 Future Technologies and Innovations (FTI) Proceedings: 4th international conference on computer applications and information security (iccais'2021) / March 19 / 2021/ Tunisia: https://fti-tn.net/iccais-2021-list-of-papers

Short Text Topic Modeling For the Arabic Language

Alhanouf Almutairi
Computer Science Department
College of Computer and Information Sciences
King Saud University
439203697@student.ksu.edu.sa

Najwa Altwaijry
Computer Science Department
College of Computer and Information Sciences
King Saud University
ntwairi@ksu.edu.sa

***Abstract*— Topic modeling is an active research area that involves statistical methods to detect patterns of words used within documents. It has been used recently on social networks, for example, Facebook, Instagram, and Twitter. Finding the characteristics of text in Twitter is significant for various applications, such as events detection, personalized message recommendation, or user recommendations, among others. Unfortunately, this has not been extensively studied for the Arabic language. In this paper, we describe a machine learning approach to conceptualize topics used by groups of users on Twitter via their Arabic tweets. We study two topic modeling models: Latent Dirichlet Allocation (LDA) and the Correlated Topic Model (CTM), which is an extension of LDA. Our models are used to discover topics a user is interested in, and tweets about. Results show that LDA outperforms CTM in terms of human interpretability and topic similarity.**

Keywords- Topic Modeling, Latent Dirichlet Alloca- tion, Correlated Topic Model, Twitter, Arabic.

I. INTRODUCTION

Online social networking has become the most important medium and information resource for many individuals around the world. A large amount of data from various sources is posted each second on social networks. Twitter is a popular social network site, enabling millions of users to keep in contact, participate in experiences, publish data, exchange views, and discuss topics. Users can follow different accounts (e.g. friends, famous people, or company accounts) to obtain the most recent information through 140- character messages. A user might conceivably wish to follow accounts that share his or her own interests. Alternatively, a company might wish to discover the interests of their customers, and make recommendations based on those interests. Another scenario involves a company that needs to discover customer sentiments towards its products and services. Thus, a technique to automatically discover the topics or interests of a Twitter account is beneficial.

Rapid accumulation of information requires suit- able tools and techniques that automatically organize, summarize, and understand a large collection of information. Topic modeling is an unsupervised content mining technique that is able to extract relevant topics from a group of documents, i.e. determine recurring patterns of words in textual documents [2]. Topic modeling studies terms within a corpus of documents that represent a mix of topics; a topic can be a likelihood distribution over words. A topic model is a generative model for a corpus, characterized by a straightforward probabilistic method. By utilizing probabilistic methodology, documents can be created and new documents produced by selecting an appropriate distribution over topics. At that point, each word in a document is assigned to a topic randomly depending on the distribution.

In this research, we use machine learning techniques to perform topic mining on Arabic language Twitter accounts. We conceptualize topics using a sample of an account's followers, starting with a random user account, using Latent Dirichlet Allocation (LDA) and the Correlated Topic Model (CTM). Understanding the topics discussed on Twitter provides significant value to various fields, such as marketing [18], education [6], and management [19]. Topic modeling provides a powerful method for categorizing text documents according to their respective topics. Consequently, we applied topic modeling to followers' tweets for a random user account to study the interests of these followers.

The aim of this paper is to develop a model that is capable of assigning a topic to a Twitter account. This topic can further be used to assist users in choosing who to follow, or to make recommendations. Specifically, this paper presents (1) a novel LDA model for Twitter users Arabic tweets, (2) a novel CTM model for Twitter users Arabic tweets, (3) a comparison of the two models, based on human interpretation as well as similarity measures, and sample output of our model. To the best of our knowledge, this is the first work on using LDA with Arabic with the intention of assigning topics to Twitter accounts.

The rest of this paper is organized as follows: Section II discusses other related works. Section III presents the two topic modeling techniques used in this paper. Section IV describes our proposed system. Section V presents the experimental results, and Section VI concludes our work.

II. RELATED WORK

Topic modeling is a common tool for understanding a large amount of unstructured data, such as detecting important topics in social networks. It is an unsupervised probabilistic model utilized to detect and assign large documents to topics [8]. Topic

modeling utilizing Twitter data has been studied by researchers. Twitter has unique challenges compared with other textual data, because of the unstructured language form and nonstandard type of language [22]. The following outlines some relevant research on topic modeling on Twitter. For Arabic topic classification, the interested reader is referred to [1], [5], [15], [23].

Hidayatullah et al. [17] applied topic modeling using LDA to tweets about football news in Bahasa, Indonesia. Through the content analysis, they found many topics, such as pre-match analysis, live match updates, football club achievements, etc. Hidayatullah and Ma'arif [16] applied topic modeling using LDA to obtain the most important topics in traffic information posted by the official Twitter accounts of traffic management centers in each region in Java Island, Indonesia. The most dominant topic found in the data set was the regular monitoring of traffic conditions. This topic was elaborated into more specific terms in 28 topical segments that are common terms to describe road traffic situations.

Rohani et al. [21] developed a practical topic model based on LDA to find topics from social media datasets with 90,527 records, in the domain of aviation and airport management. In order to improve the accuracy of the implemented topic modeling algorithm, they identified a list of generic keywords, including 645 common English and Malay words, that were excluded from the studied datasets. In the proposed method, the model accurately detected five main topics discussed on social media within the studied domain of aviation. Furthermore, the dynamics of topics were visualized per day based on a probabilistic model.

In the area of Arabic topic mining, numerous researchers worked on sentiment analysis and topic modeling, although research is lagging behind the amount of Arabic online content [20]. Brahmi et al. [11] proposed a new Arabic stemming technique and applied it to newspaper articles, then did topic modeling using LDA. The discovered topics were then used in a supervised learning manner, and the supervised learning algorithm achieved better results using the discovered topics when compared with supervised learning in the full word space.

Beseiso [7] presented a sentiment analysis model on Arabic tweets. After preprocessing, LDA was used to extract topics that were input into a Sup- port Vector Machine (SVM) classifier to output the corresponding sentiment. Their model outperformed a na¨ive Bayes classifier and an SVM with a lexicon in terms of precision and F-measure. In [24], Zarra et al. applied LDA on Facebook comments focused exclusively on Maghrebi Arabic. An interesting observation was that for their LDA model, the higher the number of iterations is, the more stable the model convergence.

Alhawarat and Hegazi [4] applied LDA and the k- means clustering algorithm to a news dataset. First, a Vector Space Model (VSM) is created for the documents based on the bag-of-words model. Then, Term Frequency-Inverse Document Frequency (TF- IDF) weighting is applied to remove unnecessary frequent terms, then data is normalized. Then the data is clustered, or LDA is applied then the data is clustered. The second combined algorithm produced better results. Alhawarat [3] applied LDA to the words of the holy Quran chapter "Joseph". They show that the best structure to use is verses, which give the least energy for data. These results suggest that topic modeling techniques failed to capture in an accurate manner the coherent topics of the chapter.

Although Arabic language works exist for topic classification, to the best of our knowledge, we did not find research on Twitter user interests in the Arabic language. In the next chapter, we present our proposed framework to study user interests according to tweets in the Arabic language using LDA and CTM models.

III. TOPIC MODELING

A topic model is a probabilistic model that is frequently utilized in the domain of computer science. It has received a lot of interest from re- searchers in several fields. Topic modeling includes processes to discover the type of words that appear within documents. Topics are typically thought of as a distribution of words, and documents include mixtures of topics. Regarding document clustering, topic modeling can be utilized to cluster documents by providing a probability distribution over several topics for every document. In this section, we will introduce the two topic modeling algorithms used in this work: Latent Dirichlet Allocation (LDA) and the Correlated Topic Model (CTM).

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised generative probabilistic technique, proposed by Blei, Ng, and Jordan [10]. It is one of the most known and used models for topic modeling. In LDA, every document appears as a probabilistic distribution through latent topics, and every latent topic also appears as a probabilistic distribution through words, where both distributions share a Dirichlet prior. It is defined as follows: a corpus D contains M documents, and documents d contains N_d words (d $\in$ 1,...,M):

1) *Pick a multinomial distribution* φ_t for topic t, t $\in$ {1,...,T}, from a Dirichlet distribution with parameter β.
2) *Pick a multinomial distribution* θ_d for document d, d $\in$ {1,...,M}, from a Dirichlet distribution with parameter α.
3) *For a word* $w_{n,}$ n $\in$ 1,..., N_d, in document d:
   a) *Choose a topic* z_n *from* θ_d
   b) *Choose a topic wn from* φ_{zn}

Words in documents are only noticed variables, while the φ and θ and their corresponding hyperparameters α and β are latent variables. The probability of D is calculated and obtained from a corpus as follows:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)$$

$$\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d \quad (1)$$

B. Correlated Topic Model

A limitation of LDA is its inability to model topic correlation because of the Dirichlet distribution. The Correlated Topic Model (CTM) [9] was developed to address this lack by allowing topics to exhibit correlation via the logistic normal distribution. CTM models words in documents using a mixture model. Every document shares the same mixture components in appropriate proportions. CTM builds a generative model as follows: Given topics $\beta_{1:K}$, a K- vector μ and a K × K covariance matrix Σ:

1) Draw $\eta_d|\{\mu,\Sigma\}\sim N(\mu,\Sigma)$
2) For $n \in \{1,\ldots,Nd\}$:
   a) Draw topic assignment $Z_{d,n}|\eta_d$ from $Mult(f(\eta_d))$.
   b) Draw word $W_{d,n}|\{Z_{d,n},\beta_{1:K}\}$ from $Mult\left(f\left(\beta_{zd,n}\right)\right)$.

where $f(\eta)$ maps a natural parameterization of the topic proportions to the mean parameterization.

IV. METHODOLOGY

This study aims to utilize machine learning approaches to conceptualize topics used by groups of users on Twitter via their tweets of text. First, data is extracted from the source (Twitter). In this phase, we retrieve all possible tweets of followers for a randomly selected twitter account. Next, the textual elements are preprocessed to remove common punctuation and stop words. Then, the LDA and CTM models are trained to learn topics from the tweets. The overall structure of the framework is shown in Figure 1.

A. Data Extraction

First, we select a random Twitter user account (@UKinSaudiArabia). Then, 100 followers are randomly selected and the 200 most recent Arabic tweets are sampled, for a dataset of size 20000 × 90. Most of the columns contain specific Twitter information such as tweet time, or favorite count, and are unneeded for our purposes. We select the user name column and the tweet text column. All tweets by a single user are then merged into one row to represent one user document.

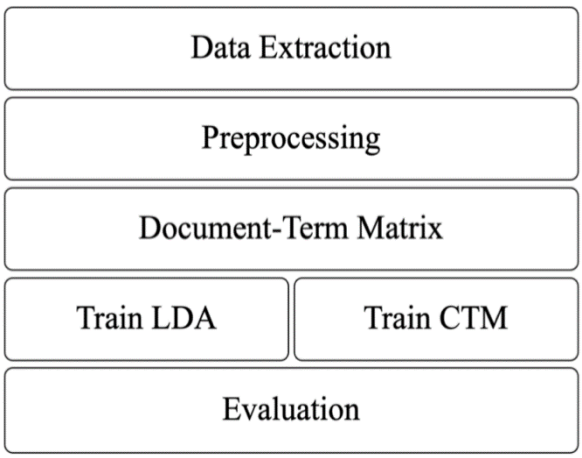

Fig. 1: Overall structure of our framework

B. Preprocessing

The preprocessing step is fundamental in any data mining task, such as text mining, as it improves accu- racy and runtime. Preprocessing Arabic text is known to be a challenging process due to the complexity of Arabic grammatical rules. Hashtags and underscores are replaced by a space, while diacritics, stop words, newlines, tabs and URLs are removed. All non- Arabic characters are removed, and the "Hamza" character is replaced with "Alif". Finally, stemming is performed.

C. Document-Term Matrix

A document-term matrix describes the frequency of terms that occur in a collection of documents. Rows represent documents, while columns represent words. Each cell represents the frequency of the words in each document. If a document does not contain a word, then the cell value is zero. A small sample of our document-term matrix is shown in Table I.

TABLE I: Document-Term Matrix Sample (User names have been anonymized).

User	الله	برنامج	خير
User 1	28	4	3
User 2	11	0	0
User 3	0	1	0
User 4	48	1	0
User 5	0	1	0
User 6	10	0	1

D. Model Training and Evaluation

We train two models: LDA with Gibbs sampling and CTM with VEM, as outlined in Section III. We evaluate both models in two steps:

1) Selection of number of topics k: The number of topics is a hyperparamter selected by the user, and it can influence the interpretability of the results. A model with too few topics may result in very broad topics, while a model with too many topics may be uninterpretable. Therefore, we used a number of metrics for selecting the number of topics as follows:
1) *CaoJuan2009* [12]: select the value of k that minimizes the average cosine similarity be- tween topic distributions.
2) *Griffiths2004* [14]: estimate LDA parameters via a Gibbs sampler and then select k such that it maximizes the harmonic mean log-likelihood of the Gibbs samples.

3) *Deveaud2014* [13]: similar to CaoJuan2009, however, it maximizes the average Jensen- Shannon distance of topic distributions.

In addition, we use perplexity, a popular method for evaluating a probabilistic model. Perplexity measures how well a probability model predicts a sample, through the log-likelihood of the test set. A low perplexity denotes a better probabilistic model. Perplexity is defined as:

$$Perplexity_{testset} = \exp(-\frac{\sum_{d=1}^{M} \log(w_d)}{\sum_{d=1}^{M} N_d}) \qquad (2)$$

where M is the number of documents in the test set, w_d are the words that appear in document d, and N_d is the number of words in document d.

The various number of topics returned by the above metrics (CaoJuan2009, Griffiths2004, De- veaud2014) is evaluated through the perplexity measure, and the final value of k is chosen to minimize perplexity.

2) Model Comparison: In order to compare the CTM and LDA models, we adopt two methods. First, we apply human interpretation for labeling models. Second, we measure topic similarity, by using cosine similarity and Jaccard coefficient.

$$Cosine(V^i, V^j) = \frac{V^i . V^j}{||V^i|| * ||V^j||} \qquad (3)$$

Where V^i and V^j are semantic vectors.

$$Jaccard(X, Y) = |X \cap Y| / |X \cup Y| \quad (4)$$

where X and Y are the two sets to be measured.

V. EXPERIMENTAL RESULTS

In terms of the candidate numbers of topics, we performed several experiments for LDA and CTM. First, we used the metrics presented in Section IV-D1. Then, we used perplexity with a variable number of topics and 10-fold cross-validation. The experiments were undertaken on a 2.9GHz Intel Core i5 processor with 8GB RAM, and OSX platform.

A. Selecting the optimal number of topics

We utilized the metrics (Griffiths2004 and Cao- Juan2009) with Gibbs for LDA, and (CaoJuan2009 and Deveaud2014) with VEM for CTM. Table II shows the experiments conducted. We selected three different ranges as in the first column. The second column shows the optimal k value via Gibbs-LDA while the third column shows the optimal k value via VEM-CTM.

TABLE II: Identifying the number of optimal topics

Selected range	Gibbs-LDA	VEM-CTM
2-20	20	13
2-30	30	13
2-50	40	13

We used perplexity with 10-fold cross-validation on the different ranges of the candidate numbers of topics. Figure 2 shows that for Gibbs-LDA, as the number of topics increases, the perplexity de- creases. However, the amount of reduction is not commensurate with the addition of topics. We select k = 20 as it was more coherent than others, with $Perplexity = 4182$. For VEM-CTM, perplexity scores show that $k = 14$ is a good number of topics, with $Perplexity = 1147$.

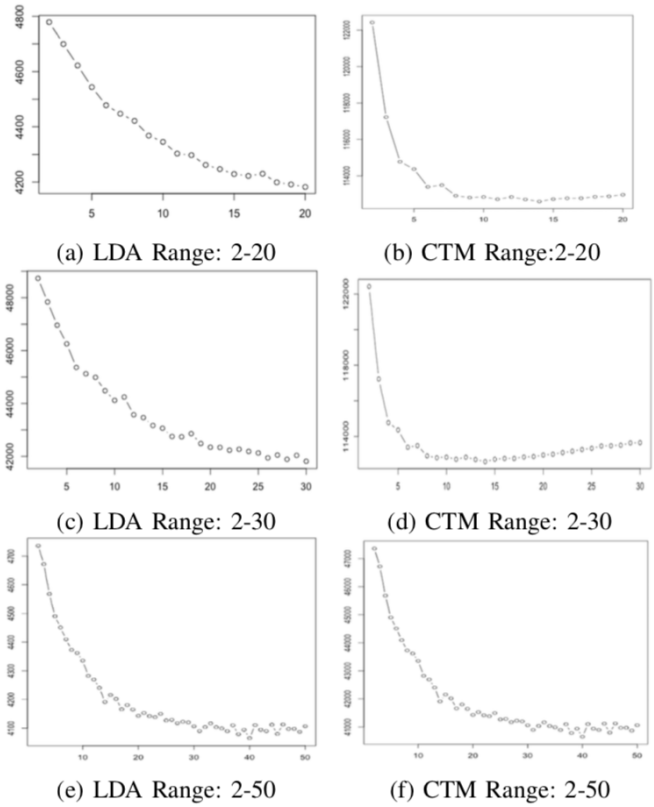

Fig. 2: Perplexity on Gibbs-LDA and VEM-CTM

B. Model Comparison

We compared LDA and CTM in two ways. First, we use human interpretation, and second, by measuring the similarity between topics in each model using cosine similarity [12] and Jaccard coefficient.

1) Human Interpretation: For human interpretation, we asked three specialists to read through a list of words describing each topic in LDA (20 topics) and CTM (14 topics). For LDA, 17 out

of 20 topics were considered coherent by all special- ists (85%) Topics include: countries, competitions, health, sport, language services, jobs, religion, life, political events, emotions, transport, history, econ- omy, poetry, education, and politics. For CTM, 12 out of 14 topics were considered coherent (85.7%). Topics include: education, work, emotions, delivery, traveling, external affairs, politics, and religion. On average, the number of topics humans could interpret was similar for both LDA and CTM. It was noticed that there was an overlap between topics for LDA and CTM, and within the topics of each. All three specialists preferred the LDA topics, as they found them easier to label and more coherent.

2) Similarity Measures: To calculate similarity, we convert the topic terms to vectors and calculate the cosine similarity and the Jaccard coefficient. Table III shows that LDA outperforms CTM on both measures, as the similarity between LDA topics is smaller than the similarity between CTM topics. This means that LDA is able to produce topics that are more dissimilar, i.e. there is less overlap of topics.

TABLE III: Similarity Measures

Model	Cosine similarity	Jaccard coefficient
LDA	0.075598	0.063447
CTM	0.246767	0.172908

3) *Sample output:* In this section, we present some sample output from our algorithm. Since LDA outperforms CTM, we chose LDA to assign to each user a topic. We randomly selected two users whose topic was "health", and show their tweets in Figure 3, and two users whose topic was "political events", and show their tweets in Figure 4. As seen from the figures, LDA was able to assign topics to users accurately.

Fig. 3: Sample tweets from users assigned to topic "health"

VI. CONCLUSION

In this paper, we studied two machine learning algorithms to discover user interests, modeled as latent topics in Twitter. We proposed a framework for applying topic modeling to Arabic tweets for a group of users and evaluated the efficiency of the proposed framework. Our proposed framework applied LDA and CTM to Twitter followers' Arabic tweets that were extracted from Twitter to study users' interests. We randomly selected 100 followers of a Twitter account, and sampled 200 recent tweets. We evaluated the models in two steps: first, the best number of topics for the two models are selected according to many experiments using CaoJuan2009, Griffiths2004, Deveaud2014, and perplexity measurements with different ranges of topics. Second, we compared the two models in terms of human interpretation by labeling the topics and similarity between topics using cosine and Jaccard similarity. The results show that the optimal numbers of a topics for LDA is 20 while CTM is 14. We note that when number of topics in CTM increases, the results are redundant topics. LDA outperformed CTM in terms of human interpretation of labeling the topics and in terms of similarity measures between topics. Results show that LDA has more independent topics and is easier to label. We assigned a topic to each user based on LDA results, and provided sample output. For future work, we intend to enhance our system by applying word embedding to improve the discovery of related words in topics models. In addition, the results can be used to build a user recommendation system on Twitter based on shared interests.

Fig. 4: Sample tweets from users assigned to topic "political events"

REFERENCES

[1] Abdullatif Alabdullatif, Basit Shahzad, and Esam Alwagait. Classification of arabic twitter users: a study based on user behaviour and interests. Mobile Information Systems, 2016, 2016.

[2] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 6(1), 2015.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange

[3] Mohammad Alhawarat. Extracting topics from the holy quran using generative models. International Journal of Advanced Computer Science and Applications, 6(12):288– 294, 2015.

[4] Mohammad Alhawarat and M Hegazi. Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. IEEE Access, 6:42740–42749, 2018.

[5] Amani Alhozaimi and Mishari Almishari. Arabic twitter profiling for arabic-speaking users. In 2018 21st Saudi Computer Society National Computer Conference (NCC), pages 1–6. IEEE, 2018.

[6] Subhasree Basu, Yi Yu, Vivek K Singh, and Roger Zimmer- mann. Videopedia: lecture video recommendation for edu- cational blogs using topic modeling. In International Con- ference on Multimedia Modeling, pages 238–250. Springer, 2016.

[7] Majdi Beseiso. New sentiment analysis model using lda for arabic tweets. In Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence, pages 212–216, 2019.

[8] David M Blei. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012.

[9] [David M Blei, John D Lafferty, et al. A correlated topic model of science. The annals of applied statistics, 1(1):17– 35, 2007.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. La- tent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.

[11] Abderrezak Brahmi, Ahmed Ech-Cherif, and Abdelkader Benyettou. Arabic texts analysis for topic modeling eval- uation. Information retrieval, 15(1):33–53, 2012.

[12] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive lda model selection. Neurocomputing, 72(7-9):1775–1781, 2009.

[13] RomainDeveaud,EricSanJuan,andPatriceBellot.Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique, 17(1):61–84, 2014.

[14] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1):5228–5235, 2004.

[15] Fouzi Harrag, Eyas El-Qawasmeh, and Pit Pichappan. Im- proving arabic text categorization using decision trees. In 2009 First International Conference on Networked Digital Technologies, pages 110–115. IEEE, 2009.

[16] Ahmad Fathan Hidayatullah and Muhammad Rifqi Ma'arif. Road traffic topic modeling on twitter using latent dirichlet allocation. In 2017 International Conference on Sustainable Information Engineering and Technology (SIET), pages 47– 52. IEEE, 2017.

[17] AhmadFathanHidayatullah,ElangCergasPembrani,Wisnu Kurniawan, Gilang Akbar, and Ridwan Pranata. Twitter topic modeling on football news. In 2018 3rd International Conference on Computer and Communication Systems (IC- CCS), pages 467–471. IEEE, 2018.

[18] Noor Farizah Ibrahim and Xiaojun Wang. A text analytics approach for online retailing service improvement: Evidence from twitter. Decision Support Systems, 121:37–50, 2019.

[19] Nicolas Pröllochs and Stefan Feuerriegel. Business ana- lytics for strategic management: Identifying and assessing corporate challenges via topic modeling. Information & Management, 57(1):103070, 2020.

[20] Ahmed Rafea and Nada A GabAllah. Topic detection approaches in identifying topics and events from arabic corpora. Procedia computer science, 142:270–277, 2018.

[21] Vala Ali Rohani, Shahid Shayaa, and Ghazaleh Babane- jaddehaki. Topic modeling for social media content: A practical approach. In 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), pages 397–402. IEEE, 2016.

[22] Asbjørn Steinskog, Jonas Therkelsen, and Bjørn Gambäck. Twitter topic modeling by tweet aggregation. In Proceedings of the 21st nordic conference on computational linguistics, pages 77–86, 2017.

[23] Fadi Thabtah, M Eljinini, Mannam Zamzeer, and W Hadi. Naïve bayesian based on chi square to categorize arabic data. In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, pages 4–6, 2009.

[24] Taoufiq Zarra, Raddouane Chiheb, Rajae Moumen, Rdouan Faizi, and Abdellatif El Afia. Topic and sentiment model applied to the colloquial arabic: a case study of maghrebi arabic. In Proceedings of the 2017 international conference on smart digital environment, pages 174–181, 2017.