World Academics Journal of Engineering Sciences
Vol.8, Issue.3, pp.28-32, September (2021)

Research Paper

E-ISSN: 2348-635X

Detecting Syriatel's Success Indicators Using Textual Data Mining for Arabic Language

Hanan Mohsen Wassouf[1*], Mohammad-Bassam Kurdy2

[1]Master in Web Science, Syrian Virtual University, Damascus, Syria
[2]Ph.D in Mathematical Morphology, Syrian Virtual University, Dijon, France

**Corresponding Author: hanan_110501@svuonline.org*

Available online at: www.isroset.org

Received: 03/Aug/2021, Accepted: 20/Sept/2021, Online: 30/Sept/2021

***Abstract*—**Despite the broad knowledge and skills of the millennial generation, recent studies indicate that 90% of Companies are classified as failing companies, and usually end in bankruptcy due to many factors, including a departure from the market need, as the percentage of companies that built their activity around processing a good idea, but far from the needs of the market to bankruptcy 42% of the total companies. In addition to others reasons like the lack of liquidity and incomes with less capital than necessary, the weakness of the founding team and the emergence of many competing companies.
All these reasons lead us to think about studying and identifying the success factors for these emerging companies in order to mitigate the possibility of failure.
The target study will be Syriatel Mobile Telecom company which is most famous telecom company in Syria.

Keywords— Indicators Success , Data Mining, Textual Analysis, Topic Modeling

I. INTRODUCTION

Despite the broad knowledge and skills of the millennial generation, recent studies indicate that 90% of Companies are classified as failing companies, and usually end in bankruptcy due to many factors, including a departure from the market need, as the percentage of companies that built their activity around processing a good idea, but far from the needs of the market to bankruptcy 42% of the total companies. In addition to others reasons like the lack of liquidity and incomes with less capital than necessary, the weakness of the founding team and the emergence of many competing companies[1].

All these reasons lead us to think about studying and identifying the success factors for these emerging companies in order to mitigate the possibility of failure.
The target study will be Syriatel Mobile Telecom company which is most famous telecom company in Syria.

II. RELATED WORK

This section exhibits a number of related previous studies as this paper adopts some of their approaches and overcome the absence of some points for Arabic in others.
Saura, Palos-Sanchez, & Grilo [1] is one of the most famous study that cited by most researches in this field,They used new techniques in the proposed research methodology to determine the key factors for the success of these companies' projects. The LDA model, which is a sophisticated modeling tool that works in Python and determines the subject of the database by analyzing the tweets that include the hashtag #Startups, was used, and then a sentiment analysis was performed using the SVM algorithm To divide the subjects into negative, positive and neutral feelings, then a textual analysis was performed on the topics in each feeling using text data mining techniques using Nvivo software.

Saura & Bennett [2] proposed a three-step research methodology based on textual data mining, this methodology can be used for Business Intelligence analysis strategies to analyze UGC in social networks and digital platforms. Then it was proposed to analyze the feelings on the results of LDA to divide the topics identified in the sample into three feelings. Finally, textual analysis with textual data mining techniques was applied to the subjects in each sense.

Saura, Reyes-Menendez, & Palos-Sanchez [3], In this study, a new and innovative data analysis process aimed at introducing digital marketing strategies based on promotions was developed with the focus on Black Friday 2018 in Spain by analyzing the contents published by technology companies on Black Friday 2018, including offers, discounts, and promotions exclusivity, as well as fraud and insults.

Vasile,Elena,Doina,& Magdalena [4] extracted large data from online Amazon reviews since Amazon is considered one of the most important online marketplaces for purchasing products.
Focused on the mobile category, this study proposes a further step in the application of sentiment analysis by

adding selection of product features and calculating additional information from online reviews by extracting the text of online user-generated reviews and increasing the accuracy of the analysis at the sentence level, this proposal will help managers and users to make better decisions in analyzing their products and purchases.

Al-Kabi,Al-Qudah,Dabour & Izzat Alsmadi [5] in this study a primary dataset was compiled in English/Arabic containing 4050 English/Arabic comments and reviews generated by users of social networking sites, and the best 45 words and phrases were used for each of the three categories (academic, news, and business).

Annoying and loud comments have been removed to avoid inconsistencies. In addition, comments and reviews have been removed to ensure the uniqueness of the content of the data set. This data set was used to create three polar dictionaries: (Arabic, English and symbols) These dictionaries were used to evaluate Social Mention experimentally.

III. METHODOLOGY

This section presents method for detecting success indicators using textual data mining for Arabic language.
We will work on using emerging technology mechanisms and technologies such as data mining, sentiment analysis, and textual analysis, for determining the main factors for Syriatel's success by analysing comments on the Facebook social network. This helps for discover indicators success and reduces failure rates for this company.

The study consists of the process as shown in figure(1):

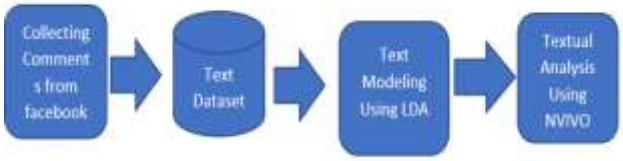


Figure1. Work methodology diagram

DETAILS OF PROPOSED METHODOLOGY:

The approach contains the main stages:

1) Use the Latent Dirichlet Allocation (LDA), a sophisticated modeling tool that works in Python and identifies the subject of the database.
2) Conducting a textual analysis of the subjects in each topic using textual data mining techniques with Nvivo software.
3) Creating nodes in NVIVO to detect indicators succeess by improving company's product and suggesting new product.

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 facebook posts for syriatel offers(sabaya and army)

Second, a Textual Analysis was performed on the results with Text Data Mining techniques using the Nvivo qualitative analysis software

Third, indicators success were discovered by coding the text and creating nodes in nvivo to improve network service in Syriatel and innovate new product for other people like Shabab or young people.

Details design of the proposed approach:
The sample for this study was structured using information from facebook posts for syriatel offers(sabaya and army) for a sample of 1200 comments stored in excel file.

1) **Topic Identifications Using LDA:**
LDA is a generalization of older approach of Probabilistic latent semantic analysis (pLSA), The pLSA model is equivalent to LDA under a uniform Dirichlet prior distribution. both methods are similar in principle and require the user to specify the number of topics to be discovered before the start of training (as with K-means clustering)

The first step of lda model 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 as below(1):

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

where bi is the distribution of a word in topic i, with total K topics; qd is the proportion of topics
in document d, with total D documents; zd is the topic assignment in document d; zdn is the topic
assignment for the nth word in document d, with total N words; wd is the observed words for document
d; and wd,n is the nth word for document d.
Then,the topics and words were identified using Equation(2).

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

*2)**Textual Analysis:***
NVivo is one of the computer-assisted qualitative data analysis softwares (CAQDAS) developed by QSR International (Melbourne, Australia), the world's largest qualitative research software developer. This software allows for qualitative inquiry beyond coding, sorting and retrieval of data. It was also designed to integrate coding with qualitative linking, shaping and modelling.

An important indicator for the analysis using Nvivo is the weighted percentage , 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:

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

*3)**Coding Text and Creating Nodes:***

The application of Nvivo software can use feature auto code

Also Arabic language is not supported in nvivo we have created three nodes and the words were grouped into these nodes according to the number of times the words were repeated in the dataset.

IV. RESULTS AND DISCUSSION

1) Latent Dirichlet Allocation (LDA) Results:

The topics identified with the LDA are shown in Figure (1) as below, In the LDA process, the words were automatically categorized into topics, and the researchers gave each topic a name after analysing the group of words. Manually naming the topics is a standard procedure in LDA-based topic identification. 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.

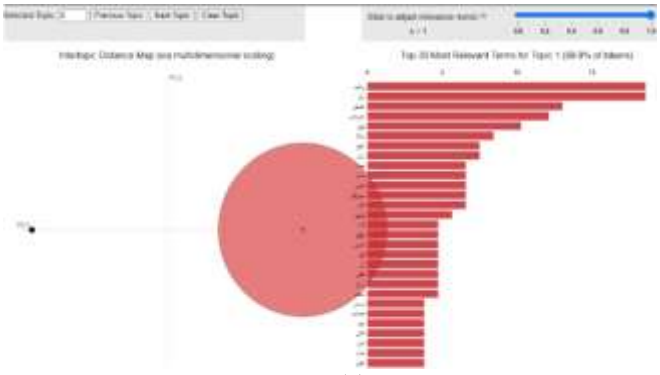
Topic	Weight
راتب	0.022
باق	0.022
عسكر	0.016
عرض	0.015
يعن	0.012
شبك	0.010
غيغ	0.009
ديار	0.009
حما	0.008
مف	0.008

Figure(2)

We can see the result in Figure(4) by using the function prepare in Figure(3) after importing the module pyLDAvis and passing lda_model as parameter to it as below:

```
>>> import pyLDAvis
>>> import pyLDAvis.gensim
>>>
>>> vis = pyLDAvis.gensim.prepare(topic_model=lda_model,
...                               corpus=bow_corpus,
...                               dictionary=dictionary)
```

Figure (3)

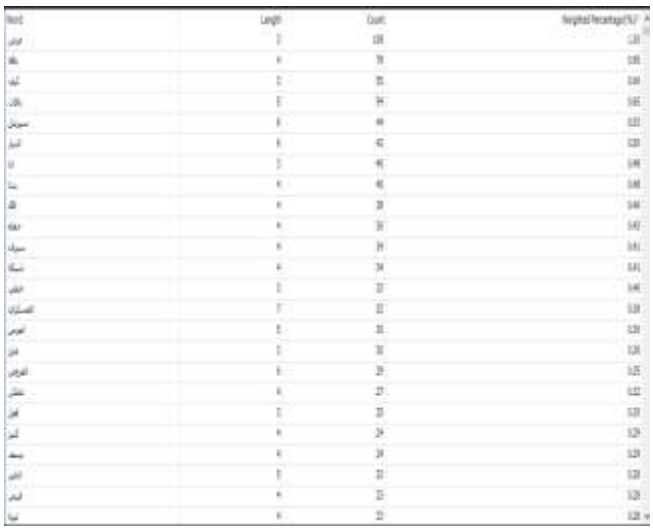


Figure(4)

1) Textual Analysis Results:

The textual analysis identified factors about Syriatel success from the topics modellings expressed in lda

we have used Word Frequency queries to list the most frequently occurring words or concepts in our file and the result was as below in Figure(5), Figure(6):



Figure(5)

Figure(6)

We have also used text seach query for (صبايا,حماة الديار) and the result was represented as Word Tree in which the results is displayed as a tree with branches representing the various contexts in which the word or phrase occurs. You may be able to find recurring themes or phrases that surround the word as below in Figure(7), Figure(8)

Figure(7)

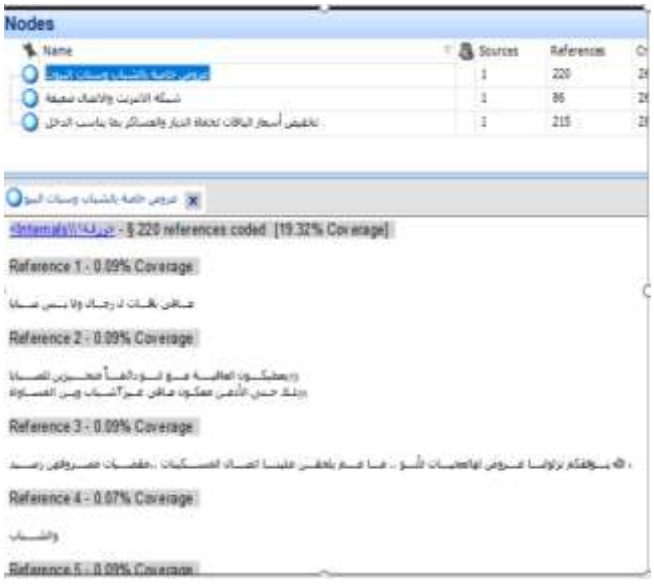
Figure(8)

1) Improving Company's Product and Suggest new Offers

We have created three nodes and the words were grouped into these nodes according to the number of times the words were repeated in the dataset.

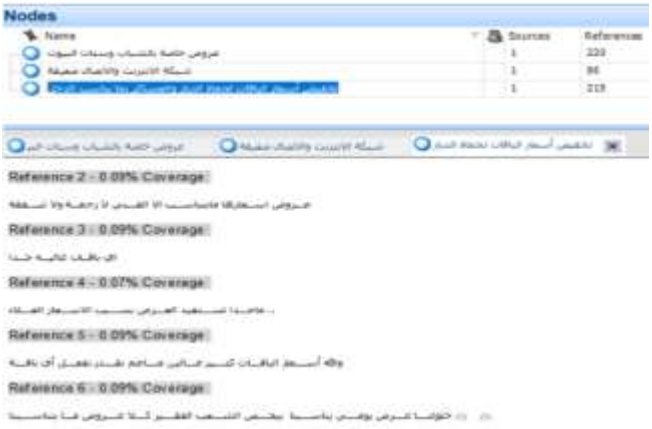
Figure(9)

For each node we can find the text reference as below in Figure(10), Figure(11),Figure(12):



Figure(10)

Figure(11)



Figure(12)

Discussion

This study is a continuation of the one in the reference (Saura, Palos-Sanchez, and Grilo, 2019) but the added value is that we will try to test a sample of feedback on startups on the network. Facebook is in Arabic and we will apply the SVM algorithm to analyze sentiments into positive, negative and neutral, then the NVIVO program will be used for textual analysis to discover the success factors of these companies.

The beneficiaries of these resulting factors are all start-up companies that are trying to establish themselves correctly and try to mitigate their failure rate and thus ensure a correct start to enter the competitive market in a strong manner.

V. Conclusion and Future Scope

In This study we have presented Detecting Syriatel's Success Indicators Using Textual Data Mining for Arabic Language which works on topics modeling technique for identifying the topics ,and then textual analysis was performed to detect the key factors Authors and Affiliations The results of this mining are demonstrated as three nodes in nvivo software for detecting indicators success by

improving the existing service and proposing new product for this company.

However, the results of the experiment showed the effectiveness of the proposed model, also Arabic language is not supported in nvivo software which lead us to code the words manually and this take long time,

Even this the proposed study gave new and good features:
1) Ability to use topic modeling and lda for Arabic language.
2) Ability to improve existing product or service by using textual analysis for Arabic language.
3) Ability to innovate a new product or service using nvivo coding and creating node in it.

In the near future, we will modify the proposed study to improve Arabic language topic modeling and textual analysis to get high accuracy in recognition and faster response time.

REFERENCES

[1] Saura, Pedro, & Grilo, "Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining", Sustainability, **11, 917, 2019.**

[2] Saura & Bennett, " A Three-Stage method for Data Text Mining: Using UGC in Business Intelligence Analysis", Sustainability **2019.**

[3] Saura, Reyes-Menendez, & Palos-Sanchez, "Are Black Friday Deals Worth It? Mining Twitter Users' Sentiment and Behavior Response", **2019**

[4] Vasile-DanielPăvăloaia, Elena-MădălinaTeodor, DoinaFotache, & MagdalenaDanilet," Opinion Mining on Social Media Data: Sentiment Analysis of User Preferences" , **2019**

[5] Al-Kabi, Al-Qudah, Dabour, & Izzat Alsmadi," Arabic / English Sentiment Analysis: An Empirical Study", 2013)

[6] Kauffmann, et al ."Managing Marketing Decision-Making with Sentiment Analysis: An Evaluation of the Main Product Features Using Text Data Mining",**2019**

[7] Saura, Reyes-Menendez, & Bennett, "How to Extract Meaningful Insights from UGC: A Knowledge-Based Method Applied to Education".

[8] Vasile-DanielPăvăloaia, Elena-MădălinaTeodor, DoinaFotache, & MagdalenaDanilet.,"Opinion Mining on Social Media Data: Sentiment Analysis of User Preferences", **2019**

[9] Mohammed N. Al-Kabi, Alsmadi, Gigieh, & Wahsheh ,"Opinion Mining and Analysis for Arabic, **2014**

[10] Mohammed Al-Kabi ,"Arabic / English Sentiment Analysis: An Empirical Study", 2013

[11] الحمد و الكردي,"Web Opinion Mining for Arabic Language", **2016**

[12] راعي و الكردي, "Based Approach for Personalizing Arabic - Semantic Web News" ,2019

AUTHORS PROFILE

Ph.D. Mohamad-Bassam Kurdy Born in Syria/Damascus July/1961, He obtained Master degree in information systems engineering from INPG - France 1986, Ph.D. in Mathematical Morphology from Mines ParisTech - France 1990,He worked at HIAST 1991-2013. He was Head of Computer Sciences at HIAST between 1997 and 2003. Country Manager for Syria of EC project EUMEDIS -Medforist. Actually Professor at SVU, ESC Dijon and ESC Rennes teaching:Advanced Data Mining, Big Data, Information Retrieval, cbIR (content based Image Retrieval) and Supervising many Master student projects (postgraduate)

Eng. Hanan Wassouf Born in Syria/Hama July/1994, She obtained bachelor degree in information systems engineering from Syrian Virtual University - Syria December, 2016,
She is student in Master in Web Science Program, Syrian Virtual University since 2018..

