A content & link Analysis of a Moroccan entrepreneurs Facebook group

***Abstract*—This article presents a study on a Moroccan entrepreneurs Facebook group using a combination of some text mining and social network analysis techniques. The aim of the study is to understand the behavior of individuals and groups within the group by analyzing the content of their posts and the connections between the group members. The study focuses on two main areas:**

- **content analysis of the Arabic dialect spoken in Morocco, known as Darija (Topic Modeling on the group posts, as well as sentiments analysis on comments to detect the value of the post)**
- **analysis of the group members network graph (Community detection; Modularity, statistical inference).**

The study highlights the importance of analyzing social media networks as they have become an integral part of modern society, with billions of people around the world using platforms such as Facebook. The use of text mining and network analysis techniques allows for a more in-depth understanding of the behavior of individuals and groups within a social media network. model accuracy: 0.96%

I. Introduction

In recent years, social networks have become an essential tool for communication, networking, and business development. analysis of this social network is a powerful tool for understanding the complex relationships and interactions within a group or community.

Content analysis is used to examine the text or other content that is shared within the network, revealing patterns and themes that may not be immediately visible in the raw data.Graph analysis,on the other hand, focuses on the structure and organization of the network itself, providing a visual representation of the relationships between individuals or groups. Together, these two approaches offer a comprehensive view of a social network.

Moroccan Arabic dialects (DA) play a crucial role in the way people communicate in Morocco, particularly on social media. With the increasing use of social media in Morocco, there is a growing need for NLP tools that can accurately understand and analyze Moroccan Arabic dialects, also known as Darija. Innovations in NLP for Moroccan Arabic dialects can help researchers, businesses, and governments to better understand and engage with Moroccan audiences on social media, and make important decisions.

Social media such as Facebook might be the most convenient source to collect DArija data, as they provide different contents that reflect the feelings of users across several topics and are written in the user native and informal Arabic dialect. however, data collected on social media should not be used in its raw form as it suffers from several issues, we can cite, for example, the problem of code-switching where users tend to borrow words or phrases from other languages(English or French), also Arabic script is written from right to left and includes diacritical marks and special characters, which can cause problems with encoding when collecting data from social media platforms.

in this article, we present a social network analysis on a Moroccan entrepreneurs group , focusing on two main areas: the analysis of the group members network and the analysis of the content of the group's interactions. This study aims to understand the role of social networks in shaping the interactions and relationships among the members and the dynamics of communication within the group.

The data collected will be analyzed using a combination of network analysis and text mining techniques in order to identify key players and patterns of communication, as well as to gain insights into the social and cultural dynamics of entrepreneurship in Morocco.

the rest of the paper is organized as follows, section 2 presents the collected data and the way it was gathered, section 3 presents the content-based analysis in which we used two applications: sentiments analysis and topic modeling, section 4 presents our application of link-based analysis.

II. Dataset

The data was gathered from a Moroccan entrepreneurs Facebook group posts, we used a python library named facebook-scraper, facebook-scraper has a function named get_posts this function needs a Facebook profile cookies.txt to gather data as well as the Id of the desired group. We can get the cookies using the extension get cookies from google chrome there is another tool for Firefox users as defined here in facebook-scraper's website [1] With get_post() we can extract as many number of posts we want the page argument as default has 10 for this number the function returns 210 posts from the group with all reactors, comments text, the commenters, comments reactors, comments replies, and post sharers, the total number of posts in our dataset is 420, also we extracted the comments of every post, the id of users who commented and reacted to every post, the reactors of comments the repliers of comments and their reactors as well so we have everythig related to the post and that can help to create the desired graph with any condition we want. Another use-full function is get_profil() this function helps us to get the reactors id, we can extract the reactors id either from the profile link or by passing the private profile name throw the get_profile function (the private profile name is the primary key of any Facebook profile generated by Facebook, not the public profile name).

	index	post_id	text	post_text	shared_text	original_text	time
0	0	5379670632102808	la أول حاجة كنبغي نقول سنة سعيدة startup stati...	la أول حاجة كنبغي نقول سنة سعيدة startup stati...		None	1666684968000
1	0	5601900936546442	لي عنده شي زهر يطلعو و يقدم المشروع ديالو قدام...	لي عنده شي زهر يطلعو و يقدم المشروع ديالو قدام...	None	None	1673280850000
2	0	5601568666579669	Incubateur Francophone Africain : Talk inspira...	Incubateur Francophone Africain : Talk inspira...		None	1673271877000
3	0	5599821916754344	المقاولة الإجتماعية واحد المفهوم قليل لي تيعرف...	المقاولة الإجتماعية واحد المفهوم قليل لي تيعرف...		None	1673210666000
4	0	5555684837834719	البريد بنك ينظم دورات تكوينية مجانية لفائدة ال...	البريد بنك ينظم دورات تكوينية مجانية لفائدة ال...		None	1671815744000
...	...	...	...	...	...	...	...
415	0	5139234859479721	عندما تود التفكير في أخذ مشروعك إلى ا #Startup...	عندما تود التفكير في أخذ مشروعك إلى ا #Startup...		None	1659268319000
416	0	5137344303002110	LESSTART-UPS AFRICAINES\nDEVELOPPENT DES SOLUT...	LESSTART-UPS AFRICAINES\nDEVELOPPENT DES SOLUT...		None	1659202138000
417	0	5136811133055427	شركة مطاعم تبيع رخص الفروع الممنوحة\n\nالفرنشايز...	شركة مطاعم تبيع رخص الفروع الممنوحة\n\nالفرنشايز...		None	1659187609000
418	0	5139053749497832	"الناس ديال جهة "مراكش آسفي\n\nPour vous aide...	"الناس ديال جهة "مراكش آسفي\n\nPour vous aide...	MARRAKECHINVEST.MA\nMarrakech Invest - CRI MS	None	1659262606000
419	0	5135134879889719			None	None	1659131298000

420 rows × 58 columns

Fig. 1. collected data

As we can see in the figure above we have many columns each one present some details about the post (user_id, comments_full, reactors, and more interesting information.......)

III. Content-based analysis

A. Methodology

To conduct the content-based analysis in this study, we carried out several NLP architectures and models, in our case DarijaBert gives good results in topic modeling and sentiments analysis of our data. DarijaBert [2] model, is the first Open Source BERT model for the Moroccan Arabic dialect called "Darija". It is based on the same architecture as BERT-base, but without the Next Sentence Prediction (NSP) objective. This model was trained on a total of 3 Million sequences of the Darija dialect representing 691MB of text or a total of 100M tokens. and it was realized by the AIOX laboratory, it was fine-tuned on 3 downstream tasks, namely Dialect Identification (DI), Sentiment Analysis (SA), and Topic Modeling (TM). In our study, we used DarijaBert to make predictions in our collected dataset for the two applications; sentiment analysis on comments and topic modeling on posts.

1) Sentiment analysis: The dataset used for this project is called the MSDA Datasets Wich is an open access NLP dataset for Arabic (Moroccan, Algerian and Tunisian) dialects : Sentiment Analysis for social media posts in Arabic dialect : Labeled data set of 50K posts. The paper [3] describes the collection, the labeling method for this dataset as well as the database use with typical NLP algorithms. In order to use

	Unnamed: 0	text	label	sentiment
0	0	what happens	neu	1
1	1	😞 😞	neg	0
2	2	ان شاء الله	neu	1
3	3	بالتوفيق ان شاء الله	pos	2
4	4	الحمدلله على كل حال	pos	2
...	...	...	...	...
52205	52205	🥺 🥺 هاي باهي وليت تفهم في اللبسة متاع الرجال	neu	1
52206	52206	ربي يصبر امها	neu	1
52207	52207	اللهم امين يارب العالمين ربي يصبر أهلهم	pos	2
52208	52208	...اللهم امين يارب العالمين مخيبها عواشر لا حول و	pos	2
52209	52209	... wangen لا حول ولا قوة الا بالله جريمة بشعة في	neg	0

52210 rows × 4 columns

Fig. 2. Sentiment analysis data

machine learning algorithms, text data needs to be converted into numerical representations. For this project, we used the BERT architecture which requires the following preprocessing steps:

- Adding special tokens to separate sentences and do classification
- Padding sequences to a constant length (In our project, choosing a maximum length of 512 tokens for the input sequences was a reasonable decision given that most reviews in our dataset had fewer than 600 tokens)
- Creating an array of 0s (pad token) and 1s (real token) called an attention mask.

Additionally, we cleaned the data to remove noise and transform it into a suitable format for machine learning. This involved:

- Removing specific keywords and symbols such as hashtags #, URLs, and http from each tweet.
- Removing Arabic stop words, including a list of 250 words from the NLTK library, over 700 words from the GitHub page, and a few additional stop-words.
- Removing emojis, except for sentiment analysis since they are important in expressing sentiments.
- Removing punctuation characters, including Arabic punctuation characters like the reverse question mark, as well as English punctuations from the string library.
- Translating non-Arabic words for dialect detection, such as "Thank you" and "what happens."
- Removing all Arabic diacritics and replacing some letters that were written in various ways with a single format.
- Calculating the length of each tweet and removing meaningless short tweets.

Finally, to ensure efficient model training, we use PyTorch to create a data loader that splits our prepared dataset into smaller batches. This approach helps to prevent memory issues and enables faster training. In addition, PyTorch data loader provides useful features like shuffling and parallel data loading, which can further enhance model efficiency and accuracy.

We pre-trained a DarijaBERT model using the cleaned and tokenized text data. The pre-training process involved masking 15% of the tokens and predicting them based on the context provided by the other tokens. The model architecture was based on the BERT-base architecture and consisted of 12 encoder blocks, 768 hidden units, and 12 attention heads. The pre-training was performed using a GPU and data loader with a batch size of 16. After pre-training, we fine-tuned the pre-trained DarijaBERT model on a smaller labeled dataset of 50,360 reviews, which was split into 45,324 reviews for training, 2,518 reviews for validation, and 2,518 reviews for testing. We constructed our sentiment classifier using the fundamental BertModel. The model was trained to classify the sentiment of each review into one of three categories: negative, neutral, and positive.

During training, we monitored the train and validation accuracy curves to ensure that the model did not overfit to the training data. The model achieved a training accuracy of 96.31% and a validation accuracy of 83% after 5 epochs of training. The accuracy curves indicate that the model did not overfit to the training data, as the validation accuracy did not decrease while the training accuracy increased.

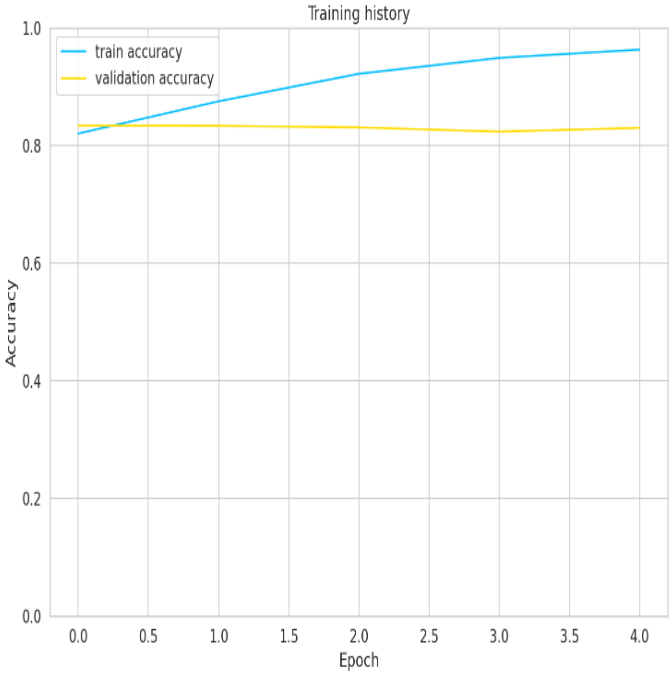

Fig. 3. Train and validation accuracy curves

To evaluate the performance of the model, we tested it on the held-out test set and achieved an accuracy of 85.06% for sentiment classification. This result indicates that our model is able to accurately predict the sentiment of reviews in Darija language.

We also generated a sentiment analysis classification report to provide further insight into the model's performance. The report shows precision, recall, and F1 score for each sentiment category. Our model achieved a precision of 82%, recall of 87%, and F1 score of 84% for the negative class, a precision of 87%, recall of 88%, and F1 score of 88% for the neutral class, and a precision of 73%, recall of 57%, and F1 score of 64% for the positive class. The high F1 scores across all classes suggest that our model can accurately classify sentiment in Darija text.

2) Sentiment analysis results:

	precision	recall	f1-score	support
negative	0.82	0.87	0.84	788
neutral	0.87	0.88	0.88	1391
positive	0.73	0.57	0.64	339
accuracy			0.84	2518
macro avg	0.80	0.77	0.79	2518
weighted avg	0.83	0.84	0.83	2518

Fig. 4. Sentiment analysis classification report

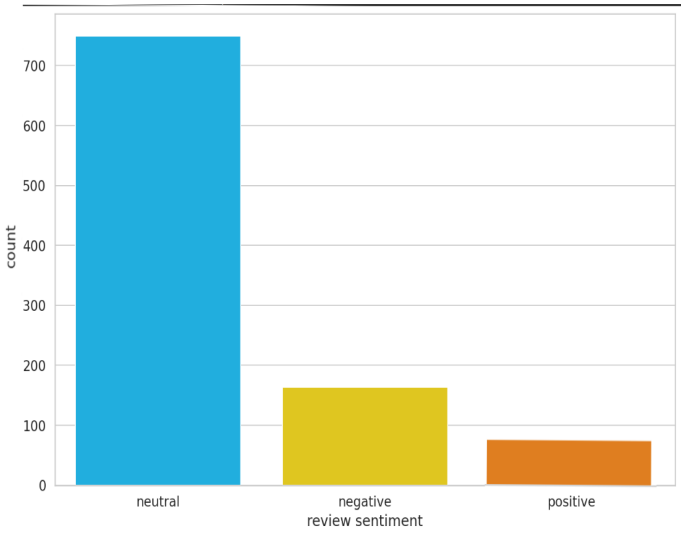

Fig. 5. Sentiment Analysis Results of the predictions made on our collected data

3) Topic modeling: The "Arabic Classification DataSet " [4] proposed by Mohamed Biniz in 2018, was used as a dataset for topic modeling. This dataset contains 111,728 text documents collected from three Arabic online newspapers (Assabah, Hespress, and Akhbarona) through a semi-automatic web crawling process. The texts are in modern Arabic and contain alphabetic, numeric, and symbolic words. The dataset is divided into five categories: sport, politics, culture, economy, and diverse, with varying numbers of documents and words per category. This publicly available dataset, version number 2, can be accessed on Mendeley Data [5] and it is useful for research in the field of Arabic natural language processing and text classification.

For model training, we fine-tuned the pre-trained DarijaBERT model on a smaller labeled dataset of 84,293 reviews, which was split into 67,434 reviews for training, 8,429 reviews for validation, and 8,430 reviews for testing. We constructed our fundamental BertModel and trained the model to classify the topic of comments made by group members in a discussions. The model was trained to identify the topic of each comment from a set of predefined topic classes, including sport, politics, culture, economy, and diverse. This allowed us to automatically categorize the comments based on their underlying topics, providing valuable insights into the discussion patterns and topics of interest among the group members.

4) Topic modeling results:

	precision	recall	f1-score	support
Colture	0.97	0.95	0.96	692
Diverse	0.96	0.98	0.97	843
Economy	0.92	0.94	0.93	721
Politic	0.94	0.92	0.93	1004
Sport	0.99	0.99	0.99	1047
accuracy			0.96	4307
macro avg	0.96	0.96	0.96	4307
weighted avg	0.96	0.96	0.96	4307

Fig. 6. Topic modeling classification report

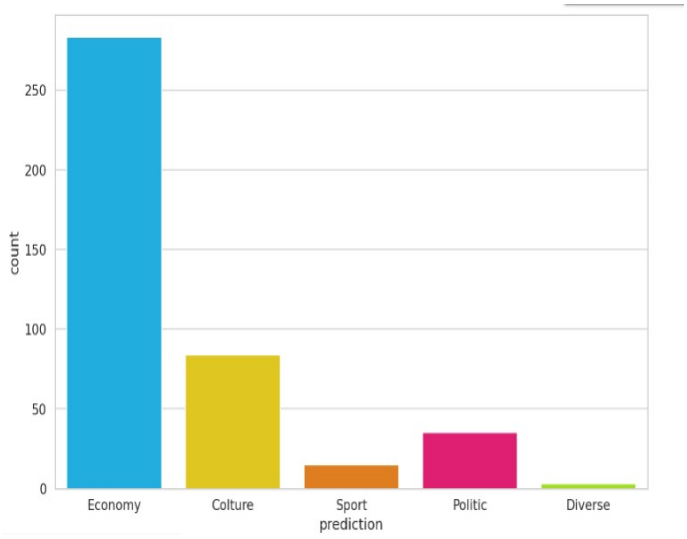

Fig. 7. Topic modeling results of the predictions made on our collected data

IV. Link-based analysis

Link-based analysis is a field of study that uses network theory to examine the structure and behavior of social networks. These networks can be represented as a graph, with nodes representing individuals or organizations and edges representing relationships between them. Social network analysis allows researchers to understand the patterns and connections within a network, as well as the influence and centrality of individual actors within the network.

A. Methodology

After gathering the- needed data especially the IDs of all the participants in any post including (commenters, reactors, sharers, repliers, comments reactors, replies reactors) and the post owner for sure, we draw the graph using Networkx and Gephi. Networkx is a Python library for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. It provides classes for graph objects, generators to create standard graphs, IO routines for reading in existing data files, algorithms to analyze the resulting networks, and some drawing tools. So it is a suitable tool to use for drawing simple graphs with node labels as well as links weight (the number of interactions between two or more users). here is a simple networkx weighted & labeled graph in the figure bellow.

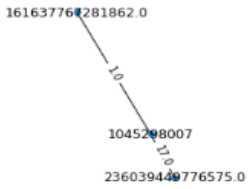

Fig. 8. Gephi interface

Then we save our graphs to use in Gephi, The Open Graph Viz Platform, Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free. Gephi offers many usefull techniques that can be used in our case such as modularity, community detection, statistical inference and other visualization tools

.

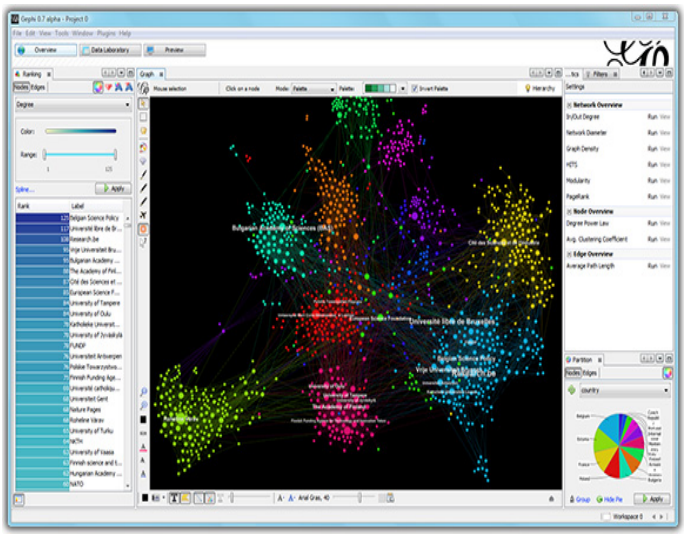

Fig. 9. Gephi interface

B. Gephi Graphs

After all data processing and Ids collection here are the graphs of post reactors, commenters, sharers, and all interactions as well. see the figures bellow

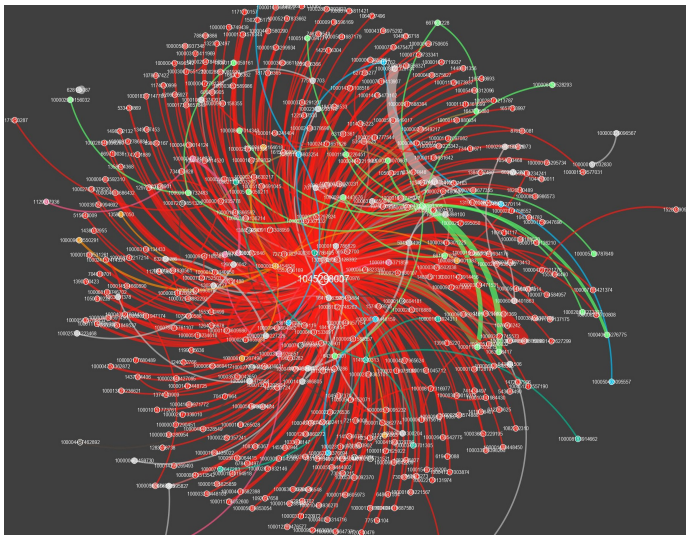

Fig. 10. Graph of commenters

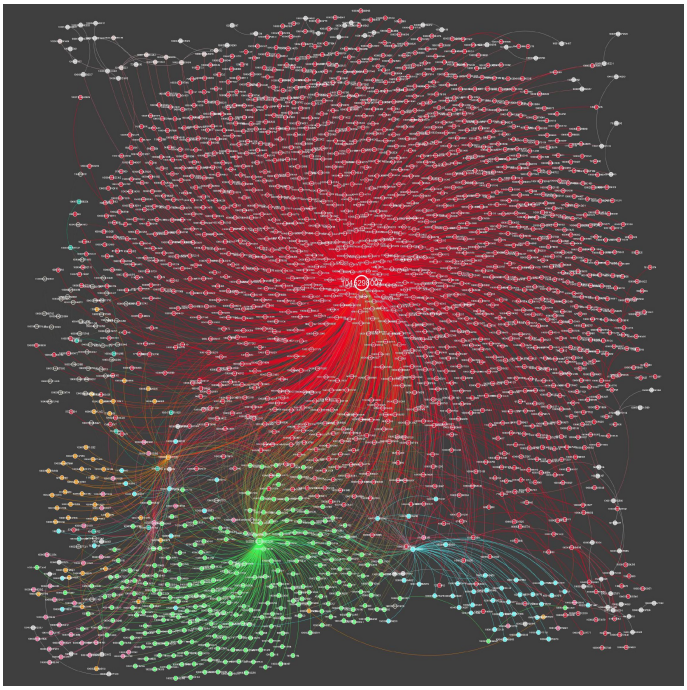

Fig. 11. Graph of reactors

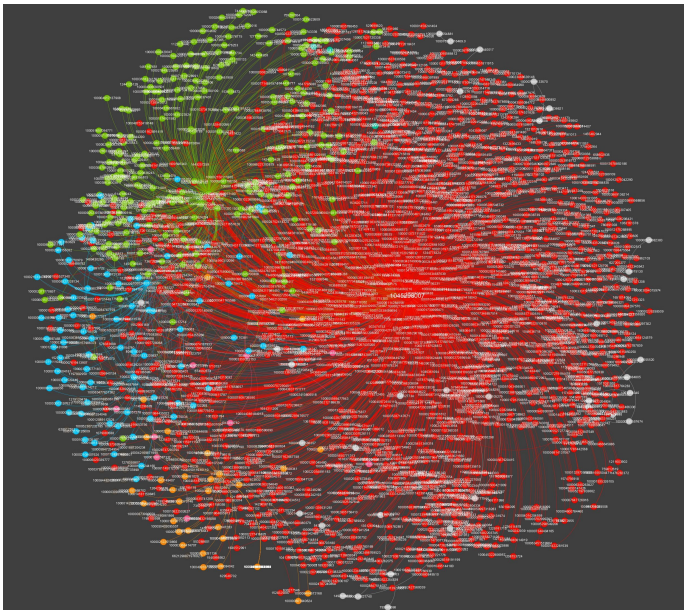

Fig. 12. All interactions graph

Fig. 13. Shares graph

Since our Topic modeling & sentiments analysis models gave good results we can use the predicted labels to draw some specific graphs such as for example a graph contains only Economy posts interactions (reactions or comments) or Sport or Politic posts. Or why not use a combination on negative comments or positive ones on Politic posts for example well we have some good news we achieved this goal we created a function that generate this graphs with all conditions we want and here are some examples in the figures bellow

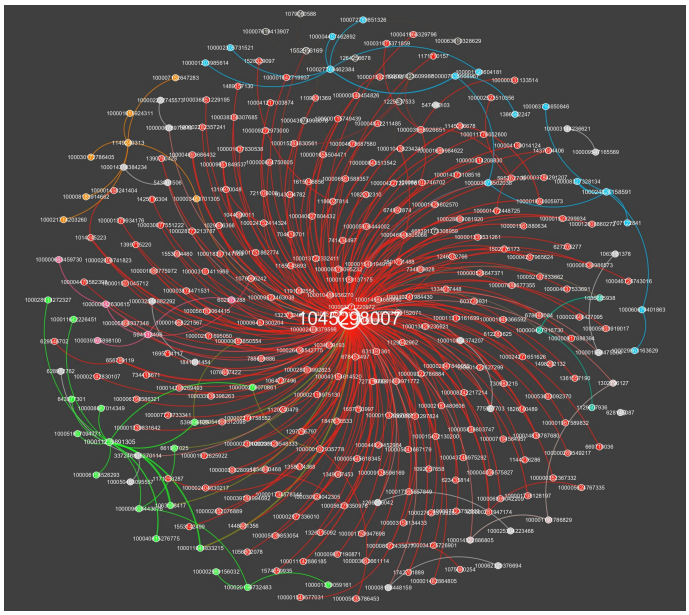

Fig. 14. The graph of commenters on predicted Economy posts

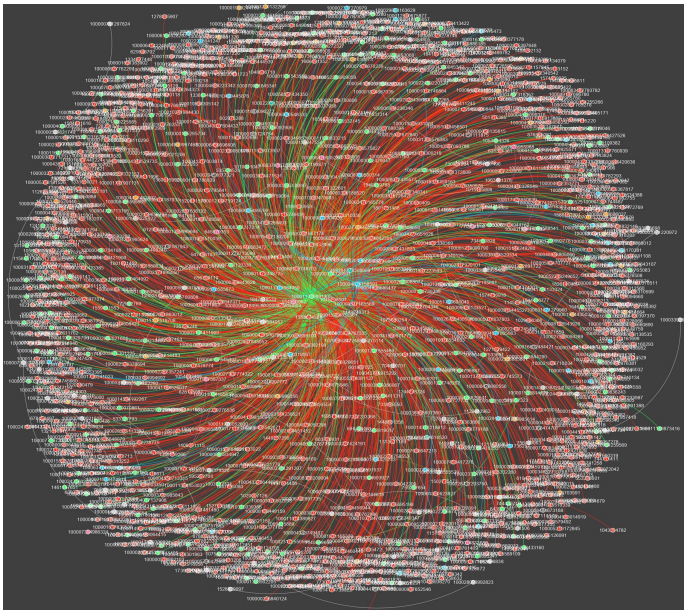

Fig. 15. The graph of reactors on predicted Economy posts

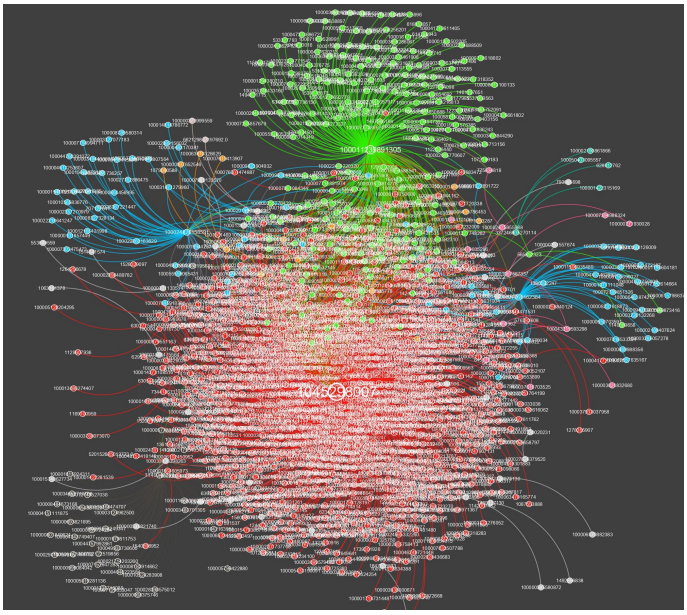

Fig. 16. The graph of all interactions on predicted Economy posts

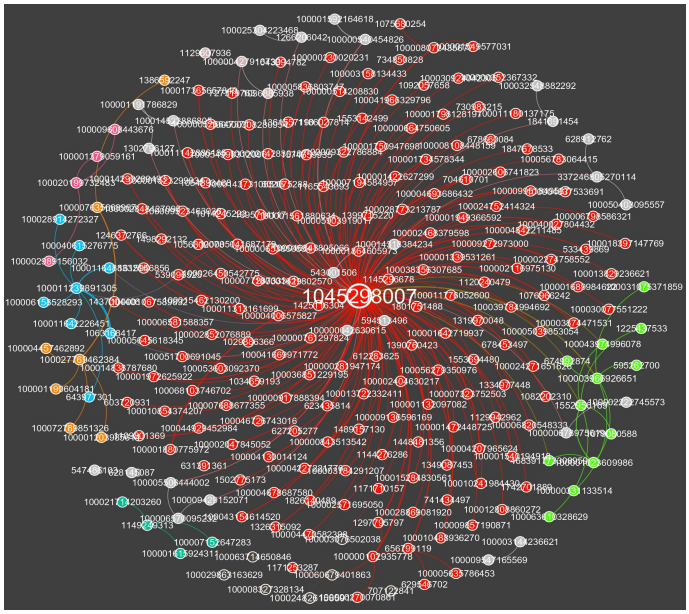

Fig. 17. neutral comments graph on Economy posts

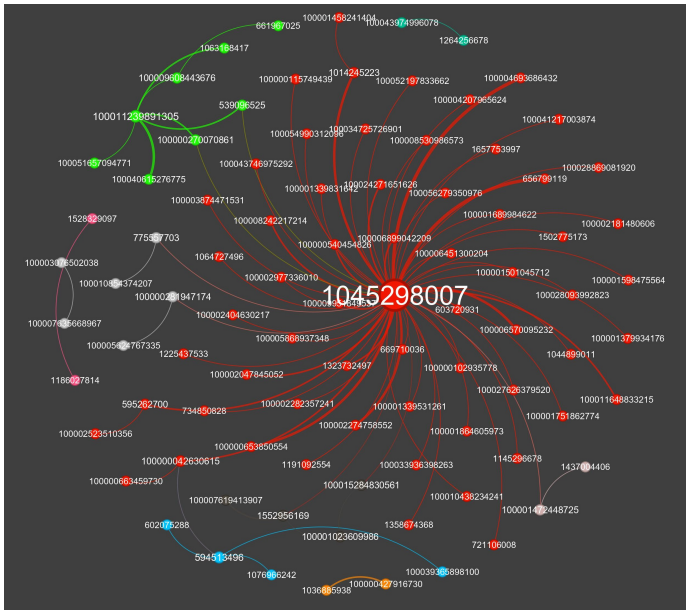

Fig. 18. negative comments graph on Economy posts

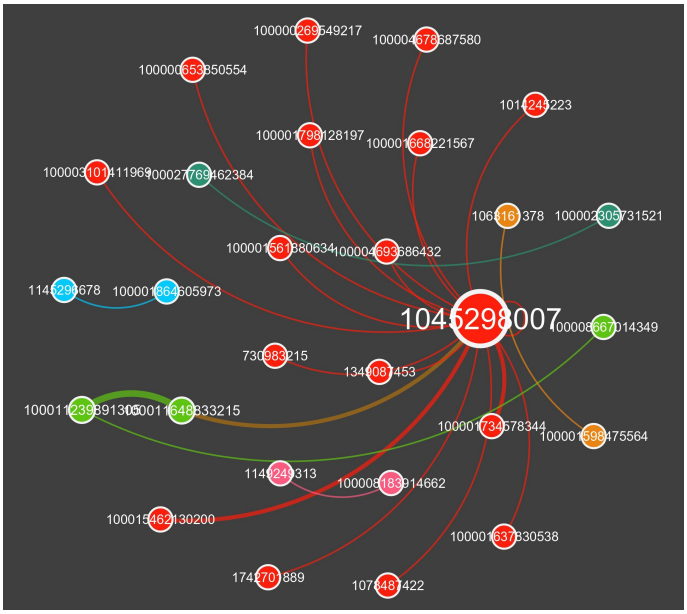

Fig. 19. positive comments graph on Economy posts

We implemented our code in google colaboratory which is a cloud-based platform that allows users to run Jupyter notebooks on Google's cloud servers. It provides users with free access to virtual machines equipped with GPUs and TPUs, we recommend to follow the same order in which the codes have been implemented to avoid any error.

V. Conclusion and Future Work

In conclusion, our analysis of the Facebook entrepreneurs Moroccan group using social network analysis (SNA) and topic modeling provided a valuable insights into the structure and content of the group's network. However, we recognize the challenges of analyzing the Moroccan dialect written in (Freanch and numbers) which needs more attention in that case we need to collect a specific data for that purpose and labeling it which may take a long time , or the presence of Moroccan dialect variations. Certain messages were written in Arabic, Darija, and FrenchEnglish we tried to translate texts in this case to arabic , so our analysis focused on the standard variety of Arabic. While this approach allowed us to better understand the structure and content of the group's network, it may not have captured the full range of linguistic and cultural variation within the group.

To address this limitation, future work could explore the use of language models specifically designed for dialectical variations and languages. Additionally, comparing the results of analyzing each language and dialect separately could provide a more nuanced understanding of the group's dynamics. Also some additional data sources such as text data (e.g., other posts, comments) or geographical data (e.g., location of users), which can provide additional insights into the social network and its dynamics. Implement real-time analysis to focuse on a dynamic social network, that can capture changes and updates in the network as they happen. This can be achieved through the use of streaming data processing tools such as Apache Kafka and Apache Flink. Finally, incorporating qualitative analysis, such as in-depth interviews with group members, could provide additional context and insights into the experiences and perspectives of the entrepreneurs in the group.

REFERENCES

[1] https://pypi.org/project/facebook-scraper/
[2] https://github.com/AIOXLABS/DBert
[3] An open access NLP dataset for Arabic dialects : data collection, labeling, and model construction, Elmehdi Boujou, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi and Ismail Berrada MENACIS 2020 conference, In press.
[4] https://data.mendeley.com/datasets/v524p5dhpj/2
[5] DOI 10.17632/v524p5dhpj.2