Figure 4.1.

Figure 4.1: Wordcloud: Hashtag frequencies

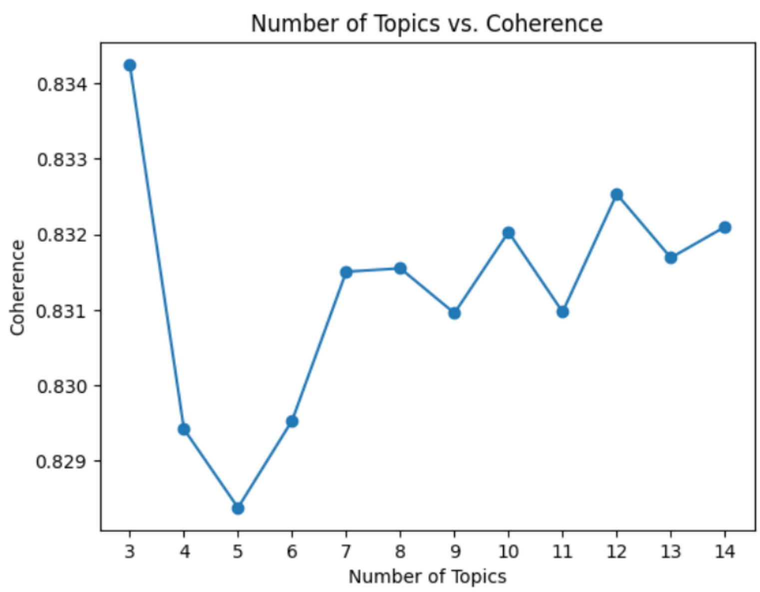


Upon initial observation, the prominent points of discussion among Twitter users regarding the startup ecosystem revolve around "acquisitions" probably of start-up companies. Also, the word "wednesdaythought" is a prominent hashtag referring to general thoughts and comments Twitter users share for the start-up ecosystem. The "freelancer" might be related to technical experts that form their own start-ups. Also, the hashtag "brand" could refer to a new product or service launched by a start-up company. Furthermore, many users are interested in "events" organized probably by tech hubs that promote innovation and "collaboration" between start-up companies in Nigeria. Also, the concept of "development" of start-up companies is extensively discussed on Twitter. In general, analyzing the hashtags used in the tweets provides a clear overview of the topics being discussed on Twitter. In the next subsection, the results of topic modeling will be presented.

4.1.2 Topic modeling

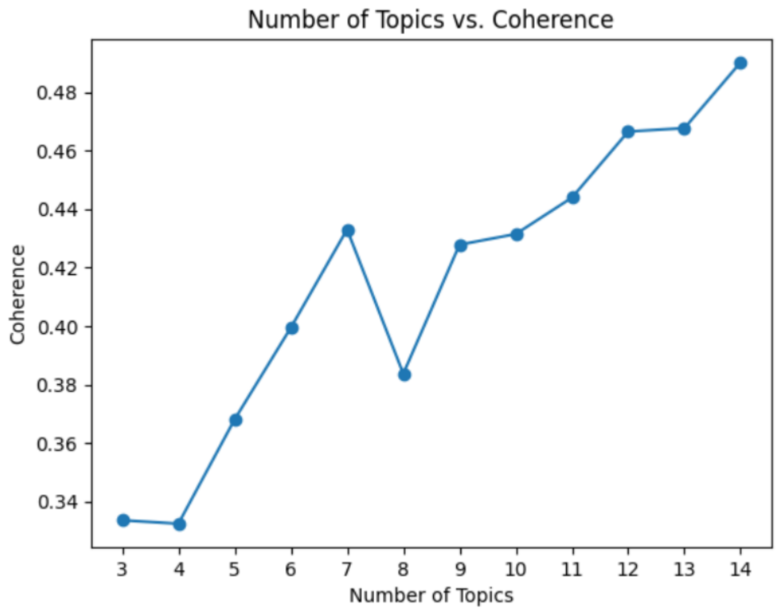
The coherence scores, depicted in Figure 4.2, demonstrate high values ranging from 0.829 to 0.834. Based on these scores, 7 topics were chosen for two reasons: firstly, there is a negligible increase beyond this number until topic 8, and secondly, the factors related to the start-up ecosystem discussed in the literature review were approximately 7.

Figure 4.2: Coherence score plot for hashtag-based LDA



In contrast, the coherence scores depicted in Figure 4.3 for the baseline LDA topic modeling are significantly lower compared to the hashtag-based approach. Therefore, the topics generated by hashtag-based topic modeling were selected due to their higher coherence scores and because the topics

Figure 4.3: Coherence score plot for baseline LDA



were more interpretable. The hashtag-based LDA topic modeling process, as described in the methods section, identified 7 topics, as illustrated in Appendix A. A.1

The first topic includes keywords about "investment", "growth", and "innovation" as well as Uganda and Ethiopia, countries that receive funding for start-ups. So, this topic discusses financial matters in start-ups. Especially, Twitter users discuss issues related to the funding for Nigerian tech start-ups. Venture capital funding usually plays a significant role to support early-stage start-ups with the potential for rapid growth. An indicative tweet highlights how venture funds flow towards start-ups with innovative ideas (see Appendix A, Table A.2). Also, another tweet emphasizes the practices negatively impacting foreign investment in Nigerian start-ups. Furthermore, Twitter users expand the discussion further by pointing out that technological hubs dedicated to innovation invest in start-ups (see Appendix A, Table A.2). A tweet example refers to the Co-Creation Hub located in Lagos city, which is one of the largest economic hubs in Nigeria (see Appendix A, Table A.2).

The primary focus of the second topic centers around the sectors in which startups are most actively involved with keywords such as "edtech", "agritech".

"lagos" which is the city with the most start-ups in Nigeria. According to pertinent tweets, startups show considerable activity in the fields of health technology, agritech, and education technology. Apart from that, this topic includes keywords like "techexperts", "developer" probably referring to tech-savvy individuals who launch start-ups or work for them. Indeed, numerous tweets share inspiring personal narratives of startup founders who have made significant contributions across the above sectors (see Appendix A, Table A.3).

The third topic includes keywords about "education", "ambition", "leadership" etc. It delves into the significance of young entrepreneurs in the startup ecosystem, the transformative impact of startups on the education sector, and the existing challenges related to technology literacy for kids compared to other nations (see Appendix A, Table A.4).

The fourth topic uses keywords such as "money", "apps" and "payments" probably focused on the fintech industry, particularly highlighting acquisitions and deals within the industry. Additionally, the keywords "rwanda" which has many fintech companies, "entrepreneur", "developers" might be related to young entrepreneurs mostly software developers. Some tweets refer to these young people who establish startups and substantiate this claim with success stories of building startups from the ground up (see Appendix A, Table A.5).

The fifth topic explores the role of internet adoption and its implications for the start-up ecosystem in Nigeria by including the keyword "internet" as examples of start-ups that leverage the internet to grow. Some keywords are about "kenya", "egypt, "UK" to draw comparisons between Nigeria's internet usage and that of other nations, understanding its relative standing (see Appendix A, Table A.6). Meanwhile, the keywords "binance" and "designtwitter" underscore how start-ups have harnessed the internet to their advantage, showcasing the positive outcomes it has brought to their operation (see Appendix A, Table A.6). Another topic discussed is the necessity of policy-making to facilitate the development of a thriving ecosystem. The keyword "endsars" refers to the anti-police brutality protests that were predominantly organized by young people in Nigeria through online platforms.

The sixth topic includes the keywords "didyouknow" which refers to tweets that provide tips to a wide spectrum of start-up issues such as investments in "blockchain" and "binance" as well as advice to "entrepreneurs" and "programmers". According to Appendix A, Table A.7. the majority of the tweets provide advice on how to attract grants and loans, how to buy stocks, how to get into the tech industry, and the role of business lobbies on start-up growth. Furthermore, another tweet highlights the significance of the founder's personal traits such as passion in determining the success of a startup.

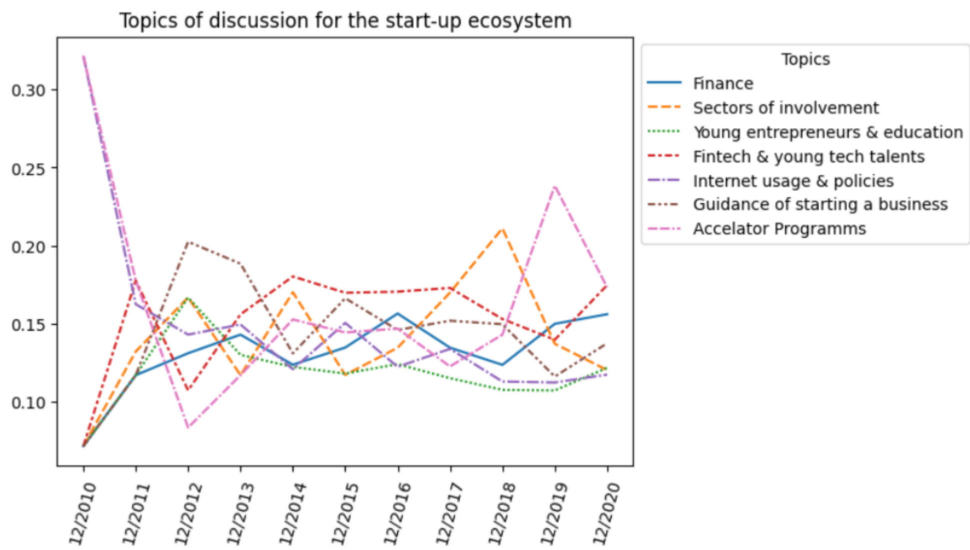
The last topic focuses on the available accelerator programs organized mainly by "south africa", a country with many available accelerator programs that foster tech entrepreneurship in a "techforgood" logic by mainly funding start-up companies through venture capital (keyword: "vc). Also, it seems that some accelerator programs are designed to support and empower women in the tech industry (keyword: "womenintech") (see Appendix A, Table A.8).

All the above 7 topics represent some factors that are found in the Nigerian start-up ecosystem, as explained in the literature review. According to Appendix A, Figure A.1, the most principal factor is the one that refers to financial issues of start-ups with more than 8000 tweets. The rest of the topics share approximately the same number of tweets (fewer than 1500 tweets).

To better capture the discussed topics over time, we will examine Figure 4.4 which shows the average topic distribution for each year. In December 2012, the peak of the discussions about the guidance of starting a business reached because the Nigerian start-up ecosystem began to develop in the years from 2012 (Oxford Business Group, 2023). As a result, the majority of the tweets were focused on discussing business ideas and tips related to starting a start-up in Nigeria.

In 2019, accelerator programs became the focal point of discussion among Twitter users. The flurry of acquisitions, expansions, and funding in the tech industry ignited a storm of conversations surrounding the topic (McKinsey & Company, 2023). In December 2018, the increased interest in sectors of involvement could be attributed to the launch of many start-ups in those sectors during that period (Disrupt Africa, 2022).

Figure 4.4: Average topic distribution for each year



4.1.3 Sentiment analysis

The last part of this study is to detect the positivity or negativity of each factor. As previously described in the methods section the tweets were annotated as positive or negative by the BERT pre-trained model and the sentiment score was accessed ranging from -1 to 1. Figure 4.5 highlights the predominance of positive sentiment. Almost 7000 (61%) out of 10947 tweets related to start-up ecosystem factors express a positive sentiment and about 4000 (39%) tweets express a negative sentiment. Similarly, positive sentiment surpasses negative sentiment for all the factors.

The "accelerator programs" factor has the highest proportion of positive sentiment (71%), followed by the factor related to "guidance of starting a business" (70%) (see Appendix B1, Figure B.2 Figure B.7). The factor related to "internet usage and policies" ranks third in terms of positive sentiment, comprising approximately 67% of the tweets (see Appendix B1, Figure B.6).

The factors related to "sectors of involvement" and "fintech and young talents" account for approximately 62% and 61% of the total tweets, respectively (see Appendix B1, Figure B.4 Figure B5). Finally, the sentiments associated with the "young entrepreneurs education" factor are relatively more balanced compared to the previously mentioned cases. Approximately 59% of the tweets are positive, while 41% are negative (see Appendix B1, Figure B.3).

Figure 4.5: Overall sentiment of the tweets

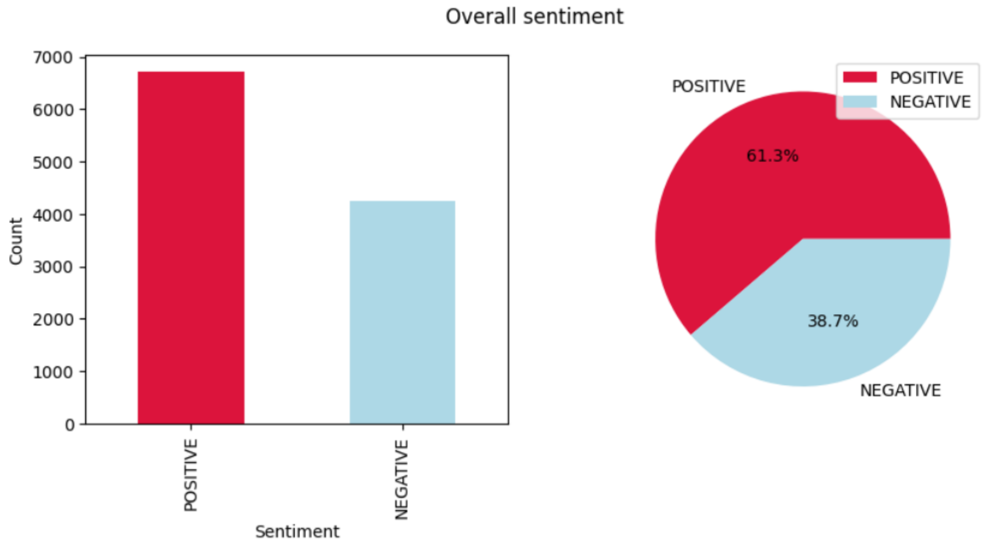
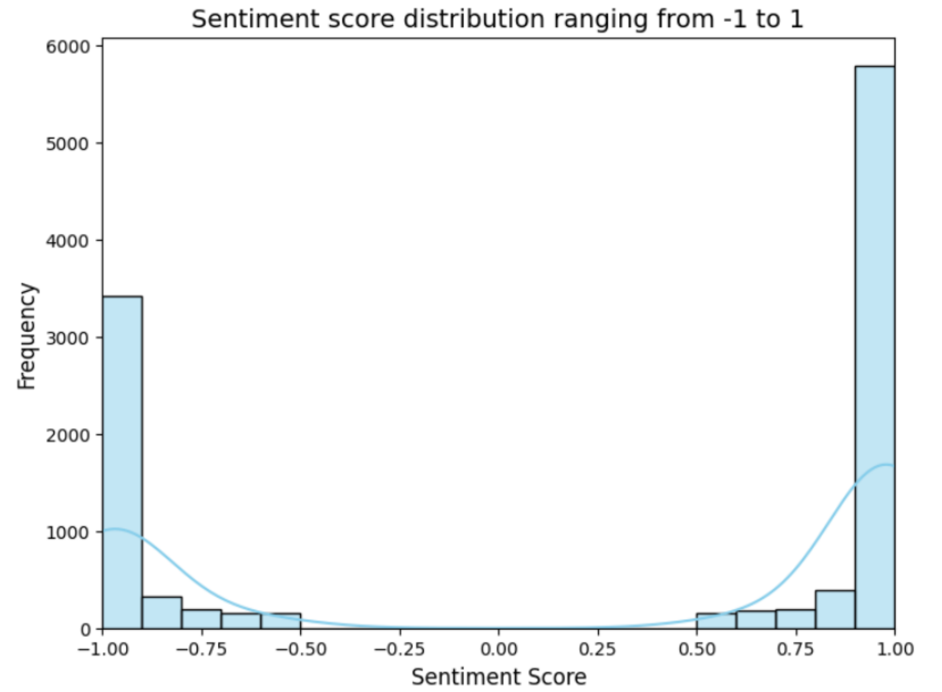


Figure 4.6 demonstrates a bimodal distribution of the sentiment score, indicating the presence of two prominent peaks at the extreme values of -1 and 1. This observation suggests a substantial occurrence of tweets in those regions. Nearly half of the tweets (around 6000) convey an overwhelmingly positive sentiment, while approximately 3200 tweets exhibit a significantly negative sentiment, approaching -1.

Figure 4.6: Sentiment score distribution



To obtain more insights into the overall sentiment score of the factors, we will examine the mean of the sentiment scores. Table 4.1 illustrates that the "accelerator programs" factor has the highest mean sentiment score, followed by the factors "guidance of starting a business" and "internet usage policies". The factors "sectors of involvement", "fintech & young talents", "finance" and "young entrepreneurs & education" have lower sentiment score averages than the first three factors.

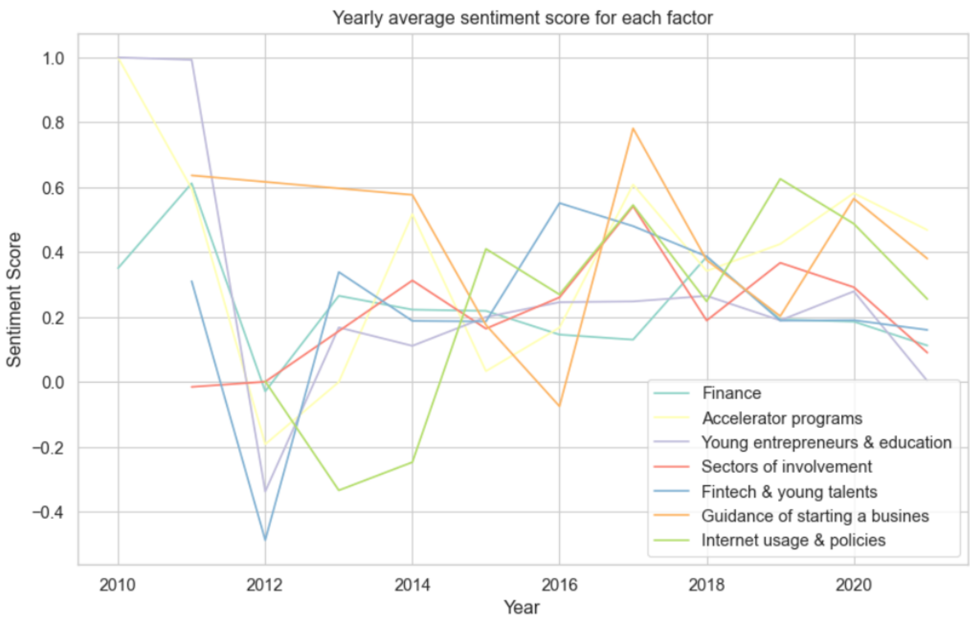
Table 4.1: Mean sentiment scores by factor

Factor	Mean Sentiment Score
Accelerator programs	0.40
Guidance of starting a business	0.38
Internet usage & policies	0.34
Sectors of involvement	0.26
Fintech & young talents	0.25
Finance	0.20
Young entrepreneurs & education	0.18

The dataset used in this study spans ten years, and it is crucial to analyze how sentiments expressed by Twitter users change over time to capture the main trends. Figure 4.7 in the study demonstrates significant changes in sentiment scores for different factors.

For the "fintech & young talents" and "young entrepreneurs education" factors, there was a notable decline in sentiment scores. In 2011, both factors had positive scores, but by 2012, the sentiment scores dropped to below -0.4 for "fintech young talents" and below -0.2 for "young entrepreneurs & education." The "guidance of starting a business" factor reached a high peak with a score of 0.8 in 2017. This peak can be attributed to the launch of many new start-ups in Nigeria during that year, as reported by Disrupt Africa. As a result, Twitter users engaged in discussions about efficient ways to foster the growth and success of these start-ups (Guardian, 2017). Furthermore, the "Internet usage & policies" factor reached a peak sentiment score of 0.6 in 2019. During this period, Nigeria experienced a 14% increase in internet users in the fourth quarter of 2018 compared to the previous year. This growth in internet usage may have had an impact on the usage of the internet by Nigerian tech start-ups, leading to discussions on Twitter about related policies and practices (BusinessDay, 2018).

Figure 4.7: Yearly average sentiment score by factor



5. Conclusion

In conclusion, the present study was an attempt to identify the main factors that shape the Nigerian start-up ecosystem in Nigeria. Hashtag-based LDA resulted in more interpretable topics than the baseline LDA. The identified topics were seven representing factors related to the start-up ecosystem in Nigeria. The most discussed factor by Twitter users was the financial issues of start-ups. Financial issues such as funding opportunities for start-ups specifically in Nigeria play a major role and investors should examine them carefully. Other factors were associated with the sectors in which start-ups are most actively involved, the young entrepreneurs, the fintech industry, and many more. The majority of the tweets expressed a positive sentiment toward the above factors. The sentiment score exhibited high peaks at both ends of the distribution. So, many tweets were either very negative (close to -1) or strongly positive (close to 1). "Accelerator programs" was the most positive factor followed by the factors "guidance of starting a business" and "internet usage & policies". Real-life events linked to the start-up ecosystem were associated with peaks in the sentiment score of the factors. However, those peaks cannot be exclusively attributed to the events. They may be just a coincidence. Based on the findings, the first research question was addressed and answered because the topics were representative of the Nigerian tech start-up ecosystem. It seems that there is an overlap between the factors "Young entrepreneurs education" and "Fintech young talents" as both topics refer to young talents in Nigeria. However, this finding showcases the importance of young talents for investors and also these topics discuss young talents in Nigeria from different perspectives. The second research question was partially answered. Even if the sentiment analysis was achieved to distinguish between positive and negative factors, the polarity of the sentiment did not provide us with adequate information about the sentiment of the factors. In future work, the different categories that demonstrate the intensity of the sentiment (i.e. positive, highly positive, etc.) can be taken into consideration for a more comprehensive understanding of the intermediate sentiment scores.

5.1 Discussion

This study makes a contribution to the existing literature on start-ups by leveraging NLP techniques. It offers valuable insights for individuals interested in funding start-up companies in Nigeria, providing an understanding of the factors that influence the start-up ecosystem and their associated sentiments as expressed by Twitter users. Additionally, the incorporation of BERT as an unsupervised tool for sentiment detection enhances the value of the unsupervised sentiment analysis process, which traditionally relies on lexicon-based methods. However, following the recent study of Saha et al. (2023), in future work the sentiment labels generated by BERT should be compared with labels obtained by human annotators and with a lexicon-based method such as VADER, in order to assess the performance of each method. Furthermore, based on those labeled data, several supervised machine learning and deep learning classifiers could be built to predict new unlabelled tweets and measure their performance. The author acknowledges that Twitter, as a social media platform, may have limitations in providing in-depth insights into such a specific subject. In future research, it is recommended to incorporate websites and blogs that specialize in start-up issues to augment the analysis. Also, hashtag-based LDA resulted in a huge information loss since 7812 tweets did not include hashtags. Due to time constraints, the researcher could not implement a more advanced topic modeling method to avoid the limitations of a hashtag-based approach. However, a Biterm topic modeling approach is suitable for short texts such as tweets. This method learns topics over short texts based on the aggregated biterms in the whole corpus to deal with the data sparsity of each document (Yan et al., 2013).

Acknowledgements

I would like to express my deepest gratitude to my thesis supervisor Dennis Nguyen, Assistant Professor for Digital Literacy and Digital Methods at Utrecht University in the Netherlands. His guidance and support throughout this thesis project have been invaluable. I am also grateful to the Onassis Foundation for their financial support with a scholarship that provided me with the opportunity to pursue my master's degree. I had the pleasure of collaborating with my fellow classmates Zhiran Dai, Luiza Swierzawska, and Liao Shoufan. Their perspectives have enhanced my understanding of the subject matter. Finally, many thanks to the second examiner, David Gauthier, Assistant Professor of Computational Media and Arts in the Department of Media and Culture Studies at Utrecht University, for his auxiliary role in this thesis project.

Appendices

In this section the appendices will be presented. Appendix A includes figures and tables related to the factors of the Nigerian start-up ecosystem. Appendix B includes pie charts about the sentiment proportions of each factor. The code of this project can be found here GitHub Repository on my GitHub account

A. Appendix A

A.1 Factors of the Nigerian Start-up ecosystem

Figure A.1: Tweets count per topic

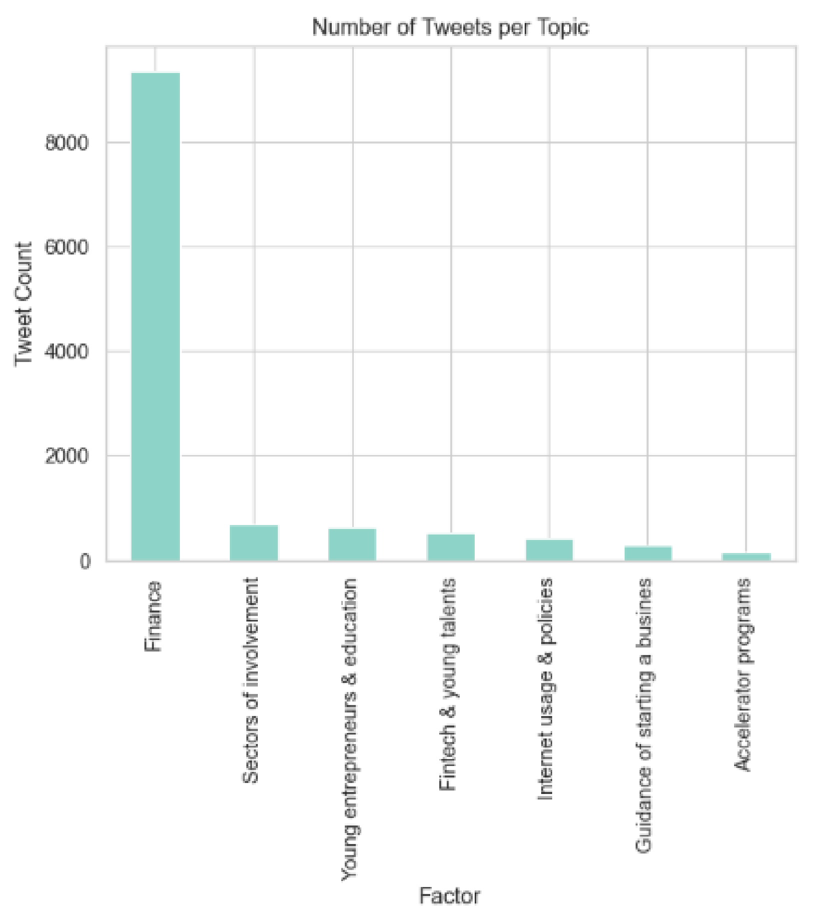


Table A.1: Factors about the Nigerian start-up ecosystem

Factor	Description	Keywords
Finance	This topic discusses the importance of venture capital, the existing challenges about investments, and the supporting role of tech hubs to start-ups	innovation, technews, investment, online, uganda, growth, investors, ethiopia, webdesign, stolen
Sectors of involvement	This topic focuses on the sectors in which start-ups are most actively involved by presenting personal narratives of founders	lagos, teckexperts, designtwitter, endsars, covid, bitcoin, developer, binance, edtech, agritech
Young entrepreneurs & education	This topic talks about the role of young entrepreneurs in the start-up scene, the impact of start-ups on education, and the challenges related to technology literacy for kids	ghana, ai, education, science, leadership, elitepath, ambition, ilovemyjob, zambia, relaxation
Fintech news & young tech talents	This topic delves into the fintech trends and young tech talents (mostly software developers) that formulate their own enterprises	entrepreneur, g, money, apps, developers, rwanda, Twitter, software, healthtech, payments
Internet usage & policies	This topic explains the role of Internet usage, its impact on the Nigerian start-up ecosystem, and the need for policymaking to grow the ecosystem	kenya, internet, fact, covid, designtwitter, endsars, egypt, uk, investors, binance
Guidance of starting a business	This topic discusses the need for information and guidance on investing and growing a start-up business	blockchain, entrepreneurs, apple, didyouknow, crypto, portharcourt, binance, programmer, artificialintelligence, techstartups
Accelerator programs	This topic centers around accelerator programs	southafrica, news, entrepreneurship, vc, socialmedia, growth, analytics, womenintech, agritech, techforgood

Table A.2: Tweets about finance factor

Tweet
This is currently going in on in tech in Nigeria, that's why we have seen foreign investment has dried up, local investors only reveal information to selected friends or introduce startup that has no hope of expansion to foreign investors, who will bypass the market eventually.
Money follows innovation returns, my friend. In 2019, VCs invested US$1.3bn into African tech startups: Nigeria: 49.5% or US $643mn Kenya: 32% or US$416mn (Sh41.6bn) Others: 18.5% Kenya has an attractive environment for innovation... talent included.

Table A.3: Tweets about sectors of involvement factor

Tweet
Are good hospitals the norm or the exception? Do we have them everywhere? I come from a very health-based area, and my first startup was in health tech, and the numbers are staggering. Comparing the health sector of the US with that of Nigeria is a stretch.
As much as I know Nigeria has the talents to move into these sectors healthtech, edtech, agritech and all, there are still underlying factors. But we have startups making waves in the health tech, edtech, and agritech.
[mention] is a medical doctor and founder of health tech start-up, flyingdoctors Nigeria. Her startup is poised to bringing emergency care to the parts of Nigeria which is unmotorable through her air ambulance which is in Lagos.
This was a good read.[mention] is doing commendable work with edtech in Nigeria through [mention]. You should read his journey here: My Life in Tech: [mention] journey to building an investable edtech startup.
[mention] is advancing health tech in Nigeria and working on affordable solutions that will improve people's lives! Congratulations on becoming our next DueDash Gold Ambassador! #community #startups #nigeria #tech #health #ambassador.
Before the fellowship, I ran a nonprofit organization, though I've had the thought to build a startup product but don't know how to. Today, I have built a health tech startup that allows people to get access to well-vetted licensed health professionals —[mention], Nigeria.

Table A.4: Tweets about young entrepreneurs factor

Tweet
[mention] is one of the largest tech community, actively educating, inspiring, and connecting more than 10,000 young entrepreneurs in over 36 universities in Nigeria. powered by M.I.T, fluxtechafrica and Lite House Studios Coming soon... #Tech #startup #africa #entrepreneurs
[mention]. Because tech is everyday life for the Chinese, 12yr olds are designing standard apps already unlike Africa were most kids don't even have access to quality education. The Chinese hv millions of tech startups already. Besides Africa(etc Nigeria) need to start making its own food.
Listen to the Lagosian in New York City podcast: guest is [mention], founder PLEDRE | Reinventing education in Africa using digital technology.
New tech startups are creating jobs in Kenya, Nigeria, Ghana, Uganda and elsewhere with young entrepreneurs.
[mention] has launched a Startup Academy for employees

Table A.5: Tweets about fintech news & young tech talents factor

Tweet
Nigerian fintech Abeg faces its biggest test yet after blitz scaling to millions of users: Piggy Tech Is the parent company of PiggyVest, Nigeria’s most popular savings app and one of the country’s most valuable fintech startups
Kenya’s insurance tech startup GrassRoots Bima and Nigeria’s payments company, Flutterwave and financial management app, Riby have been named amongst the 50 best emerging fintech companies in the world.
Global technology company... is a startup that specializes in training software developers in Africa and hiring them out to tech companies around the world in need of their skills and talents
More than a third of the total funding into African tech went to Nigerian startups, though the number of deals was less than three other countries. Nigeria also stands out as Africa’s fintech hub receiving 66% of total fintech investments into the continent in 2019
The release of 400 developers may be welcome in Africa’s most active tech hubs, such as Nigeria and Kenya, where rapid startup formation and funding is starting to outpace software engineering talent — according to a number of founders

Table A.6: Tweets about internet usage & policies factor

Tweet
Global #internet usage: #Nigeria and #SouthAfrica ahead of #India #digital #tech #africa #web #startup #ecosystem
Hotel startup is doing study abroad admissions, an internet startup that can't get their service to average standard in Nigeria just raised $3m to expand across Africa. Wo I don't even know what's going on in this Tech ecosystem. Wash wash wash!!!!
Tech Startups: 5 biggest drawbacks to Nigeria's Internet ecosystem A technology ecosystem cannot be well harnessed...
We understand how important it is for startups to be at the front of the policy making process when creating policies to enable the tech ecosystem in Nigeria. That is why we created EDS, a space for startups; government to discuss policy recommendations for a thriving ecosystem
Unpopular opinion. 10+ years of startup culture in Nigeria had built the most dynamic startup ecosystem in Africa. But what are the merits/demerits of not having the startup ecosystem as an element in a science and tech ecosystem? This convos need to happen without emotions. [mention] I mean in really investing in the tech Startups in Nigeria, specifically internet!
IBAKA TV set to debut plug and play internet TV in Nigeria #tech #startup

Table A.7: Tweets about the guidance of starting a business factor

Tweet
Follow us now and as we share unique business ideas, tips, tech startups, career management and insight on Nigeria business. Have you started a business or about to start one? A major step in attracting grants, loans, and investments for your business is COMPANY REGISTRATION, send us a DM and get your registration sorted ASAP #Flutterwave #StartUp #tech #investments #Africa #Nigeria
Na only relationship this once sabi talk for space. Nobody wan talk about how to buy us stocks, how to survive in Nigeria with a particular job , how to expand your small scale business, how to get into the tech business , how to invest in startups,how to japan
With $3 million from a pre-Series A funding, Nigeria's leading wealth management startup can expand throughout Africa, only 4 years after its founding... Cowrywise, drop these cowries on the mat. Make us as wise. Show us the way. Cut soap for us!
[mention] Tech-hub Incubation Programme 2020 for Nigeria Startups -Cash prizes Teams will have access to shared office space with Internet, hands-on business advisory support, $10,000 credit and business support up to $5,000 from AWS Apply.
A new kind of business lobby is needed in #Nigeria to fight for their rights. Driven by Social Media and inclusive of the entire spectrum of business from the [mention] to the Tech Startup cc [mention] [mention] [mention] [mention]
What other feature will You like on your business webpage? #MerryChristmas #business #ourarea #startup #Marketing #Nigeria #MondayMotivaton #madeinnigeria #ourarea.com.ng #tech TedTalk
Course of study, ought not to be D catalyst for starting a business, Dt is D role for D passion of D business owner. Some tech giants of today, were started by people who dropped out of College. To get Nigeria working, establish Startup accelerators in Universities, NYSC; States

Table A.8: Tweets about accelerator programs factor

Tweet
The good news in Lagos. US tech giant @Google launched the new accelerator project to help the 12 african startups!!! #Nigeria #startups #funding #acceleratore #Africa #Investment #Google
Na only relationship this once sabi talk for space. Nobody wan talk about how to buy us stocks, how to survive in Nigeria with a particular job , how to expand your small scale business, how to get into the tech business , how to invest in startups,how to japan
Tech start-ups blossom as #accelerator programmes fill #funding gap in #Africa#startup #vc #Nigeria via [mention]
#News: [mention] launches #Itanna, an accelerator program for tech #startups in #Nigeria; The accelerator is launching with four companies and each of them will receive $25,000 - via [mention] #AAAPAfrica #Innovation #Africa #Technology
[mention] from Nigeria and [mention] from Ghana have been selected for F-Lane accelerator, which is Europe's first accelerator with a focus on #startups by and for women in the tech #Africa #Nigeria #Ghana #Accelerator]
UK-Nigeria Tech Hub launches COVID-19 startup accelerator programme #startups #accelerator #venturecapital #seedinvestment #funding
[mention] and [mention] : call for applications into the FbStart Accelerator. The accelerator (Nigeria's first of its kind) is designed for innovative startup and student teams using deep-tech to build solutions
African Tech Startups Selected for the Catalyst Fund's Accelerator #HSNews #HorsesStableNews #HorsesStable #Africa #nigeria #paymenow #southafrica #fintechstartups #gatesfoundation #bfa #WellaHealth

B. Appendix B

B.1 Sentiment proportions of each factor

Figure B.1: Finance: Sentiment proportions

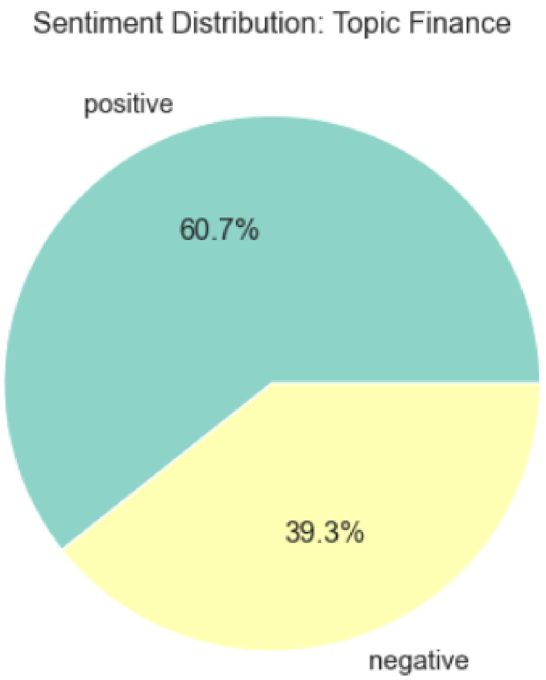


Figure B.2: Accelerator programs: Sentiment proportions

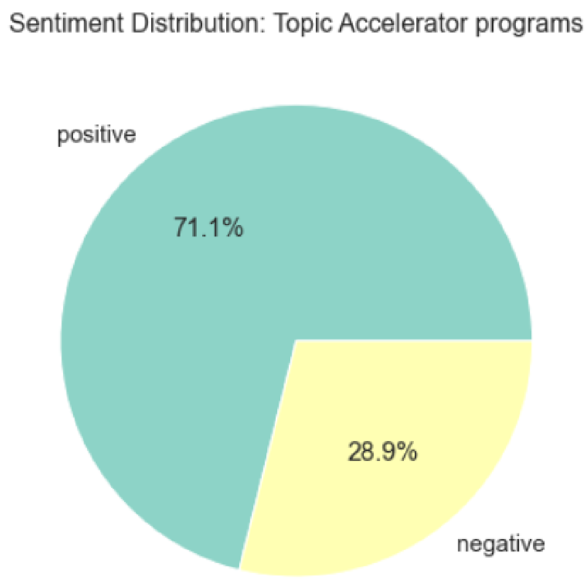


Figure B.3: Young entrepreneurs & education: Sentiment proportions

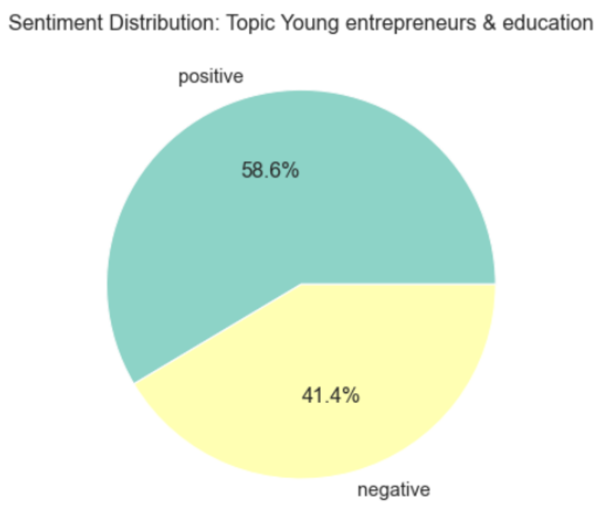


Figure B.4: Sectors of involvement

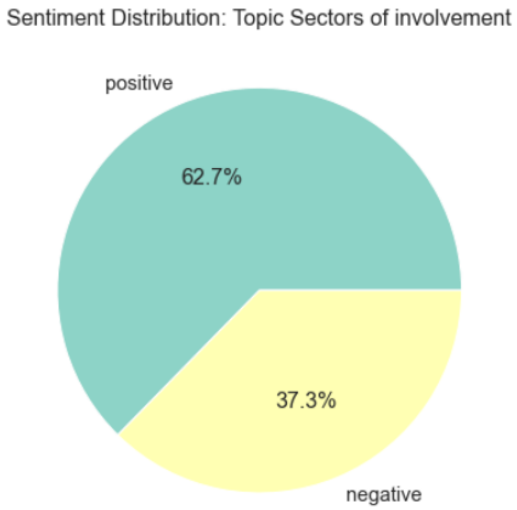


Figure B.5: Fintech & young talents: Sentiment proportions

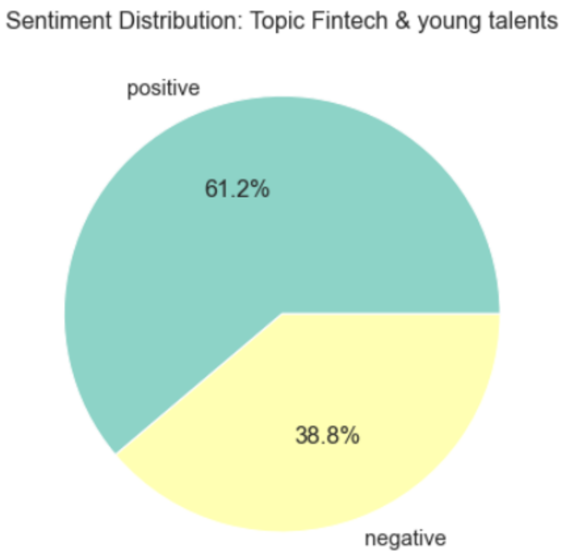


Figure B.6: Guidance of starting a business: Sentiment proportions

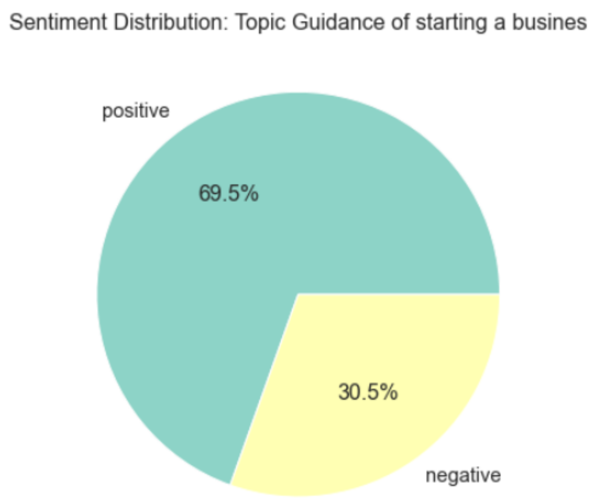
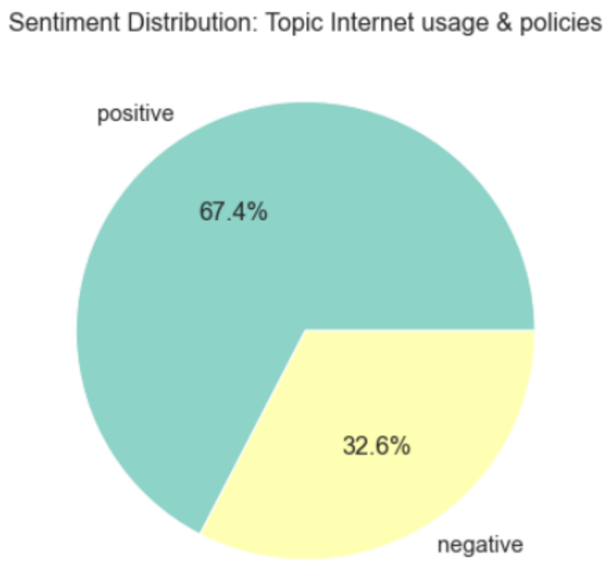


Figure B.7: Internet usage & policies: Sentiment proportions



Bibliography

Disrupt Africa. (2022). *The nigerian start up ecosystem report*. Retrieved May 5, 2023, from https://disrupt-africa.com/wp-content/uploads/2022/09/The-Nigerian-Startup-Ecosystem-Report-2022.pdf

Fowler, G. (2021). *African startups going global for success*. Forbes. Retrieved June 12, 2023, from https://www.forbes.com/sites/forbesbusinessdevelopmentcouncil/2021/05/24/african-startups

Nigeria start-up act. (n.d.). Nigeria Start-up Act. Retrieved June 12, 2023, from https://nigeriastartupact.ng

Stam, E., & Van De Ven, A. H. (2019). Entrepreneurial ecosystem elements. *Small Business Economics*, *56*(2), 809–832. https://doi.org/10.1007/s11187-019-00270-6

Stam, F., & Spigel, B. (2016). Entrepreneurial ecosystems. *USE Discussion paper series*, *16*(13).

Mattes, V., Leible, S., & Schallmo, D. (2022). Start-up ecosystems: An approach for analysis and success factors. *ResearchGate*. https://www.researchgate.net/publication/366199902_Start-up_Ecosystems_An_Approach_for_Analysis_and_Success_Factors

Ziakis, C., Vlachopoulou, M., & Petridis, K. (2022). Start-up ecosystem (stupeco): A conceptual framework and empirical research. *Journal of Open Innovation: Technology, Market, and Complexity*, *8*(1), 35. https://doi.org/https://doi.org/10.3390/joitmc8010035

Japan International Cooperation Agency (JICA). (2022). *Title of the report*. Japan International Cooperation Agency (JICA). Retrieved June 12, 2023, from https://www.jica.go.jp/nigeria/english/office/topics/gh13tc000000sz4y-att/220328_01.pdf

Steamledge. (2023). *Future of tech in nigeria: Trends and opportunities in the nigerian tech space*. https://steamledge.com/articles/future-of-tech-in-nigeria-trends-and-opportunities-in-the-nigerian-tech-space/

Ruhi, U. (2014). Social media analytics as a business intelligence practice: Current landscape & future prospects. *Umar Ruhi (2014)," Social Media Analytics as a Business Intelligence Practice: Current Landscape & Future Prospects", Journal of Internet Social Networking & Virtual Communities, 2014*.

Curran, K., O'Hara, K. L., & O'Brien, S. M. (2011). The role of twitter in the world of business. *International Journal of Business Data Communications and Networking*, *7*(3), 1–15. https://doi.org/10.4018/jbdcn.2011070101

Asgari, T., Daneshvar, A., Chobar, A. P., Ebrahimi, M., & Abrahamyan, S. (2022). Identifying key success factors for startups with sentiment analysis using text data mining. *International Journal of Engineering*

Business Management, *14*, 18479790221131612. https://doi.org/10.1177/18479790221131612

Saura, J. R., Palos-Sanchez, P., & Grilo, A. (2019). Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability*, *11*(3). https://www.mdpi.com/2071-1050/11/3/917

Saura, J. R., Reyes-Menéndez, A., deMatos, N., & Correia, M. B. (2021). Identifying startups business opportunities from ugc on twitter chatting: An exploratory analysis. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(6), 1929–1944. https://www.mdpi.com/0718-1876/16/6/108

Saura, J. R., Palos-Sanchez, P., & Rios Martin, M. A. (2018). Attitudes expressed in online comments about environmental factors in the tourism sector: An exploratory study. *International Journal of Environmental Research and Public Health*, *15*(3), 553. https://doi.org/10.3390/ijerph15030553

Pournemat, M., & Weiss, M. (2022a). Identifying business opportunities using topic modeling and chance discovery. *LUT Scientific and Expertise Publications:* https://www.researchgate.net/publication/362782899_Identifying_business_opportunities_using_topic_modeling_and_chance_discovery

Singh, S., Chauhan, A., & Dhir, S. (2019). Analyzing the startup ecosystem of india: A twitter analytics perspective. *Journal of Advances in Management Research*, *17*(2), 262–281. https://doi.org/10.1108/jamr-08-2019-0164

Kao, S., & Luarn, P. (2020). Topic modeling analysis of social enterprises: Twitter evidence. *Sustainability*, *12*(8), 3419. https://doi.org/10.3390/su12083419

Reyes-Menendez, A., Saura, J. R., & Thomas, S. H. (2020). Exploring key indicators of social identity in the metoo era: Using discourse analysis in ugc. *International Journal of Information Management*, *54*, 102129. https://doi.org/10.1016/j.ijinfomgt.2020.102129

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in artificial intelligence*, *3*. https://doi.org/10.3389/frai.2020.00042

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. https://doi.org/10.1145/2484028.2484166

Culmer, K., & Uhlmann, J. (2021). Examining lda2vec and tweet pooling for topic modeling on twitter data. *WSEAS transactions on information science and applications*, *18*, 102–115. https://doi.org/10.37394/23209.2021.18.13

Ko, N., Jeong, B., Choi, S. C., & Yoon, J. (2018). Identifying product opportunities using social media mining: Application of topic modeling and chance discovery theory. *IEEE Access*, *6*, 1680–1693. https://doi.org/10.1109/access.2017.2780046

Liang, Y., Liu, Y., Feng, D., & Jiang, Z. (2018). Extracting topic-sensitive content from textual documents—a hybrid topic model approach. *Engineering Applications of Artificial Intelligence*, *70*, 81–91. https://doi.org/10.1016/j.engappai.2017.12.010

Liu, B. (2015, June 1). *Sentiment analysis: Mining opinions, sentiments, and emotions*. http://ebooks.cambridge.org/ref/id/CBO9781139084789

Yadollahi, A., Shahraki, A. F., & Zaïane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, *50*(2), 1–33. https://doi.org/10.1145/3057270

Wankhade, M., Rao, A. C. S., & Kulkarni, C. A. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys*, *49*(2). https://doi.org/10.1145/2938640

Goularas, D., & Kamis, S. (2019). Evaluation of deep learning techniques in sentiment analysis from twitter data. *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*, 12–17. https://doi.org/10.1109/Deep-ML.2019.00011

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A. Q., Duan, Y., Al-Shamma, O., Santamaría, J. V. G., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00444-8

Alash, H. M., & Al-Sultany, G. A. (2020). Improve topic modeling algorithms based on twitter hashtags. *Journal of physics*, *1660*(1), 012100. https://doi.org/10.1088/1742-6596/1660/1/012100

Burscher, B., Vliegenthart, R., & Vreese, C. H. D. (2016). Frames beyond words. *Social Science Computer Review*, *34*(5), 530–545. https://doi.org/10.1177/0894439315596385

Pano, T., & Kashef, R. (2020). A complete vader-based sentiment analysis of bitcoin (btc) tweets during the era of covid-19. *Big data and cognitive computing*, *4*(4), 33. https://doi.org/10.3390/bdcc4040033

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Pournemat, M., & Weiss, M. (2022b). Identifying business opportunities using topic modeling and chance discovery.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. https://doi.org/10.1145/2684822.2685324

Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174. https://doi.org/10.1109/DSAA.2017.61

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Oxford Business Group. (2023). Innovation destination: The domestic start-up ecosystem continues to attract investment as public-private partnerships seek to improve capacity. *Oxford Business Group Report*. https://oxfordbusinessgroup.com/reports/nigeria/2023-report/ict/innovation-destination-the-domestic-start-up-ecosystem-continues-to-attract-investment-as-public-private-partnerships-seek-to-improve-capacity-analysis/

McKinsey & Company. (2023). Harnessing nigeria's fintech potential. *McKinsey & Company Featured Insights*. https://www.mckinsey.com/featured-insights/middle-east-and-africa/harnessing-nigerias-fintech-potential

Guardian, T. (2017). Nigeria's tech startups raise $114.6m in 2017. *The Guardian*. https://guardian.ng/technology/nigerias-tech-startups-raise-114-6m-in-2017/

BusinessDay. (2018). Nigeria's internet users rise by 14% in q4 2018. *BusinessDay*. https://businessday.ng/technology/article/nigerias-internet-users-rise-by-14-in-q4-2018/

Saha, S., Showrov, M. I. H., Rahman, M. M., & Majumder, M. Z. H. (2023, January 1). *Vader vs. bert: A comparative performance analysis for sentiment on coronavirus outbreak*. https://doi.org/10.1007/978-3-031-34619-4_30

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. https://doi.org/10.1145/2488388.2488514