PalArch's Journal of Archaeology of Egypt / Egyptology

SENTIMENT ANALYSIS WITH LDA ALGORITHM FOR GOVERNMENT POLICY ANALYSIS USING TWITTER

*Dedy Rahman Prehanto[1,2], Suryono[3], Sularto[4],
[1]Student of Doctor of Information System, School of Postgraduate Studies, Universitas Diponegoro
[2]Universitas Negeri Surabaya
[3,4]Universitas Diponegoro
dedyrahman@unesa.ac.id,dedyrahman@students.undip.ac.id,
suryono@fisika.fsm.undip.ac.id, sulartorb@lecturer.undip.ac.id,
*Coressponding Author

Sentiment Analysis With Lda Algorithm For Government Policy Analysis Using Twitter-- Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(18) 699-711. ISSN 1567-214x

Keywords: Latent Dirichlet Allocation, Sentiment Analysis, Government policy, Twitter

ABSTRACT:
The government needs to formulate public social policy opinion, a source of information to improve performance. People use Twitter to post their views about an object or event. When using this opinion, proper analysis is required to use the information generated for policy decisions. The purpose of this study is to use the Latent Dirichlet Allocation (LDA) algorithm and Twitter social media data obtained in real-time using the API provided by Twitter to classify public comments that determine government policy. The results of the analysis show that people's perceptions of government policy on the opinion on Twitter with latent self-allocations formed into 26 topics with a coherence value of 0.53049and the topic that is often discussed is topic 1 with a percentage score of 8.6%, namely regarding government efforts inequality and access to education, health, employment, and infrastructure also contains information on government policies that facilitate business actors in expanding the MSME market.

PRELIMINARY

Democracy is based on the assumption that citizens are educated enough to play a wise role in participation and deliberation (Elena and Gracia, 2020). Government view policy controls are used to control the review process's timeliness, including but not limited to feedforward controls, real-time controls, and feedback control (JunhaiMaa and ChunyongMa, 2020). if there is no appropriate and appropriate way to obtain information

related to governance, such informed participation and deliberation will not be possible (Flores, 2017).
Information systems are a way of representing information that can provide added value. Additional value can be obtained in the form of information based on real data, which is processed so that it is useful for the recipient (Prehanto et al., 2020). In refining this policy, the government needs a public opinion, which is a source of information to improve performance (Harwood and Thower, 2019) People use Twitter to write opinions about objects or events. This opinion can be used to find information. However in its use, it requires proper analysis so that the information generated can help various aspects to support decision making or choices (Gu and Kurov, 2020). According to a recent report released by We Are Social and Hootsuite in July 2019, more than 3.5

Billion people on the planet have joined social media. Meanwhile, Twitter is in fourth place with the 13-17 age group with 20.2 million users. The total number of Twitter users reaches 254 million. In this case, Twitter data can be used to analyze public sentiment towards government policies. The resulting data analysis can show positive, negative, and neutral sentiments of the community, and then the government can use these sentiments to formulate policies for the community.

II. LITERATURE REVIEW

A. Policy.

The government is focused on allocating public policies necessary to achieve the best governance in education and mapping a balanced growth path (Rey and Garcia, 2020). Examples of use are to plan, initiate, organize, control activities, and present information based on data processing. Generate useful information as a reference for determining the final decision. The simple definition of an information system is that it must have input, process, and output. (Prehanto et al., 2020). Information policy includes laws, regulations, doctrinal positions and decisions, and other practices that have a constitutive role in society as a whole, involving the creation, processing, flow, access, and use of information (Braman, 2011).

B. Sentiment Analysis.

Sentiment analysis is to classify each tweet according to its positive, negative, or neutral sentiment (Cimino, 2016). As a result, for every tweet, we get the possibility to fall into the sentiment category. After manual analysis, we used the probability threshold to filter out low confidence predictions; that is, we could not classify high confidence tweets that could not be classified as positive or negative as neutral (Filippoa et al., 2018).

C. Social Twitter

Twitter is a social media and Weibo service that allows users to send messages in real-time. This message is usually called a tweet. (Agarwal et al., 2014). The previous tweet process by removing mentions (@ characters), URLs, product tags, emojis, and single characters (Filippoa et al., 2018). Twitter's sharing structure refers to disseminating research results on Twitter over time and consists of original tweets, retweets, and retweet links. The original tweet is defined as Twitter, which refers to a scientific publication originally issued by Twitter users, and retweet refers to the repartition of the original tweet by Twitter users (Fang and Costas, 2020).

D. Latent Dirichlet *Allocation (LDA)*

Latent Dirichlet *Allocation (LDA)* is a probability model of generating a corpus set. The basic idea is to represent a document as a mixed model of various topics, also called latent topics, where each topic is characterized by a word. based on Blei (2018), Figure 1 below illustrates LDA's working principle.

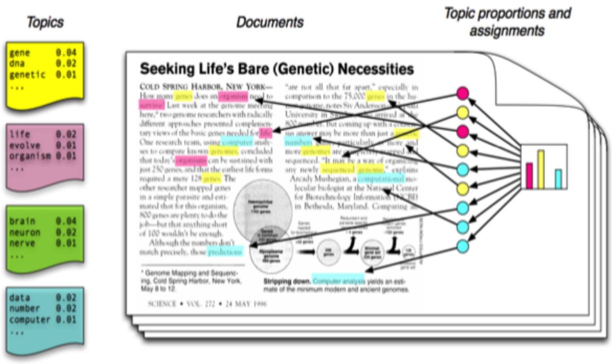


Figure 1. Principles of LDA

LDA assumes that the process of creating each w document in the corpus is as follows:
1. Select N ~ Pissson (x),
2. Select θ ~ Dir (α),
3.For every N own words,
a. Select Topic in ~ Multinomial (θ),
b. Choose a word Wn from p (Wn | Zn, β).

Several simplifying assumptions are made in the distribution of (latent) topics known to follow the k of the Dirichlet distribution. Second, word probability is a matrix of β of size kx V where bij = p (WJ = 1 | Zi = 1). While k as the Dirichlet distribution has a density function, it can be seen in equation (1) as follows:

$$(PQ\ |a) = \frac{r\left(\sum_{i=1}^{k} a_1\right)}{\prod_{i=1}^{k} r(a_1} \theta_1^{a_1-1} + \cdots \theta_k^{a_k-1} \text{.. (1)}$$

As for the form in the joint distribution of Topic mixture θ of N topics z and N-words w conditional α and β can be seen in equation (2) as follows:

$$(PQ, z, w|a, \boldsymbol{\beta}\ ) = p(\theta\ |a) \prod_{n-1}^{N} p(Z_n|\theta) p(w_n|Z_n, \boldsymbol{\beta}\ ) \text{ ... (2)}$$

The shape represented by the LDA model can be illustrated in the figure and can be seen in Figure 2.

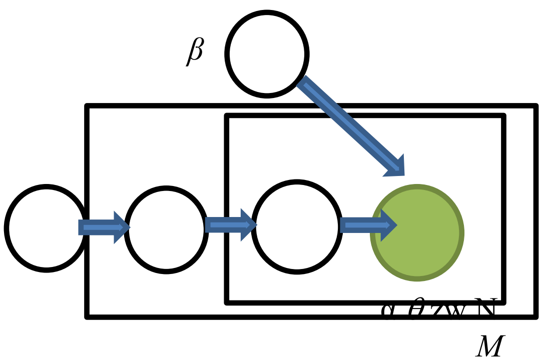


Figure 2. Representation of the LDA Model

The form of the marginal distribution of p (w |α,β) obtained by integrating equation (2) with θ can produce equation (3):

$$(\quad |PWA, \beta) = \int (\theta \ |a) \left(\prod_{n=1}^{N} \sum_{Z_n} p(Z_n|\theta\,) p\left(W_n \middle| Z_{n,}\beta\right) \right) d\theta \ \ldots (3)$$

Finally, we obtain the product of the marginal density for a document, which will obtain the marginal probability of a corpus of equation (4) as follows:

$$(\quad |PDA, \beta) = \prod_{d=1}^{M} \int p(\theta_d \ \ |a) \left(\prod_{n=1}^{N_d} \sum_{Z_n} p\left(W_n \middle| Z_{n,}\beta\right) \right) d\theta_d .. \ (4)$$

E. Preprocessing

The preprocessing process in this study includes two stages, namely tokenization and stemming.

1. *Tokenization*
   Tokenization is a processing stage where input text is divided into small units called tokens or individual blocks. Certain characters will be deleted at this time, such as punctuation marks, characters other than numbers and letters. Another process that occurs in tokenization is folding uppercase or converting all letters to lowercase (lowercase) and removing components that do not match the document (for example, tags, links, and HTML tags), also known as the cleaning process (Aso, Takamichi, and saruwatari, 2020).

2. *Stemming*
   Stemming is the process of changing the form of a word into its root. Due to the different forms of language, the stemming algorithms of each language are also different. Stemming in some prefix variants is omitted to get the root word. (Sun, Zhang, and Ouyang 2020).

F. Evaluation

After completing the LDA training and testing process, conduct an *assessment*. The evaluation process aims to get the best model by calculating the accuracy based on the matrix configuration. Accuracy is calculated using equation (8).

$$Accurancy = \frac{TP+TN}{TP+TN+FP+FN} \ldots\ldots\ldots (8)$$

If several folds are found with the same highest accuracy value, the sensitivity and specificity values will be calculated to determine *the* best model (Gangadharan and Gupta, 2020). The sensitivity and specificity values are calculated using formula (9) and formula (10).

$$Senzitivity = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

$$Specifivity = \frac{TN}{TN+FP} \ldots\ldots \ldots\ldots\ldots (10)$$

Where :
TP: true positive on document

TN: true negative in the document
FP: false positive on document
FN: false negative on the document

III. RESEARCH METHODS

The research method is needed so that the research is more structured so that the results obtained are following the research objectives. The stages of the research method are shown in Figure 3 (Jonasson, 2019):

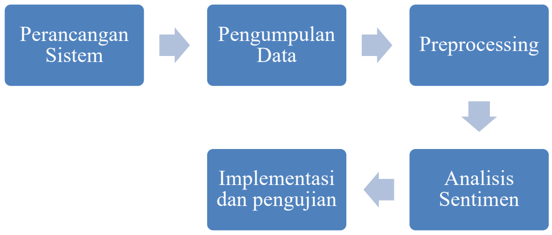


Figure 3. Research Methodology

A. System planning

The sentiment analysis system is designed as follows:

1. Use the Twitter API to retrieve tweets.
   a. Enter keywords related to "government policy."
   b. Save crawling data from various Twitter accounts.

B. Data collection

The data source used comes from a collection of tweets from Twitter users who use the Twitter Application Programming Interface (API) in Indonesia. The data sample is drawn from the two words "government policy" that appear on Twitter.

C. *Preprocessing*

Perform text preprocessing, including (Jonasson, 2019):

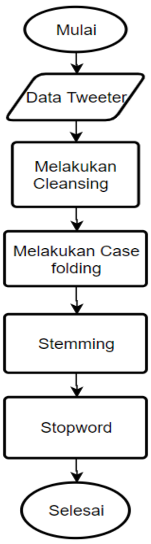


Figure 4 Preprocessing Stages.

a. Remove usernames, hashtags, RTs, blank lines, punctuation marks, spaces, and extra URLs.
b. Perform case folding of case folding, change all characters to lowercase,
c. Perform stemming analysis and convert words into root words
d. Performs a stopword.

D. Sentiment analysis using algorithms Latent Dirichlet Allocation (LDA)

This study uses training data from pre-determined tweet data up to 100 classification data using the TF-IDF method (i.e., word weight). The test data used to predict classification, or data with unknown classification is data obtained in real-time from Twitter using the API (Momtazi, 2018). By determining the number of topics and the number of iterations, the LDA method can model these topics. And model topics based on the number of topics with a coherence value (Joshi and Kanseri, 2020).

E. Implementation and Testing

At this stage, all the content planned in the previous design and design will be applied. This stage will also determine the success of the system to be built. The testing topic modeling method using the Gibbs sampling algorithm is done by dividing the data into k subsets with the same amount of data. In this study, the amount of data used was 100 data records. In this test, the data will be divided into 1, 2, 3, and 4 topics. One topic will be selected from each topic according to the number of topics as testing data, and the other topics will be used as training data.

IV. RESULTS AND DISCUSSION

Modeling Topics Using the Gibbs Sampling algorithm

Wordcloud is one way to find out how many terms (words) appear in the analysis. Here is a word cloud of Twitter users' analysis of government policy.

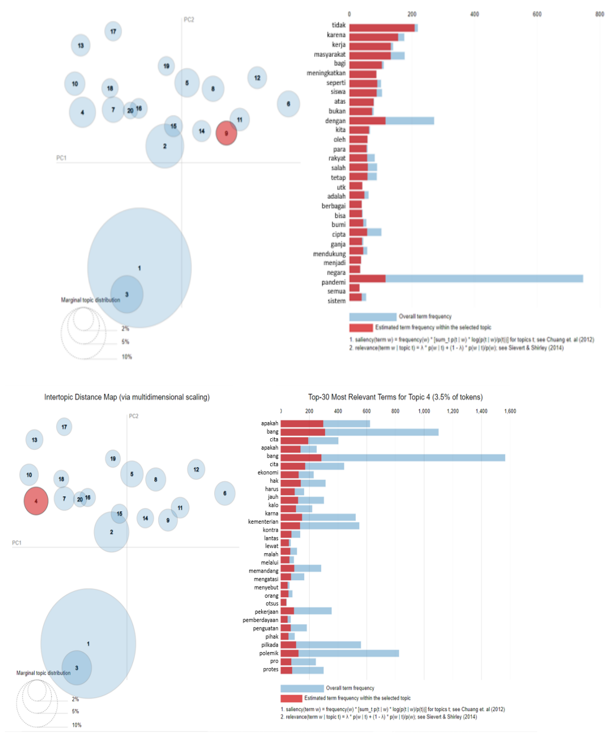
Figure 5. Wordcloud data Twitter "government policy."

Figure 5 shows that the word that appears frequently is 86 times the word Policy, and the word government is 82 times. Other words that often appear are "this, the" "in," "from," "Indonesia." After knowing which words appear frequently, the coherence value is used to select the best word (many groups) through various iterations.

Table 1 Iteration Results

Iteration	Group optimization	Value of Coherence
100	69	0.49
200	26	0.53049
300	17	0.46634
400	15	0.45981
500	23	0.45984
600	11	0.49772
700	14	0.34521
800	13	0.47998
900	12	0.43021
1000	19	0.45964

Table 1 shows that the best coherence value after 200 iterations was formed as many as 26 groups and the coherence value of 0.53049. Once you know that many groups formed, you know what words formed in that group. The following are the results of the LDA.



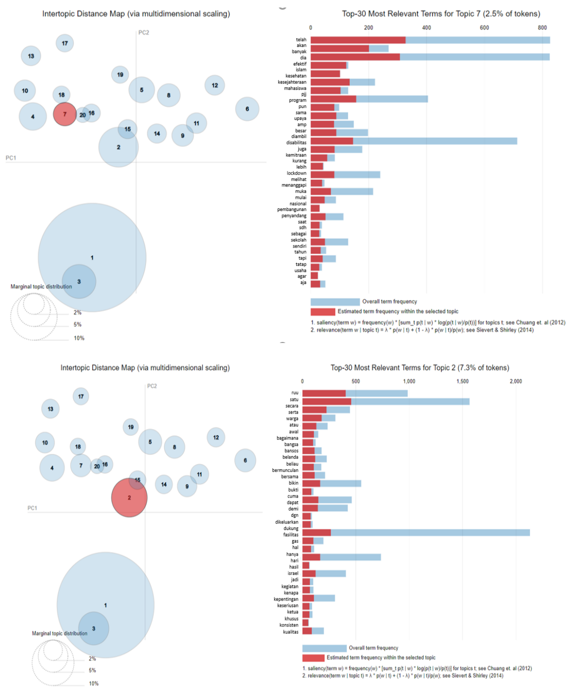


Figure 6.Latent Dirichlet Allocation (LDA) Visualization

Figure 6 shows the results of topic modeling using the latent Dirichlet allocation algorithm that makes up 26 topics. It shows that several topics overlap, which means they have the same few words. Topic 2 and Topic 9 are synonymous with the terms "policy, governance, improvement and society." The 4 biggest topics formed by the topic modeling analysis are also discussed as follows.

1. Topic 1 LDA

   Table 2 shows the results of the analysis of modeling topic 1 using the LDA algorithm.

Table 2. Topic 1

Percentage	Word	Opportunity
	and	0.0248
	this	0.0219
	that	0.0207
	to	0.0178
	in	0.0152
	from	0.0140
Topic 1	Indonesia	0.0134
	that	0.0130
8.6	there is	0.0117
	Papua	0.0114
	What	0.4846
	on	0.0173
	not	0.0146
	Public	0.0140
	Upgrade	0.4856

Table 2 shows that topic 1 contains several words, such as "what," "community," and "improvement," indicating that topic 1 scores a percentage value of 8.6% of the total number of topics formed. The word most likely to appear on topic 1 increases its value by 0.04856. Topic 1 concerns government efforts on equity and access to education, health, jobs, and infrastructure. Topic 1 also contains information on government policies that facilitate business actors in expanding the MSME market.

2. Topic 2 LDA

Table 3 shows the results of the analysis of modeling topic 2 using the LDA algorithm.

Table 3. Topic 2

Percentage	Word	Opportunity
Topic 2 5.6	for	0.0094
	college	0.0089
	student	0.0089
	as	0.3889
	students	0.0130
	on	0.0115
	not	0.0110
	with	0.0105
	we	0.0090
	by	0.0075
	the	0.0085
	people	0.0185
	wrong	0.0080
	permanent	0.0083
	for	0.3

Table 3 shows that the word "student" has the greatest chance of topic 2, equal to 0.3889. Topic 2 also contains words like "by," "student," and "person." Topic 2 score percentage value of 5.6% of the total number of topics formed. Topic 2 contains face-to-face school policy information because the PJJ system is considered ineffective.

3. Topic 3 LDA

Table 4 shows the results of the analysis of topic 3 using the LDA algorithm.

Table 4. Topic 3

Percentage	Word	Opportunity
Topic 3 5.3	various	0.0111
	can	0.0099
	earth	0.0094
	create	0.009
	marijuana	0.0077
	support	0.0073
	Becomes	0.0053
	country	0.0093
	pandemic	0.0069
	all	0.0069
	system	0.3000

	has been	0.0111
	will	0.0077
	Lots	0.0073
	he	0.007

Table 4 shows that topic 3 scores a percentage value of 5.3% of the total number of themes formed. The fourth topic contains several words, such as the word "system," "various," "copyright," etc. Topic 3 shows policies related to copyright of works, omnibus law priorities, job creation laws, and tax facilities.

4. Topic 4 LDA

Table 5 shows the results of the analysis of topic 4 using the LDA algorithm.

Table 5. Topic 4

Percentage	Word	Opportunity
Topic 4 4.8	effective	0.0060
	Islam	0.0057
	health	0.0055
	well-being	0.0067
	college	0.0053
	student	0.0056
	hj	0.2333
	program	0.0087
	even	0.0077
	same	0.0073
	effort	0.0042
	amp	0.0036
	big	0.0034
	taken	0.0033
	disabilities	0.0036
	too	

Table 5 shows that topic 4 scores a percentage value of about 4.8% of the total number of topics formed. The fourth topic contains several words, such as "program," "health," "welfare," and so on. Topic 4 is about maintaining public health and implementing effective policies to avoid worsening pandemic in Indonesia.

V. CONCLUSION

In government policy, the LDA clustering method (latent Dirichlet allocation) dividing Twitter data into 26 topics, while the 4 largest topics formed were topic 1 at 8.6% namely about government efforts inequality and access to education, health, jobs, and infrastructure. Topic 1 also contains information on government policies that facilitate business actors in expanding the MSME market. Topic 2, with a score of 5.6% of the total number of topics formed. Topic 2 contains face-to-face school policy information because the PJJ system is considered ineffective. Topic 3, with a score of 5.3%, indicates policies related to copyright of works, omnibus law priorities, employment creation laws, and tax facilities. Lastly, Topic 4, the percentage score of about 4.8%, is about maintaining public health and implementing effective policies to avoid the worsening of Indonesia.

DAFTAR PUSTAKA

Agarwal.A, Xie.B, Vovsha.I, Rambow.O, and Passonneau.R, , 2014. Sentiment Analysis of Twitter Data, Dep. Comput. Sci. Columbia

Aso, Masashi, Takamichi, Shinnosuke and Saruwatari, Hiroshi(2020) Acoustic model-based subword tokenization and prosodic-context extraction without language knowledge for text-to-speech synthesis. Speech CommunicationVolume 125, Pages 53-60.

Berliana.G, Shaufiah, dan Sa'adah, Siti . 2018. Klasifikasi Posting Tweet mengenai Kebijakan Pemerintah Menggunakan Naive Bayesian Classification. e-Proceeding Eng., vol. 5, no. 1, pp. 1562–1569.

Blei, D.M. 2018. Probabilistic topic models.Comunications of the ACM,

[Braman, sandra (2011) defining information policy. Journal of information policy 1 vol 1-5 Univ. New York, NY 10027 USA

Chiarello, Filippoa ; Bonaccorsi, Andreaa ; Fantoni, Gualtieroa ;Ossola, Giacomoa ; Cimino, Andreab and Dell'Orletta, Feliceb (2018). Technical Sentiment Analysis: Measuring Advantages and Drawbacks of New Products Using Social Media. 2nd International Conference on Advanced Research Methods and Analytics (CARMA2018) Universitat Politecnica de Val ` encia, Val ` encia` DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8336

Cimino-Isaacs, C. (2016). Trans-Pacific Partnership: An Assessment. Washington: Peterson Institue for International Economics.

Prehanto, Dedy Rahman, Aries Dwi Indriyanti, Ginanjar Setya Permadi, Tanhella Zein Vitadiar, F D Jayanti. (2020). Library Book Modeling Data Using the Association Rule Method with Apriori Algorithm in determining Book Placement and Analysis of Book Loans. International Journal of Advanced Science and Technology, 29(05), 1244 - 1250.

[Elena Del Rey and Garcia, Miguel-Angel Lopez (2020) On government-created credit markets for education and endogenous growth. Economic Modelling.

Fang, Zhichao dan Costas, Jonathan Dudek Rodrigo(2020) The stability of Twitter metrics: A study on unavailable Twitter mentions of scientific publications. Journal of the Association for Information Science and TechnologyVolume 71, Issue 12.

Flores, Jorge Alberto Rosas(2017) Elements for the development of public policies in the residential sector of Mexico based in the Energy Reform and the Energy Transition law. Energy Policy9 February 2017Volume 104 ,Pages 253-264.

Gangadharan, veena and gupta, deepa (2020) Recognizing Named Entities in Agriculture Documents using LDA based Topic Modelling Techniques. Procedia Computer Science Volume 171, Pages 1337-1345.

Gu, Chen and Alexander Kurov (2020) Informational role of social media: Evidence from Twitter sentiment. Journal of Banking & Finance.Volume 121 Article 105969

Harwood, Chris G. and Sam N. Thrower(2019)Performance Enhancement and the Young Athlete: Mapping the Landscape and Navigating Future . Kinesiology Review

Jonasson, johan (2019) Slow mixing for Latent Dirichlet Allocation. Statistics & Probability Letters Volume 129 Pages 96-100

Joshi, Stuti and Kanseri, Bhaskar (2020) Spatial coherence properties of down converted biphoton field generated using partially coherent pump beam. Optik, Volume 217, Article 164941.

JunhaiMa and ChunyongMa(2020) Factor Analysis Based On The COSO Framework And The Goverment Audit Performance Of Control Theory. Procedia Engineering. Volume 15, 2011, Pages 5584-5589

Momtazi, Saeedeh (2018) Unsupervised Latent Dirichlet Allocation for supervised question classification. Information Processing & Management Volume 54, Issue 3 .Pages 380-393

Sun, Heng , Zhang, Xiaolei, and Ouyang, Hongwei (2020) Sodium lactate promotes stemness of human mesenchymal stem cells through KDM6B mediated glycolytic metabolism. Biochemical and Biophysical Research Communications. Volume 532, Issue 3 Pages 433-439

Shabbir, M. S., Abbas, M., Aman, Q., Ali, R., & Orangzeb, K. (2019). Poverty Reduction Strategies. Exploring the link between Poverty and Corruption from less developed countries. Revista Dilemas Contemporáneos: Educación, Política y Valores. http://www.dilemascontemporaneoseducacionpoliticayvalores.com/

Shabbir, M. S., Abbas, M., & Tahir, M. S. (2020). HPWS and knowledge sharing behavior: The role of psychological empowerment and organizational identification in public sector banks. Journal of Public Affairs. https://doi.org/10.1002/pa.2512

Shabbir, M. S., Asad, M., Faisal, m., & Salman, R. (2019). The Relationship between Product Nature and Supply Chain Strategy: An Empirical Evidence. International Journal of Supply Chain Management, 8(2), 139-153. http://excelingtech.co.uk/

Shabbir, M. S., Bait Ali Sulaiman, M. A., Hasan Al-Kumaim, N., Mahmood, A., & Abbas, M. (2020). Green Marketing Approaches and Their Impact on Consumer Behavior towards the Environment-A Study from the UAE. Sustainability, 12(21), 8977. https://doi.org/10.3390/su12218977

Shabbir, M. S., Siddiqi, A. F., Kassim, N. M., Mustafa, F., & Salman, R. (2019). A Child Labour Estimator: A Case of Bahawalpur Division. Social Indicators Research, 147(1), 95-109. https://doi.org/10.1007/s11205-019-02146-4

Shahid, K., & Shabbir, M. S. (2019). HOLISTIC HUMAN RESOURCE DEVELOPMENT MODEL IN HEALTH SECTOR: A PHENOMENOLOGICAL APPROACH. Polish Journal of Management Studies, 20(1), 44-53. https://doi.org/10.17512/pjms.2019.20.1.04

Siddiqi, A., Muhammad, S., Shabbir, M. S., Khalid, F., Salman, R., & Farooq, M. (2019). A short comment on the use of R 2 adj in Social Science-7907. REVISTA SAN GREGORIO, 30, 24-31.

Sulaiman, B. A., Shabbir, M. S., & Rana, S. (2020). Oman's ability to Attract FDI: Dunning Instrument Survey Analysis. Propósitos y Representaciones, 8(SPE2). https://doi.org/10.20511/pyr2020.v8nspe2.640

Sulaiman, B. A., Ahmed, M. N., & Shabbir, M. S. (2020). COVID-19 Challenges and Human Resource Management in Organized Retail Operations/Desafíos del Covid-

19 y la administración de recursos humanos en operaciones minoristas organizadas. Utopia y Praxis Latinoamericana, 25, 81-92. http://doi.org/10.5281/zenodo.4280092

Ul-Hameed, W., Shabbir, M. S., Imran, M., Raza, A., & Salman, R. (2019). Remedies of low performance among Pakistani e-logistic companies: The role of firm's IT capability and information communication technology (ICT). Uncertain Supply Chain Management, 369-380. https://doi.org/10.5267/j.uscm.2018.6.002

Zulfikar, Muhammad Taufiq and Suharjito (2019) Detection Traffic Congestion Based on Twitter Data using Machine Learning. Procedia Computer Science1 October 2019Volume 157 Pages 118-124