Determining Students Preparation for College Entrance Examinations in Indonesia From Twitter Data Using Exploratory Data Analysis

Kartika Maulida Hindrayani
Data Science
UPN “Veteran” Jawa Timur
Surabaya, Indonesia
kartika.maulida.ds@upnjatim.ac.id

Tresna Maulana F
Data Science
UPN “Veteran” Jawa Timur
Surabaya, Indonesia
tresna.maulana.ds@upnjatim.ac.id

Prismahardi Aji R.
Data Science
UPN “Veteran” Jawa Timur
Surabaya, Indonesia
prismahardi.aji.ds@upnjatim.ac.id

Kartini
Informatics
UPN “Veteran” Jawa Timur
Surabaya, Indonesia
kartini.if@upnjatim.ac.id

***Abstract*— Nowadays, educational data can be learned not only for those in Education but also in Information Technology. This happened because education and technology can no longer be separated. Senior high school graduates will take College Entrance Examination to be admitted to public institutions in Indonesia. Sometimes, they share their progress, target, and complain on social media. In this research, we collected data from Twitter. We explore the data to determine student's preparation using Exploratory Data Analysis. The results are positive words in both English and Indonesia, word count, word cloud, and geographical data plot.**

Keywords—College Entrance Examination, Twitter, Exploratory Data Analysis, Educational Data

I. Introduction

College Entrance Examination or sometimes called Ujian Tulis Berbasis Komputer (UTBK) - Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN) in Indonesia will be held in April 2021. Senior high schools graduate mostly will be preparing themselves by studying hard. Hoping they will achieve their goal by being accepted in their chosen university or institution. Of course, to fulfill their target, they need to score higher than their competitors.

UTBK – SBMPTN held at the same time nationally once a year. Previously, the examinee used 2B pencils and a paper to do the test. Since 2009, the test makes use of a computer-based test to do the college entrance examination. The examinee can choose their test location based on their domicile. However, they can’t choose their test schedule because it’s been arranged by the test committee.

A fresh senior high school graduate can participate in the test for a maximum of 3 years. After that, the opportunity to join the test is no longer available. There are three test types: science and technology (saintek), social humanities (soshum), and mix of saintek and soshum (campuran). The examinee could choose a maximum of two departments.

Indonesian internet users have reached 196.7 million people in 2020. The user's number is still growing. The advantage of using internet are connected people and distance is no longer an obstacle. The economy is supported by offline and online store. E-commerce users are still growing. Nowadayas, it is at trend to use electronic money.

The age of the Examinee is around 18, they belong to a gen Z. A gen Z is inseparable from social media, such as Instagram, Twitter, or even Tik Tok. They expressed themselves by posting on social media. Sometimes, they share their life goals on social media including their UTBK - SBMPTN goal.

In this paper, we will collect Twitter data about UTBK – SBMPTN. We hope that we could catch a glimpse of the Examinee’s preparation in facing the college entrance examination.

II. Literature Review

A. Indonesia Higher Education

Education System in Indonesia is the fourth largest in the world. Behind Indonesia, there are the USA, China, and India. Indonesia also has 270.6 million people with most of the people are in productive age. Therefore the need for higher education in Indonesia is inevitable. The challenges of higher education in Indonesia are quality, skill required, affordability, and equality [1]. Improving education quality, indirectly affect human resources capabilities and wealth [2].

There are two types of higher education in Indonesia. Private institution in Indonesia around 4.500. Top institution in

Indonesia also from private institution such as Universitas Bina Nusantara, Telkom University, and Universitas Muhammadiyah Yogyakarta. Public institution in Indonesia around 122 institutions. Top higher education in Indonesia are mostly from public institution. Due to the smaller number of public institution and the lower tuition, students ultimately choose public institution.

B. UTBK - SBMPTN

UTBK – SBMPTN held since 2009. The computer-based test can make the test more effective and efficient. Location of UTBK – SBMPTN test usually held in Public Institution building. Each public institution has its UTBK Center to handle the test. The holding board of the test is the Institution of University Entrance Exams (Lembaga Tes Masuk Perguruan Tinggi). The dates of the test happened at the same time and simultaneously.

C. Previous Research

The use of Information Technology in educational data has been widely researched. Many methods can be used in educational data. The data that can be used in research also varies.

Business Intelligence (BI) can be used to understand educational data [3]. BI visualize educational data such as students, grades, research, etc. BI supports the executives to make decision making.

The behavior of students in the Learning Management System has been analyzed [4]. Using Exploratory Data Analysis and Machine Learning Approach. The pattern resulted in three clusters. The results show a low number of a student interacting with another students and student interacting with lecturers.

Maseleno et al using Mathematical Theory of Evidence in Educational Entrance Examination [6]. This approach is designed to be sustainable learning to achieve a favorable outcome. Interactive mobile application as an innovative design of this approach.

Banica et al developed a model of “Smart Universities” [7]. Determining how internet-of-things affected educational institutions. A real-time and limited area are the optimal technical solution of the conditions.

Distance learning or online learning is a challenge faced by educational institution because of COVID-19 pandemic. Habib et al identifying success factors and system's limitation of the Learning Management System [8]. The research used mix-methodology.

IT adoption in higher education has been conducted by john [9]. The functional needs such as learning class, impact, resistance, and acceptance came into focus. Their acceptance of the technology is significantly influenced by experience, compatibility, and advantage of the technology.

D. Exploratory Data Analysis

Exploratory Data Analysis (EDA) used to understand and explore the available data. The data explored are values of data, visualizations, and derived statistics [10]. The expert that conducts EDA called data analyst.

EDA helps maximize data values [11]. EDA differs from Confirmatory Data Analysis (CDA) although it is too fuzzy. Minimum of the data used in CDA.

EDA is a data driven discovery that can initiate a research [12]. EDA could provide an insight that more intuitives. EDA used in relational data, other data, or social media data.

III. Methodology

In this chapter, we discussed the methodology used in this research. The proposed systems of the methodology can be seen in Fig 1.

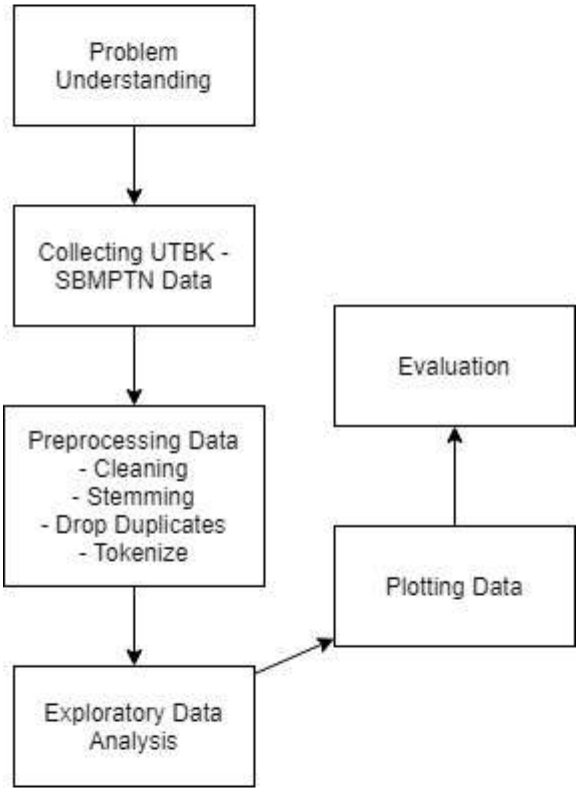


Fig. 1. Research Methodology

A. Problem Understanding

In the first step, problem understanding is needed. Literature review of higher education in Indonesia, UTBK/SBMPTN, previous research, and EDA conducted in this step. After problem understanding is done, we can go to the next step to collect the raw data needed.

B. Collecting UTBK – SBMPTN Data

UTBK-SBMPTN data collected from social media Twitter. Twitter provide API, and we have our own consumer key, consumer secret, access key, and access secret. Those key can be obtained by activating developer account. Before collecting the raw data, we need to specify the required data. Such as username, description, location, number of following, followers, total tweets, time created, retweet count, tweet text, and hashtags.

The words collected are ‘UTBK’ and ‘SBMPTN’. Because of the limitation from twitter, the data scraped are just tweets about a few days ago. Tools used in this step are Jupyter Notebook, tweepy and pandas. Pandas helps us handle dataframe. The raw data that has been collected stored in csv format file.

C. *Preprocessing Data*

Preprocessing Data in this steps contains : cleaning, stemming, drop duplicates, and tokenize. Cleaning the data from numbers, username, simbols, retweets, and hashtags. Stemming in Indonesian and English because many users use both of this language. Drop duplicates to ensure no data is redundant.

D. *Exploratory Data Analysis*

EDA or Exploratory Data Analysis conducted after preprocessing the data. After the data being cleaned, the data can be explored. Values of the data will be the focused in this step.

E. *Plotting Data*

Visualizing or plotting the data can be done in this step. Word count graphics visualize according to how many the word appears. In word cloud we can see the word according to the size and numbers of word appears. Visualizing in geographical of the locations tweeted also can be done.

F. *Evaluation*

The last step of this research is evaluation. Evaluation of the methodology and the model proposed. Conlusions and further research is discussed in this step.

IV. Results and Discussion

The word counts can be seen in Fig 2. Most of the tweets are in English. A few words in Indonesian are lolos, snmptn, terima, and semangat. Most of the Indonesian words are positive sentiments that reflect their hope to be accepted in the test.

The words in english also mostly positive sentiment. There are good, luck, fighting, hope, want, take, study, and well. No negative sentiment on the word count.

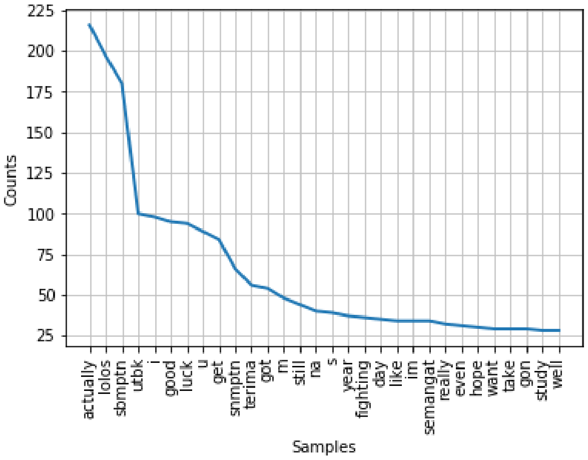


Fig. 2. Word Counts

Wordcloud of UTBK-SBMPTN can be seen in Fig. 3. The larger the size means that the word more appear in tweets. There are also the smaller size word that represents negative sentiment such as crying, forget, and rest.

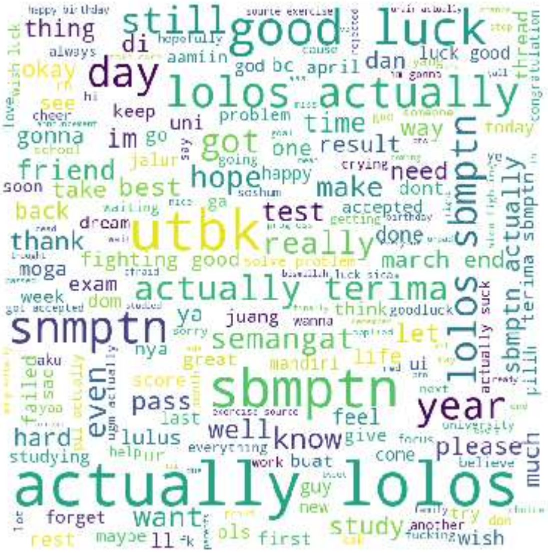
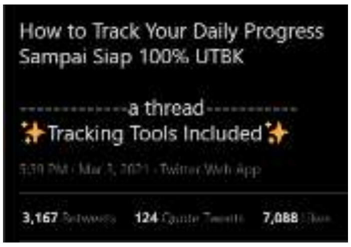


Fig. 3. UTBK-SBMPTN Wordcloud

The most retweeted thread on twitter is about students preparation to face the test as seen on Fig. 4.



How to Track Your Daily Progress
Sampai Siap 100% UTBK

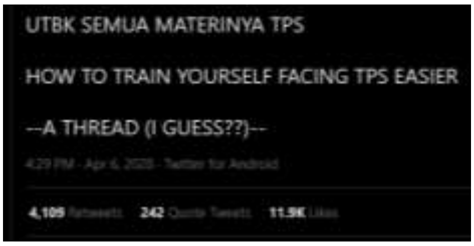
-------------a thread-----------
✨Tracking Tools Included✨

5:30 PM · Mar 3, 2021 · Twitter Web App

3,167 Retweets 124 Quote Tweets 7,088 Likes

Fig. 4. Retweeted Thread

Rumors about UTBK-SBMPTN shared on twitter is also popular to retweet. Many users retweed this rumors as seen on Fig. 5.



UTBK SEMUA MATERINYA TPS

HOW TO TRAIN YOURSELF FACING TPS EASIER

--A THREAD (I GUESS??)--

4:29 PM · Apr 6, 2020 · Twitter for Android

4,109 Retweets **242** Quote Tweets **11.9K** Likes

Fig. 5. UTBK Rumors

And the students need a good laugh about their struggle as seen on Fig. 6 about quota in UTBK.

sch! utbk struggle

[redacted]

Kak Kalo Kursi UTBK Sudah Habis
Apa Boleh Bawa dari Rumah??

Fig. 6. UTBK Rumors

Users that tweets about UTBK-SBMPTN with the largest number of followers can be seen in Table I. The users who has over a hundred thousand followers are schfess and EdukaSystem. They tweet about education in general.

Table I. Popular Users

No	*Users*	
	Username	***Followers***
1.	schfess	226063
2.	EdukaSystem	137202
3.	munconvo	64388
4.	ButetMengajar	41053
5.	Quipper_ID	32038
6.	kenshafpf	20751
7.	cuvtein	18197
8.	kampusinfo	14578
9.	k_luv20_06	14496
10.	studyyhard	14217
11.	beepbeepbio	11882
12.	beepbeepbio	11880
13.	lightaes	10538
14.	yaelahcae	10143

Tweet Locations Participations on Indonesia map can be seen in Fig. 7. From the map, we can see that the tweet location from the users are mostly from Java Island and almost none from Kalimantan.

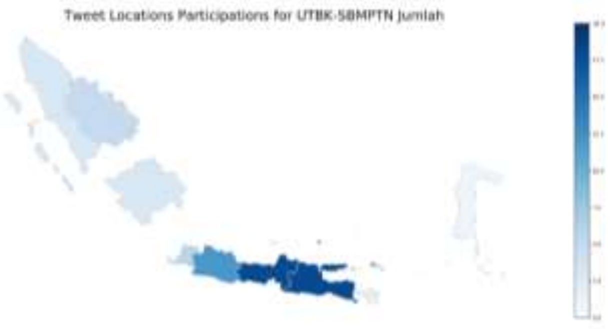


Fig. 7. Tweet Locations

V. CONLUSIONS

Exploratory Data Analysis (EDA) is a suitable method to explore the scraped data from twitter. The data is utilized to determine the preparation of the students facing UTBK-SBMPTN test. Positive words can be seen in word count and word cloud. However, the tweet locations need more data so the geographical data could be précised.

ACKNOWLEDGMENT *(Heading 5)*

REFERENCES

[1] C. Logli, “Higher Education in Indonesia: Contemporary Challenges in Governance, Access, and Quality,” *Palgrave Handb. Asia Pacific High. Educ.*, pp. 1–691, 2016.

[2] A. S. Ristapawa Indra, Martin Kustati, “Evaluation of National Examination (UN) and National-Based School Examination (USBN) in Indonesia,” *Eur. J. Educ. Res.*, vol. 9, no. 3, pp. 1063–1074, 2018.

[3] K. M. Hindrayani, “Business Intelligence For Educational Institution : A Literature Review,” no. September, pp. 22–25, 2020.

[4] T. Purwoningsih, H. B. Santoso, and Z. A. Hasibuan, “Online Learners’ Behaviors Detection Using Exploratory Data Analysis and Machine Learning Approach,” *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, no. Icic, pp. 1–8, 2019.

[5] H. M. Kartika and M. Ahmad, “Self Adaptive and Simulated Annealing Hyper-Heuristics Approach for Post-Enrollment Course Timetabling,” *J. Phys. Conf. Ser.*, vol. 1577, no. 1, 2020.

[6] A. Maseleno *et al.*, “Combining the previous measure of evidence to educational entrance examination,” *J. Artif. Intell.*, vol. 10, no. 3, pp. 85–90, 2017.

[7] L. Banica, E. Burtescu, and F. Enescu, “the Impact of Internet-of-Things in Higher Education,” *Sci. Bull. Econ. Sci.*, vol. 16, no. 1,

pp. 53–59, 2017.

[8] M. N. Habib, W. Jamal, U. Khalil, and Z. Khan, “Transforming universities in interactive digital platform: case of city university of science and information technology,” *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 517–541, 2021.

[9] S. P. John, “The integration of information technology in higher education: A study of faculty’s attitude towards IT adoption in the teaching process,” *Contaduria y Adm.*, vol. 60, pp. 230–252, 2015.

[10] K. Wongsuphasawat, Y. Liu, and J. Heer, “Goals, process, and challenges of exploratory data analysis: An interview study,” *arXiv*, 2019.

[11] A. T. Jebb, S. Parrigon, and S. E. Woo, “Exploratory data analysis as a foundation of inductive research,” *Hum. Resour. Manag. Rev.*, vol. 27, no. 2, pp. 265–276, 2017.

[12] T. Lynn, P. Rosati, B. Nair, and C. M. an Bhaird, “An exploratory data analysis of the #crowdfunding network on Twitter,” *J. Open Innov. Technol. Mark. Complex.*, vol. 6, no. 3, 2020.