ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 155-159 (2020)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Inferring Topics within Social Networking Big Data, Towards an Alternative for Socio-Political Measurement

Khalid Ait Hadi*,[1], Rafik Lasri[1], Abdellatif El Abderrahmani[2]

[1]*Department of Computer Sciences, Polydisciplinary Faculty of Larache, Abdelmalek Essaadi University, 93000, Morocco*

[2]*Department of Computer Sciences, Dhar El Mahraz Sciences Faculty, Sidi Mohamed Ben Abdellah University, 30050, Morocco*

ARTICLE INFO

Article history:
Received: 03 September, 2020
Accepted: 28 October, 2020
Online: 08 November, 2020

Keywords:
Language processing
Data Mining
Topics detection

ABSTRACT

This research sought to measure some socio-political indicators using millions of opinionated messages from social network sourced big data. Thus, and using an enhanced mixed method for sentiment analysis and a fusion model algorithm to infer topics from short text, this study attempted to demonstrate the value of computational approaches in measuring some phenomena in the real social world and quantifying public opinion fluctuations in response to certain socio-political issues. The validity of the experimental results was examined by comparing them with data obtained from representative surveys, thus providing a better understanding of the relationships between online and offline opinion dynamics. This contribution is intended to be multidisciplinary, both useful for policymakers and opinion analysts to explore public trends and to inquire into socio-political issues.

1. Introduction

Nowadays, in the communication fields, market research and public opinion analysis, as in many other disciplinary domains, social media are mobilized as new societal trends observatories [1, 2]. The evolution of these networking platforms in the first decade of the 21st century has enabled Internet users without specific technical skills to easily publish and share content on their concerns [3]. The participation architecture on which these services are based has facilitated the information co-production, and has offered spaces for socio-political engagement [4].

In this context, millions of people interact daily on social networks, producing several hundreds of millions of messages of all kinds and of any content. These messages represent valuable clues concerning practices, representations and Internet users' opinions. As such, they form a particularly interesting material to investigate when looking at citizen public behavior studies. Furthermore, due to the development of technical and computational tools that have made it possible to collect, archive and analyze huge data volumes, namely big data, these numerical phenomena can be objectified and quantified, on a large scale, for the investigation's needs. Social media analytics open up a promising way for the quantitative study of certain substantial subjects of declarative surveys, such as quantifying a social object, capturing behavioral representations and uncovering political intentions.

In this work, we seek to exploit data extracted from Facebook and use computational methods to measure some socio-political indicators. This article does not claim to provide affirmative answers about the effective correlations between the phenomena observed on social media and those perceived in the real world. On the contrary, it aims to help support the observation indicating that social networking data constitute a particularly interesting material to investigate when trying to study citizens' behavior, public opinion's fluctuations and social trends, particularly where declarative surveys and others conventional approaches turn out to be often costly, time-consuming and labor-intensive [5].

This paper is an extension of work originally presented in 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS) [6]. An initial work, which was interested in measuring of notoriety capital on Facebook of a political figure. Here, we will not be limited to measuring political popularity, but we will improve this research to be able to evaluate other socio-political indicators. Likewise, we will not settle for the use and study of messages formulated only in one language, as was the French case in [6], but we will extend the prospecting to include also social networking data expressed in classical Arabic and dialect as well.

*Corresponding Author: Khalid Ait Hadi, khalid.aithadi@gmail.com

Another improvement of the present work is linked this time to the observation relating to the fact that topic models, like Latent Dirichlet Allocation (LDA) and its derivatives, are not efficient in the analysis of short texts due to data sparsity [7]. Alas, it is clear that the overwhelming majority of data extracted from social networks consists of short texts. Here, we will overcome this inconvenience by opting for an improved Biterm Topic Model (BTM) model based on the heat matrix [8].

The rest part of the paper is structured as follows. Section 2 presents an overview of related work. Section 3 is dedicated to the data collection process description and the suggested approach implementation. An experimental evaluation of the method is also presented in the same section. Finally, discussions, conclusions and suggestions for future research openings are made in the last section.

2. Related Work

Studies seeking to examine predictive capacity of social media data and to establish existence of possible correlations between the facts perceived on social media and those observed in the effective world have been subject to a rich and growing literature since the beginning of the 21st century [2].

Many recent studies have used social media data as a new sensor of societal trends and a new predictor of economic or political phenomena, ranging from stock market volatility to box office performance or, in a certain perspective that interests us more directly, public opinion fluctuations and social dynamics [9, 10]. Research in this area has been driven by a positivist approach that is certainly strengthened owing to the unprecedented possibilities offered by these new social platforms, in the sense that they allow unheard-of access to the data shared daily by people as well as to their networks, regardless of any temporal or geographic consideration.

In this context, a review article published in 2014 [11] has listed more than a hundred studies devoted solely to political messages published on the social network Twitter. The penchant of researchers for Twitter can be explained by easier access to data, where, unlike Facebook, the vast majority of accounts are “accessible” and by the brevity of posts on Twitter, limited to 280 characters.

Always on Twitter, a meta-analysis developed by Gayo-Avello [12] examined the predictive usefulness of data published on this social network, leading to the conclusion that such data have relatively predictive power and provide some information referring to electoral consultations results.

It should be noted on the socio-political aspect that the most commonly studied subject is that concerning electoral outcomes prediction using social media data [13]. In the opposite, studies relating to the political popularity evaluation, for example, or to the measurement of social indicators are not numerous, and here too Twitter has often been called upon to make forecasts. In addition, and almost globally, these studies have not examined the validity of the predictions by making comparisons to conventionally more admissible results types, such as those from censuses and surveys [14].

It should also be emphasized, according to the existing scientific literature [1], that most results in the field of forecasting with social media data claim to have produced predictive results to some extent, registering a positive statistical correlation between offline and online data. However, several scientists have criticized these results, which draw up a fundamental limitation to these works that is linked to the lack of sociological representativeness of the populations recorded on social networks [15]. Though, dismissing social media data on this ground would fail to capture the opinion forming dynamics. Some studies have even argued that representativeness search is no longer consubstantial with the need to sample populations in the big data era. The arguments put forward in this wake point to fact that debates led by certain politically active groups prevail over those who develop in society at large [16], and that individuals active on social networks are thought leaders, more politicized than average, influential in their entourage and acquaintances and whose opinions count more than those of "ordinary" individuals. To poll these opinion leaders would indirectly mean polling their entourage and, ultimately, the entire population. In addition, discussions on certain social platforms would above all reflect the concerns and themes put on the agenda by the mainstream media [17]. In many ways, the social network appears to be an echo chamber for the media field, and the associated media agenda would indirectly influence citizens' concerns [18].

Regarding the diversity of methods used in measuring public opinion, whether sentiment-based or social network-based approaches, it should be underlined that there is no unanimity about the most efficient methods in terms of prediction, and several researchers have reported some performance with different approaches and implementations.

Thus, and particularly in the category of sentiment-based studies, where opinions tone is used as a behavior’s indicator, two lines of research stand out in this context, one is based on pre-established lexicas [19], while the other relies on new sentiment models specifically for political messages [20].

Although concerning the dictionary-based sentiment classifiers, some studies have criticized the reliability of these approaches for predictions [21, 22], advancing in this regard some deficiencies such as the incorrect classification of the word in the lexicon, the lack of words disambiguation, and the neglect of contextual inference. Imperfections that are more accentuated in the socio-political context, with the emergence of difficulties especially linked to the lack of subtleties of socio-political language [23].

Through this literature review, it is seen that the majority of research has focused on the electoral and public opinion forecasts, without identifying and analyzing the reasons behind these opinions development, problematic that the present work is concerned with. In addition, there are two more obstacles that we will overcome in this paper, firstly by using a data source other than Twitter, Facebook in this case, and secondly by comparing our results with those of a declarative survey [24].

3. Methodology and Experimental Results

As previously announced, this research aims precisely to support the predictive interest of data extracted from Facebook. This involves, on the one hand, developing algorithm to analyze a very large amount of messages with socio-political connotation in

order to identify their polarity and target and, on the other hand, to compare these information to opinion poll statistics so as to better grasp similarity degrees between online and offline opinion dynamics. For this, we will consider the same period covered by the public opinion survey 'The Arab Barometer' [24] conducted in Morocco from May 7 – June 11, 2016.

Data collection has been made using '*Rfacebook*' (a package that provides a series of functions that allow R users to access Facebook's API to collect public status updates that mention specific tags). Therefore, and in order to cover a very large sample of socio-politically concerned Moroccan Facebook users, the choice has been made so as to extract both messages and comments written in French or Arabic, even dialect. On this register, we admit that debate around socio-politically charged issues is also, if not largely, done in the mother tongue and that disregarded educational level, people remain always sensitive to socio-political themes and use informal language to take part in the debate.

Therefore, and to test our approach, we have looked at two socio-political issues: (I1) 'opinions about politics' and (I2) 'perceptions of freedoms'. For the first theme, we have collected data (posts and comments) corresponding to the following queries: 'politique', 'gouvernement', 'politique gouvernementale', 'سياسة', 'حكومة', and 'السياسة الحكومية'. The second theme was examined using three tags: 'libertés', 'حرية', and 'الحريات'.

3.1. Sentiment analysis

It would be necessary before approaching the opinion mining phase, and in the ultimate objective to recover pure personal opinion, to operate some text preprocessing and cleaning steps: stemming, removal of URLs, punctuations, stopwords, screen names, special caracters and duplicate comments.

For sentiment analysis, two approaches will be used, one reserved for data written in French and the other for those formulated in Arabic or dialect language. For the first category, the choice focused on the use of Carousel greedy algorithm with Cat Swarm Optimization based Functional Link Artificial Neural Networks (CSO-FLANN), a technique for sentiment mining, enhanced in term of accuracy and computational effort [25]. For the other category of data, we used 'SentiArabic', a sentiment lexicon package for standard Arabic [26] coupled with the 'MADAR' corpus, a collection of parallel sentences covering the dialects of 25 cities from the Arab World [27], and for our case we choose the package relating to the Moroccan city 'Fez'. Using these tools, we found 1.15 million positive, 1.48 million negative, and 3.37 million neutral posts and comments for the first issue (I1) 'opinions about politics', and 2.58 million positive, 6.02 million negative, and 4.4 million neutral posts and comments for the second issue (I2) 'perceptions of freedoms'.

3.2. Inferring Topics

Our approach focuses on detecting the main reasons behind a positive or negative impression expressed by Facebook users on a given theme.

Unlike the Latent Dirichlet Allocation (LDA)-based models, which are relatively inefficient in short text processing, where they generate a high dimensional and sparse data problem [7], we opt here for a model called HMBTM (Heat Matrix based Biterm Topic Model), based on Biterm Topic Model (BTM) and improved by introducing the heat matrix. This is also merged with VSM (Vector Space Model) for counting similarities, thus inducing accuracy improvement and dimensionality reduction [8]. Applying an implementation of VSM- HMBTM to reveal the topics from the set of negative posts and comments for the first issue (I1), and from the set of positive posts and comments for the second issue (I2), and using thereafter Maximum Likelihood Estimation (MLE) [28], 906 topics were selected as the optimum number of topics for the first issue (I1), and 798 for the second issue (I2). We admitted here that people express their agreement with positive sentiments as they express their critical positions with negative sentiments.

By examining the most used terms, and despite overlaps and similarities between several themes, we try to assign a unifying label in order to bring together the most relevant subjects. For example, we have assigned for the second issue (I2) 'perceptions of freedoms', the label "freedom of association" to topics containing: joint associations, civil associations, civil organizations… and the label "political participation" to a set including the terms: political parties, free election, partisan activities…

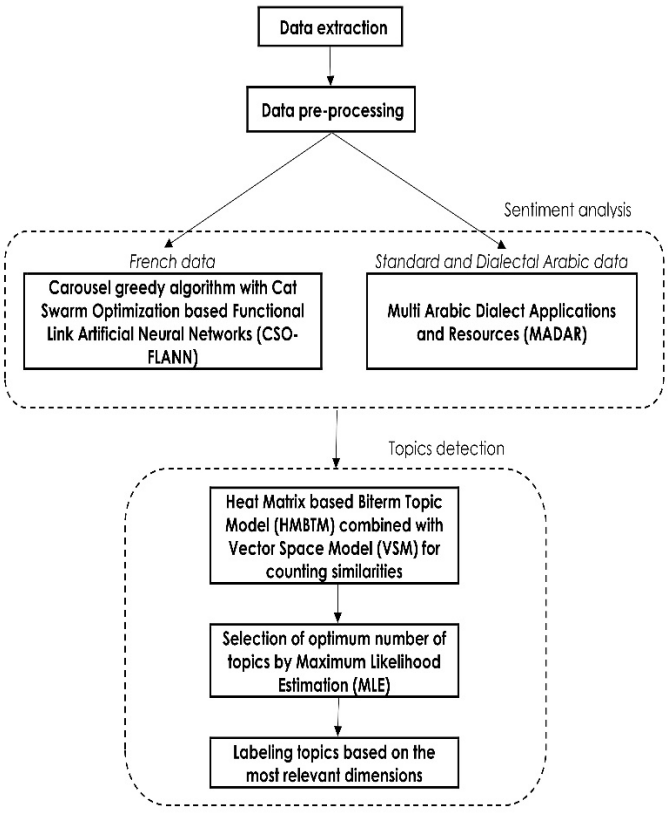

Figure 1: The construction process of our approach

We adopt the same process for the first issue (I1) 'opinions about politics. By examining the main related words such as those in Table 1, we agreed to assign a label based on the most relevant dimensions. For example, we reserved the label "complexity of politics" to topics containing: blurry political landscape, multiplicity of political concepts, number of political parties...

Table 1: A Sample of First Issue's Topics

Complexity of politics	Government not concerned with citizens
multiplicité innombrable ضبابية مبهم	opportunisme détachement لامبالاة ابتعاد
Freedom to criticize the government	Obligation to support the government
reproches critiquable انتقاد خوف	soutien supporter تأييد مساندة

The construction process of our approach is illustrated in Figure 1.

Subsequently, we explore the distribution of labels to get an idea of the topics' distribution and weight for Facebook users. Therefore, Figure 2 represents the rates corresponding to the four relevant subjects that emerge from the analysis, compared to those established by of a declarative survey [24]. Obviously, the values have been scaled given the difference in measurements and scales adopted in each situation.

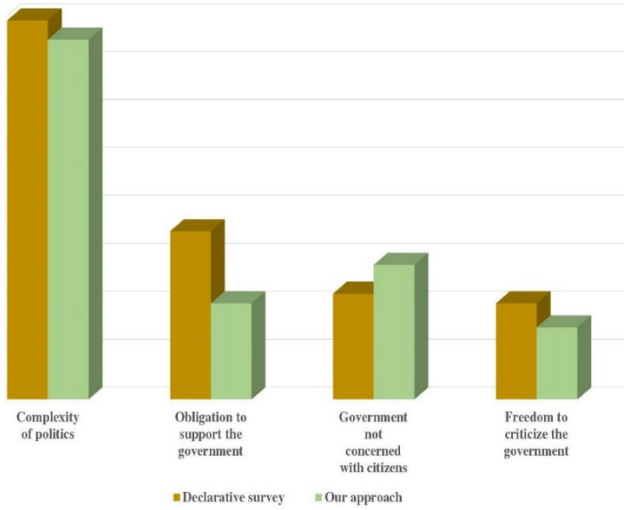

Figure 2: Distribution of First Issue's Topics

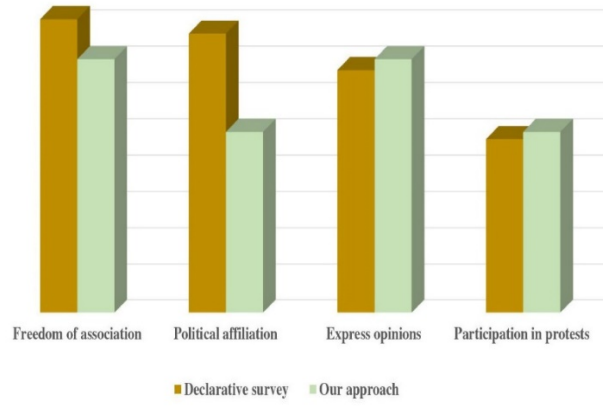

Figure 3: Distribution of Second Issue's Topics

For the second issue (I2) 'perceptions of freedoms', four topics were the most salient: freedom of association, political affiliation, express opinions and participation in protests. Figure 3 compares the weight of these topics to that resulting from the declarative survey [24].

4. Discussion and Conclusions

We have presented a research approach whose objective is to compare the results obtained via two techniques for measuring opinions: the declarative survey, on the one hand, and the analysis of messages published on Facebook, on the other hand. The work sought to situate the data produced by computational methods in relation to the results obtained by surveys. Although the current results do not release certainties, but it appears that, our approach opens up promising horizons, in the sense that it provides very indicative results about the configuration of future declarative surveys results, particularly when the latter are costly in time and even materially. The results presented through this research demonstrate the potential of data extracted from social networks to provide basic and essential information about public opinion dynamics.

However, the results as they stand confirm the observation so much raised by several researchers, indicating that the analysis of social networks would not replace, at least now, public opinion studies based on conventional polls [14]. Nevertheless, it offers indicator data and informative tools to refine our assimilation of public opinion and socio-political behavior. It also opens up perspectives for better capturing indicators relating to engagement and trends measurement, and assessing by the way, the feasibility of many promises in terms of renewing mobilization practices in socio-political field.

One of the possible openings of the present work is to diversify the platforms of data extraction, something that could play, in our point of view, a crucial role to improve the analyses precision. Broadening the base of opinionated data would undoubtedly amount to increasing the representativeness index, so much criticized in this area. One of the concerns that are shaping up recently is the future unavailability of social networks data when the confidentiality of private life is increasingly questioned. End-to-end encrypted messaging applications are attracting new users every day. Social platforms further restrict access to user data to ensure confidentiality of privacy or simply for purely commercial reasons.

One of the possible directions of research is also the one linked to the following question: does not reasoning in terms of positive or negative tones lead to neglecting an essential dimension of the analysis of socio-political messages on social networks, namely the strong majority of messages considered neutral (something that can be seen from the results above)? This leads us to question the silence or neutrality of users and to apprehend its socio-political significance, which would constitute a major issue for future promising research.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. F. Schober, J. Pasek, L. Guggenheim, C. Lampe, F. G. Conrad, "Research Synthesis: Social Media Analyses for Social Measurement" Public Opinion Quarterly, **80**(1), 180–211, 2016. https://doi.org/10.1093/poq/nfv048

[2] H. Schoen, D. Gayo-Avello, P. Takis Metaxas, E. Mustafaraj, M. Strohmaier, P. Gloor, "The Power of Prediction with Social Media" Internet Research, **23**(5), 528–543, 2013. https://doi.org/10.1108/IntR-06-2013-0115
[3] D. Cardon, V. Jeanne-Perrier, F. Le Cam, N. Pélissier, "Présentation" Réseaux, **24**(137), 9–25, 2006.
[4] N. A. Jackson, D. G. Lilleker, "Building an Architecture of Participation? Political Parties and Web 2.0 in Britain" Journal of Information Technology & Politics, **6**(3-4), 232–250, 2009.
[5] W. Wang, D. Rothschild, S. Goel, A. Gelman, "Forecasting Elections with Non-representative Polls" International Journal of Forecasting, **31**(3), 980–991, 2015. https://doi.org/10.1016/j.ijforecast.2014.06.001
[6] K. Ait Hadi, R. Lasri, A. El Abderrahmani, "Social Media Reputation and Political Popularity: Study of a Moroccan Case" in 2019 IEEE International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco, 2019. https://doi.org/10.1109/ISACS48493.2019.9068920
[7] L. Jiang, H. Lu, M. Xu, C. Wang, "Biterm Pseudo Document Topic Model for Short Text" in 2016 IEEE International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, USA, 2016. https://doi.org/10.1109/ICTAI.2016.0134
[8] Q. Liqing, J. Wei, L. Haiyan, F. Xin, "Microblog Hot Topics Detection Based on VSM and HMBTM Model Fusion" IEEE Access, **7**, 120273–120281, 2019. https://doi.org/10.1109/ACCESS.2019.2932458
[9] E. Kalampokis, E. Tambouris, K. Tarabanis, "Understanding the Predictive Power of Social Media" Internet Research, **23**(5), 544–559, 2013.
[10] D. Rousidis, P. Koukaras, C. Tjortjis, "Social Media Prediction: a Literature Review" Multimed Tools Appl, **79**, 6279–6311, 2020. https://doi.org/10.1007/s11042-019-08291-9
[11] A. Jungherr, "Twitter in Politics: A Comprehensive Literature Review" Social Science Research Network, 2014. https://doi.org/10.2139/ssrn.2402443
[12] D. Gayo-Avello, "A Meta-analysis of State-of-the-art Electoral Prediction from Twitter Data" Social Science Computer Review, **31**(6), 649–679, 2013. https://doi.org/10.1177/0894439313493979
[13] S. C. McGregor, R. R. Mourão, L. Molyneux, "Twitter as a Tool for and Object of Political and Electoral Activity: Considering electoral context and variance among actors" Journal of Information Technology & Politics, **14**, 154–167, 2017. https://doi.org/10.1080/19331681.2017.1308289
[14] M. M. Skoric, J. Liu, K. Jaidka, "Electoral and Public Opinion Forecasts with Social Media Data: A Meta-Analysis" Information, **11**(4), 187, 2020. https://doi.org/10.3390/info11040187
[15] T. H. McCormick, H. Lee, N. Cesare, A. Shojaie, E. S. Spiro, "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing" Sociol Methods Res. **46**(3), 390–421, 2017. https://doi.org/10.1177/0049124115605339
[16] A. Dür, "How Interest Groups Influence Public Opinion: Arguments Matter more than the Sources" European Journal of Political Research, **58**(2), 514–535, 2019. https://doi.org/10.1111/1475-6765.12298
[17] J. T. Feezell, "Agenda Setting through Social Media: The Importance of Incidental News Exposure and Social Filtering in the Digital Era" Political Research Quarterly, **71**(2), 1–13, 2018. https://doi.org/10.1177/1065912917744895
[18] Y. Tsfati, J. Cohen, Perceptions of Media and Media Effects, The International Encyclopedia of Media Studies: Media Effects/Media Psychology, 1st Edition, Blackwell Publishing Ltd, 2013.
[19] M. Ibrahim, O. Abdillah, A. F. Wicaksono, M. Adriani, "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation" in 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 2015. https://doi.org/10.1109/ICDMW.2015.113
[20] D. Contractor, T. A. Faruquie, "Understanding Election Candidate Approval Ratings using Social Media Data" in 2013 International Conference on World Wide Web (www '13 Companion), Rio de Janeiro, Brazil, 2013. https://doi.org/10.1145/2487788.2487883
[21] J. E. Chung, E. Mustafaraj, "Can Collective Sentiment Expressed on Twitter Predict Political Elections?" in 2011 AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 2011.
[22] B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series" in 2010 AAAI Conference on Weblogs and Social Media, Washington, DC, USA, 2010.
[23] D. Gayo-Avello, "Don't Turn Social Media into Another 'Literary Digest' Poll" Communications of the ACM, **54**, 121–128, 2011. https://doi.org/10.1145/2001269.2001297
[24] The Arab Barometer, "Morocco Five Years after the Arab Uprisings", 2017. https://www.arabbarometer.org/wp-content/uploads/Morocco_Public_Opinion_Survey_2016.pdf
[25] K. Ait Hadi, R. Lasri, A. El Abderrahmani, "An efficient Approach for Sentiment Analysis in a Big Data Environment" International Journal of Engineering and Advanced Technology (IJEAT), **8**(4), 263–266, 2019.
[26] R. Eskander, "SentiArabic: A Sentiment Analyzer for Standard Arabic" in 2018 Language Resources and Evaluation Conference (LREC'18), Miyazaki, Japan, 2018.
[27] H. Bouamor, et al., "The MADAR Arabic Dialect Corpus and Lexicon" in 2018 Language Resources and Evaluation Conference (LREC'18), Miyazaki, Japan, 2018.
[28] W. H. Greene, Econometric Analysis, 7th Edition, Pearson Education, Prentice Hall, 2012.