Twitter Topic Modeling on Football News

Ahmad Fathan Hidayatullah, Elang Cergas Pembrani, Wisnu Kurniawan, Gilang Akbar, Ridwan Pranata
Department of Informatics
Universitas Islam Indonesia, UII
Yogyakarta, Indonesia
e-mail: fathan@uii.ac.id, {14523290, 14523264, 14523091, 14523124} @students.uii.ac.id

***Abstract*—Along with the development of social media today, Twitter has become the one of the social media that is used as a provider of current information about football. Football is the most popular sport in Indonesia. People always curious about some football news update, such as match prediction, match results, transfer, rumors, etc. In this paper, we apply topic modeling to determine the topic of the tweets about football news in Bahasa Indonesia. The data used in this study were taken from several official Indonesian Twitter accounts that always update about the football and we have selected before. *Latent Dirichlet Allocation* (LDA) was used as the topic modeling method to determine what kind of topics on Twitter. According to the content analysis, we obtained several insightful topics such as pre-match analysis, live match update, football club achievements, etc. Generally, the topics posted by the Twitter account of football news provider give information about football competition in some countries such as Indonesia, England, Spain, Italia, and Germany.**

Keywords-component; topic modeling; latent dirichlet allocation; twitter; football news

I. INTRODUCTION

Social media has become the most important medium and information resource for people around the world in the last decade. There are a lot of information about the latest news or current events, that posted in every second from social media. Along with the development of social media today, Twitter has become the one of the social media that is used as a provider of current information about football.

Football is the most popular sport in Indonesia. People always curious about some football news update, such as match prediction, match results, transfer, rumors, etc. Moreover, the Indonesian people are not only curious about information updates from the national football league but also international football league, especially some top major leagues in Europe. There are four top major leagues in Europe that people are interested in getting updates, such as Premier League in England, German Bundesliga in Germany, Spain's La Liga in Spain and Serie A in Italia. Therefore, it is very important for sport news providers to share information updates from those major leagues via Twitter. For the citizens, Twitter is the one of the fastest media for getting the latest news about football.

Those huge Twitter data provide some hidden topics and important information. The topics obtained from the tweets can also illustrate and represent the trends, hot topics, etc. To get the topic from the corpus, topic modeling method can be applied. In this paper, we apply topic modeling to determine the topic of the tweets about football. The data used in this study were taken from several official Indonesian Twitter accounts that always update about the football and we have selected before. We utilize *Latent Dirichlet Allocation* (LDA) as the topic modeling method to determine what kind of topics on Twitter.

The remainder of this paper is organized into the following structure. Section 2 describes the related work. We explain our research methodology in section 3. The results and discussions are explained in section 4. Finally, section 5 describes the conclusion of our research.

II. RELATED WORK

Topic modeling has been widely applied by researchers in various fields including transportation research area [1], medical and health [2][3], bioinformatics [4], politics [5], etc. Topic modeling using Twitter data has also been conducted by some researchers before. Topic modeling of tweet data has its own challenges compared with other text data due to their unstructured language form and non-standard type of language [6]. LDA method has been applied to find topics on Twitter and there were some new approaches to improve the performance of LDA [7][8][9].

Yoon, et al [5] analyzed public opinion from Twitter about political issues in Korea by identifying the most discussed topics via LDA topic model. Yang and Rim [9] proposed new method for topic modeling called Trend Sensitive-Latent Dirichlet Allocation to extract latent topics from the contents by modeling temporal trends on Twitter over time. Lim, et al [10] proposed the Twitter-Network topic model to concurrent model the text and the social network in a fully Bayesian nonparametric way.

III. METHODOLOGY

In this section, we describe our methodology that conducted in our research.

A. Data Retrieval

Twitter provides an API that enables people to gather the tweets. This research utilizes Twitter API v1.1 and GetOldTweets-python library [1] to obtain the tweets. The advantages of this library compared to the other Twitter

[1] https://github.com/Jefferson-Henrique/GetOldTweets-python

978-1-5386-6350-9/18/$31.00 ©2018 IEEE

library is to collect data based on the time range we specify as desired, easy to use, and the results of the data are neatly arranged in csv format.

The data used in this study were retrieved taken from reliable Indonesian Twitter accounts that posted about football news. These accounts include: @bolanet, @detiksport, @goal_id, @panditfootbal, @vivabola. The total data obtained from those Twitter accounts are 120,639 tweets with a time span from 1st January 2017 to 24th December 2017. Table 1 shows the dataset of this research.

TABLE I. TWITTER DATASET

No	Twitter Account	Number of Tweets
1	@VIVAbola	31564
2	@panditfootball	15270
3	@GOAL_ID	34204
4	@detiksport	25541
5	@Bolanet	14060
Total		120639

B. Preprocessing

The preprocessing step in this research is based on the previous research about Twitter preprocessing tasks [11]. The preprocessing tasks in this research are case folding, removing HTML tags and Unicode characters, removing symbols and emoticons, removing non ASCII characters, removing special Twitter characters, removing URLs, removing punctuations, removing numbers, and removing stop words.

C. Topic Modeling Using LDA

Topic modeling is the one of the most powerful methods in text mining that aims to identify patterns and find relationship among data from a collection of text documents [12]. The most popular method in topic modeling is LDA. LDA has been proven to be an effective unsupervised learning methodology for finding different topics in text documents [13]. LDA topic modeling is an unsupervised technique in machine learning which first introduced by Blei, et al [14] as a generative probabilistic model for text corpus.

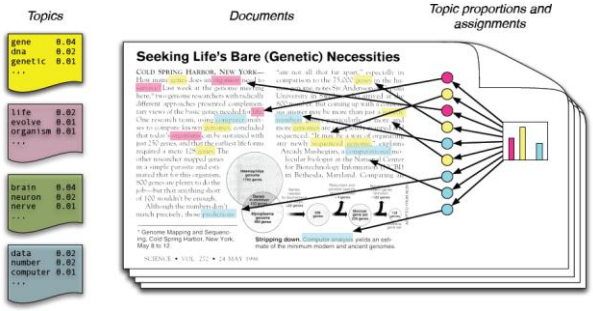

Figure 1. LDA Model [15].

The LDA model is used to find the thematic structure on a document. The purpose of LDA method is to find the topics from the collection of documents where each topic is a distribution over words or fix vocabulary, each document is a blend of corpus-wide topics, and each term is taken from one of those topics. The topic is an entity that illustrates the relationship between words as shown in Figure 1.

D. LDA Visualization

The topic model result will be visualized using Gensim and pyLDAvis library in Python. The pyLDAvis is a web-based interactive topic model visualization using LDA which is built from LDAvis using a combination of R and D3 [16]. By using pyLDAvis, we are allowed to explore the relationship between topic and terms to understand the LDA model. PyLDAvis has two panels, the distribution map of each topic and the intensity graph that represents the most frequent terms in the corpus.

IV. RESULT AND DISCUSSION

A. LDA Visualization Analysis

This section discusses the result of our experiments. To conduct the LDA topic modeling, we use LdaModel library provided by Gensim library in Python. We choose ten topics as a parameter. The inter-topic distance map visualization of our model is shown in figure 2.

Figure 2. Intertopic Distance Map Visualization.

According to figure 2, there are some topic clusters that are mutually exclusive, for instance, between topic 6, 9, and 10; topic number 4 and 6; and between topic number 1 and 8. The topic clusters that mutually exclusive indicate that the topics have similarity. On the other hand, there are other topics that can independently clustered, such as topic cluster number 2, 3, 5, and 7. Those clusters have covered specific topics that can be seen from the distance between the clusters. It also indicates that the distribution and frequency the word in the topic is very unique.

Figure 3 shows the top 30 most salient words in the corpus. We can also see that there are five terms that

frequently appeared in the corpus such as “MU”, “Madrid”, “liga”, “Chelsea” and “vs”. This indicates that Manchester United, Real Madrid and Chelsea are quite often to be discussed topic in the news. In addition, the word “liga” and “vs” indicate the topic about football matches in the league.

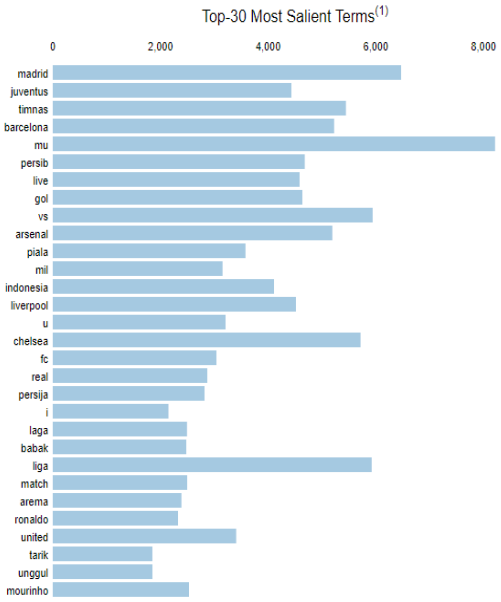

Figure 3. Top-30 Most Salient Terms.

B. *Topic Analysis*

The data visualization of each topic is illustrated using word cloud. Word cloud is a visualization composed of words in a particular text data. Word cloud displays how frequent the words appear in a collection of text. The size of each word indicates its importance which means the larger the word size, the more frequent the word in a topic. In addition, the words that dominate the word cloud is most likely directly related to the topic of the word cloud.

- Topic #0: Pre-match analysis in the Premier League

  Topic #0 talks about pre-match analysis in the English Premiere League. There are some dominated words in figure 4 that represent the topic about pre-match analysis, such as *“vs”, “jelang”, “laga”, “rekor”, ”fakta”*. The words like Chelsea, Tottenham, MU, Arsenal, premier, league illustrate about English Premier League.

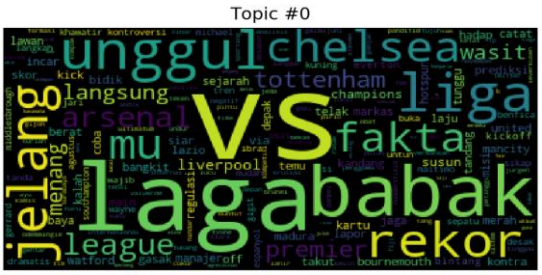

Figure 4. Word Cloud Visualization of Topic #0.

- Topic #1: Live match update in Liga 1

  Figure 5 shows the word cloud of the topic #1. We found the word live, match, and Persib as the most dominant topic. Other words like Arema, Persiba, PSM, and Bhayangkara indicate the football club in Indonesian League. From those dominant words, it can be concluded that the topic is about live matches in the Indonesian football league (Liga 1).

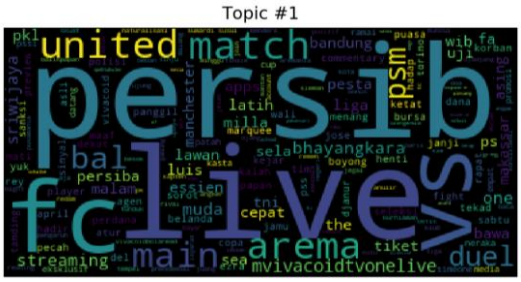

Figure 5. Word Cloud Visualization of Topic #1.

- Topic #2: English Premiere League

  The word cloud visualization of topic #2 is shown by figure 6 below. The words Chelsea, Liverpool, City, ManCity, MU are dominating in this topic segment. In addition, there are also listed some English Premier managers such as Conte, Guardiola, and Klopp. According to those words, it can be concluded that topic #2 talks about English Premier League.

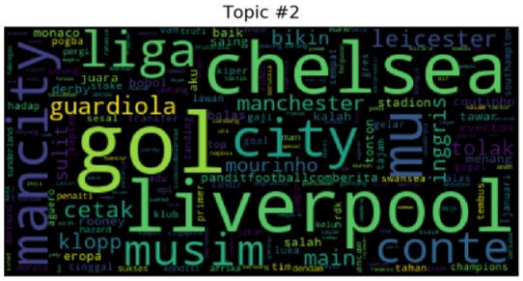

Figure 6. Word Cloud Visualization of Topic #2.

- Topic #3: Rivalry between MU and Arsenal

  The topic discussed in topic #3 is clearly about the rivalry between Manchester United and Arsenal. It can be seen from the terms Arsenal, MU, Mourinho, and Wenger. The word cloud of topic #3 can be seen in figure 7.

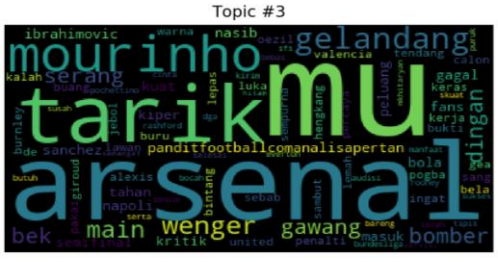

Figure 7. Word Cloud Visualization of Topic #3.

- Topic #4: Indonesian national team and Indonesian League

  According to the word cloud in figure 8, the most appropriate topic to topic #4 is about Indonesian national team and Indonesian league. The words timnas and Indonesia indicate the topic about Indonesian national team. Moreover, there are some other words such as Persipura, Persija, hasil, and klasemen that illustrate the topic concerning the Indonesian league.

Topic #4

Figure 8. Word Cloud Visualization of Topic #4.

- Topic #5: Serie A Italia

  Figure 9 shows the word cloud of topic#5. From the terms in the word cloud, it can be concluded that the corresponding topic of topic #5 is about Serie A Italia.

Topic #5

Figure 9. Word Cloud Visualization of Topic #5.

- Topic #6: World Cup

  The words in topic #6 deals with the world cup which can be seen from two dominant words in the word cloud, *"piala"* and *"dunia"*. The word cloud of topic #6 is shown by figure 10.

Topic #6

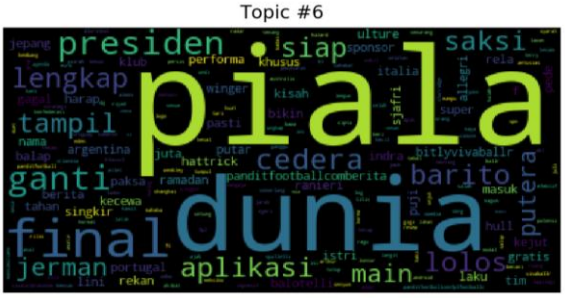

Figure 10. Word Cloud Visualization of Topic #6.

- Topic #7: El Classico (Real Madrid vs Barcelona)

  Topic #7 relates to El Classico's rivalry between Real Madrid and FC Barcelona, which can be seen from two dominant words in its word cloud, Madrid and Barcelona. Smaller words also discuss about El Classico, like the names of famous players of both two clubs, Ronaldo, Messi, and Neymar.

Topic #7

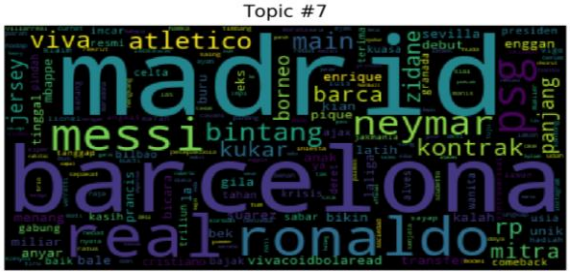

Figure 11. Word Cloud Visualization of Topic #7.

- Topic #8: Indonesia Football and Semen Padang

  Topic #8 talks about certain club in Indonesia, Semen Padang. In addition, this topic also illustrates about Indonesian football.

Topic #8

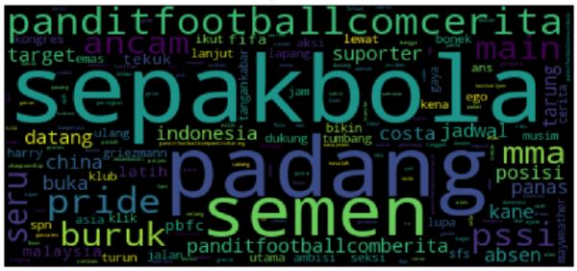

Figure 12. Word Cloud Visualization of Topic #8.

- Topic #9: The achievement of some great football clubs in Europe

  Topic #9 relates to the achievement of some great clubs in Europe. It can be seen from the dominant words in its word cloud, such as the club names like Bayern, AC (means AC Milan), Mil (possibly the origin of the word Milan affected by the stemming process), Inter, Chelsea.

Topic #9

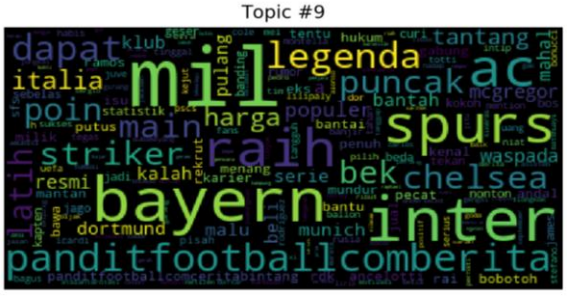

Figure 13. Word Cloud Visualization of Topic #9.

V. CONCLUSION

This paper explored the use of topical models to be applied to Twitter messages that talking about football using *Latent Dirichlet Allocation* method. According to the content analysis, we obtained several insightful topics such as pre-match analysis, live match update, football club achievements, etc. Generally, the topics posted by the Twitter account of football news provider give information about football competition in some countries such as Indonesia, England, Spain, Italia, and Germany.

REFERENCES

[1] L. Sun and Y. Yin, "Discovering themes and trends in transportation research using topic modeling," *Transp. Res. Part C*, vol. 77, pp. 49–66, 2017.

[2] X. P. Zhang, X. Z. Zhou, H. K. Huang, Q. Feng, S. B. Chen, and B. Y. Liu, "Topic model for chinese medicine diagnosis and prescription regularities analysis: Case on diabetes," *Chin. J. Integr. Med.*, vol. 17, no. 4, pp. 307–313, 2011.

[3] S. Wang, M. J. Paul, and M. Dredze, "Exploring Health Topics in Chinese Social Media : An Analysis of Sina Weibo," in *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 20–23.

[4] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *Springerplus*, vol. 5, no. 1, p. 1608, 2016.

[5] H. G. Yoon, H. Kim, C. O. Kim, and M. Song, "Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling," *J. Informetr.*, vol. 10, no. 2, pp. 634–644, 2016.

[6] A. O. Steinskog, J. F. Therkelsen, and B. Gambäck, "Twitter Topic Modeling by Tweet Aggregation," in *Proceedings of the 21st Nordic Conference of Computational Linguistics*, 2017, no. May, pp. 77–86.

[7] G. Lansley and P. A. Longley, "Computers , Environment and Urban Systems The geography of Twitter topics in London," *Comput. Environ. Urban Syst.*, vol. 58, pp. 85–96, 2016.

[8] K. Sasaki, T. Yoshikawa, and T. Furuhashi, "Online Topic Model for Twitter Considering Dynamics of User Interests and Topic Trends," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1977–1985.

[9] M. C. Yang and H. C. Rim, "Identifying interesting Twitter contents using topical analysis," *Expert Systems with Applications*, vol. 41, no. 9, Elsevier Ltd, pp. 4330–4336, 2014.

[10] K. W. Lim, C. Chen, and W. Buntine, "Twitter-Network Topic Model: A Full Bayesian Treatment for Social Network and Text Modeling," pp. 1–6, 2016.

[11] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," in *IOP Conf. Series: Journal of Physics*, 2017, vol. 801.

[12] H. Jelodar, Y. Wang, C. Yuan, and X. Feng, "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey," 2017.

[13] L. Bolelli, Ş. Ertekin, and C. L. Giles, "Topic and trend detection in text collections using latent dirichlet allocation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5478 LNCS, pp. 776–780, 2009.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[15] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[16] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 63–70.