Technical Report

The Seventh Research Dive on Artificial Intelligence and Machine Learning for Estimating Poverty

September 2018

Artificial Intelligence

KNOWLEDGE SECTOR INITIATIVE

Australian Government

PULSE LAB JAKARTA

Executive Summary

The Government of Indonesia has made significant progress in reducing poverty over the past few years, recording its lowest poverty rate of ten per cent in 2017 measured by income. Many citizens still remain vulnerable given their marginal position above the national poverty line. But how governments go about estimating poverty, in order to better target programmes, has never been an easy task. Today, technological advancements are enabling researchers to use new and efficient methods to learn more about people's quality of life. In particular, with more and more big data sources emerging, researchers are seeing the benefits of big data analytics for reducing poverty and improving citizens' well-being.

From 15-18 July 2018, Pulse Lab Jakarta research dive brought together academics, public officials and researchers to dive into a few big data sets to develop new methods and insights on burning policy questions around poverty reduction. An underlying goal of this Research Dive was to support the Indonesian Government's development agenda using artificial intelligence and machine learning, specifically efforts geared towards achieving Sustainable Development Goal number one on zero poverty. There were four teams and each was assigned a different dataset with a unique research focus: (1) Measuring Vulnerability to Poverty Using Satellite Imagery, (2) Estimating City-level Poverty Rates Based on E-commerce Data, (3) Using Twitter Data to Estimate District-Level Poverty in Greater Jakarta, and (4) Exploring the Connection Between Social Media Activities and Poverty.

This report outlines the research findings from the research sprint and is structured as follows:

1. The first paper describes the data sets that were assigned to the participants.
2. The second paper explores satellite data as a means to measure vulnerability to poverty. The team analysed nighttime light imageries from satellite over Yogyakarta.
3. The third paper looks on estimating city-level poverty rates in Java island by examining 2016 e-commerce data from 118 cities. The group also tested the accuracy of using e-commerce data to estimate poverty, by comparing the results with official government data of poverty levels.
4. The fourth paper discusses how poverty may be estimated at the district level using social media content and user profiles. The team used natural language processing to conduct content analysis of extracted public tweets that contained food and poverty sensitive keywords.
5. The last paper explores the relationship between social media activities and poverty (based on survey and census data at the village and individual level for the Greater Jakarta area).

Pulse Lab Jakarta is grateful for the cooperation of Ministry of National Development Planning, SMERU, Humanitarian Data Exchange, The National Team for the Acceleration of Poverty Reduction (TNP2K), Directorate of Central Data and Information Ministry of Social Affairs, OLX Indonesia, The National Institute of Aeronautics and Space (LAPAN), Universitas Padjadjaran, Universitas Gadjah Mada, Universitas Muhammadiyah Gorontalo, Universitas Udayana, World Food Programme, Institut Teknologi Sepuluh Nopember, National Statistics Agency (BPS), Telkom University, Bina Nusantara University, STMIK Akakom Yogyakarta, Pertamina University, and Sam Ratulangi University. Pulse Lab Jakarta is grateful for the support from Knowledge Sector Initiative (KSI), the Artificial Intelligence Journal and the Department of Foreign Affairs and Trade (DFAT) Australia.

Advisor Note

Estimating Poverty with New Alternative Methods

Any sudden, unfortunate event, such as a flood or a death in the family, could disrupt family finances. For example, a change in weather may mean no income for construction workers in urban areas or damaged crops for rural farmers. Official poverty-related figures though are only reported March and September every year, therefore to understand poverty throughout Indonesian districts we may need to wait even much longer. Considering this, it is necessary for us to find alternative estimates of poverty so that assistance and programmes to help poor people can be more timely and effective.

One way of doing so is by way utilizing new sources of data that have higher frequency such as from social media, online transactions or other non-conventional data sources. These sources of data, often called 'big' data, can be analysed and filtered into models based on artificial intelligence (AI), to produce sophisticated tools for more frequent and better monitoring. Pulse Lab Jakarta took on the challenge through their Research Dive initiative, which assembles a selected group of Indonesians from various disciplines to come up with new ideas of how to combine big data and AI to improve poverty monitoring, as well as to produce relevant information to help Indonesia's poverty alleviation agenda. I was very proud to be part of this initiative.

Prof. Arief Anshory
Economics Advisor

He currently serves as a Professor of Economics in the Department of Economics at the Faculty of Economic and Business, Universitas Padjadjaran and is the founder of the Centre for Sustainable Development Goals Studies (SDGs Center). He also works as the Director of Economy and Environment Institute (EEI) Indonesia. Most of his research focuses on the economics of the environment and natural resource management as well as economic development, particularly related to poverty and inequality. Prof Arief received his PhD in Economics from Australian National University. He holds a master's degree in Environmental and Resource Economics from the University College London and bachelor's degree in Economics from Universitas Padjadjaran.

Big Data for Poverty Analysis

Understanding poverty issues is essential for human development and to effectively do so, more collaboration is needed from various experts and citizens. Looking on how PLJ's Research Dive is organised, there are several opportunities to meet researchers from different studies and join forces to answer burning policy questions through the use of big data. I hope the results can contribute to the policy making decision in the country, and help to improve government institutions that are responsible for monitoring and evaluating poverty reduction programmes. Currently in Indonesia not many researchers conduct poverty analysis with big data, so this is a step in the right direction for the country.

Prof. Dedi Rosadi
Statistics Advisor

Dedi Rosadi is a statistics professor from Universitas Gadjah Mada. He completed his PhD at Vienna University of Technology, and before that his master's degree at the University of Twente. His research interests include time series, statistical computing and mathematical finance. He has also written books related to statistics and econometrics, for instance *Econometrics and Time Series Analysing using R, Time Series Analysis and Introduction to Statistical Modeling using R.*

Advisor Note

An Effective Two-way Communication Between Advisors and Participants

This is my second time participating in PLJ's Research Dive for Development. This time around was even more amazing and eye-opening, partly because it was not directly connected with my areas of work. The deep-diving analysis and discussions around "artificial intelligence and machine learning for estimating poverty" gave me a lot of insights that can be transferred to my own work. For me, the Research Dive atmosphere and set up allowed an effective two-way communication and learning experience between advisors and participants that is also very enriching. I hope to see more research dives, more practitioners involvement as well as continuation and improvement of work from research dive alumni. Thank you Pulse lab Jakarta for allowing me to be a part of this interesting event.

Faizal Thamrin
Remote Sensing Advisor

Faizal Thamrin works for DMInnovation as a Disaster Management Specialist. He has also been a Data Manager at Humanitarian Data Exchange, focused on strengthening data collaboration with humanitarian partners, governments and academia. Previously, he spent around 10 years working with United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA) in Bangladesh, Indonesia, Philippines, and Pakistan as a Geographic Information Officer and Information Management Officer.

Research Dive

Advisors

Prof. Arief Anshory Yusuf	Universitas Padjadjaran
Faizal Thamrin	DM Innovation
Prof. Dedi Rosadi	Universitas Gadjah Mada

Researchers

Group 1 – Estimating Poverty at the Provincial Level with Satellite Data

Benny Istanto	World Food Programme
I Wayan Gede Astawa Karang	Universitas Udayana
Nursida Arif	Universitas Muhammadiyah Gorontalo
Pamungkas Jutta Prahara	Pulse Lab Jakarta

Group 2 - Estimating Poverty at the City Level with E-Commerce Data

Ana Uluwiyah	Central Statistics Agency
Dedy Rahman Wijaya	Telkom University
Dwi Rani Puspa Artha	LPEM UI
Ni Luh Putu Satyaning Pradnya Paramita	Institut Teknologi Sepuluh Nopember
Anissa Zahara	Pulse Lab Jakarta
Muhammad Rheza	Pulse Lab Jakarta

Group 3 - Estimating Poverty at the District Level with Social Media Data

Lili Ayu Wulandhari	Bina Nusantara University
Sri Redjeki	STMIK AKAKOM Yogyakarta
Widaryatmo	Central Statistics Agency
Yunita Sari	Universitas Gadjah Mada
Muhammad Rizal Khaefi	Pulse Lab Jakarta

Group 4 - Estimating Poverty at the Household Level with Social Media Data and Household Survey Results

Eka Puspitawati	Pertamina University
Eko Fadilah	TNP2K
Hizkia H. D. Tasik	Sam Ratulangi University
Nurlatifah	Central Statistics Agency
Rajius Idzalika	Pulse Lab Jakarta

Table of Contents

Research Dive Artificial Intelligence and Machine Learning for Estimating Poverty

Zakiya Pramestri
Pulse Lab Jakarta
Jakarta, Indonesia
zakiya.pramestri@un.or.id

Dikara Alkarisya
Pulse Lab Jakarta
Jakarta, Indonesia
dikara.alkarisya@un.or.id

Lia Purnamasari
Pulse Lab Jakarta
Jakarta, Indonesia
lia.purnamasari@un.or.id

ABSTRACT

When studying poverty, researchers tend to rely on data related to economic livelihoods, especially to analyse poverty trends and patterns throughout developing countries. Findings and insights from such analyses are then typically used to design poverty reduction programmes and formulate public policies. However, collecting data on economic livelihoods is often a tall order for researchers and governments, in part due to the high costs associated with it. To address this issue, Pulse Lab Jakarta chose to explore big data - new and diverse digital data sources - that come with the benefits of increased accuracy, timeliness and cost-efficiency. In this paper, several research methodologies are discussed, which demonstrate the ability (and advantages) of using these data sources to estimate consumption expenditure, wealth, and other poverty-related indicators.

To support the Indonesian Government's agenda on reducing poverty, Pulse Lab Jakarta organised its 7th Research Dive for Development, focused on exploring artificial intelligence and machine learning approaches for Estimating Poverty. Among the invited participants were researchers, public officials and academics from across Indonesia who mashed up various big datasets to develop methods and insights on burning policy questions. The datasets included satellite imagery, e-commerce data, social media data, and household survey results.

1 INTRODUCTION

Poverty continues to be one of the most pressing issues on the global development agenda. And with global efforts geared towards achieving Sustainable Development Goal number one on zero poverty, governments continue to invest huge amount of resources to reach that goal. Yet, the growing inequality also adds another dimension to the issue. Poverty brings consequences. Inversely, reducing poverty can improve other crucial aspects, such as public health, economic development and social cohesion. Even though the number of people living below the global poverty line has gradually decreased over the last two decades [1], governments around the world are continuing their efforts to make effective policies to fully eliminate poverty. Because the causes and effects of poverty differ according to contexts, there is also the need to produce more customized policy to address each unique situation.

Poverty reduction has been a major initiative in Indonesia since the country's economic crises in 1997 and 1998. For the central government, poverty reduction has been the main focus of each national medium development plan (RPJMN) [2]. Since the decentralisation era, local governments have also been contributing significantly to poverty reduction through regional poverty reduction strategies (SPKD) [3]. In addition, many local and international development stakeholders have been contributing to poverty alleviation in Indonesia. As a result, Indonesia has made enormous progress by cutting the poverty rate to more than half since 1999, reaching a low 10.9% in 2016 [4].

To measure poverty, household income and consumption data are critical for researchers and policymakers in order to design effective and inclusive public policies. However, obtaining such data through traditional methods such as surveys can be costly. For instance, some developing countries have been conducting similar surveys for decades, yet still failed to produce timely data [5]. With the help of rising technologies, researchers across various disciplines have come up with new approaches and techniques to explore and utilise new datasets generated by modern technologies [6] – commonly referred to as "big data". While these new datasets have indeed shown potential to estimate poverty, they are not intended to replace conventional datasets. Furthermore, poverty estimation can be more useful and accurate if researchers can combine big data with traditional data sources, such as household survey data.

One of the Research Dive's underlying goals is to support the Government's development agenda; in particular, efforts geared towards achieving Sustainable Development Goal number one on zero poverty. Pulse Lab Jakarta invited several participating researchers from universities, government institutions, international organizations, and think tanks across Indonesia. The various disciplines represented included computer science, economics, geographic information system (GIS), and statistics. The researchers collaborated with Pulse Lab Jakarta to develop methods and insights on burning policy questions, specifically: (i) estimating poverty at the provincial level with satellite data, (ii) estimating poverty at the city level with e-commerce data, (iii) estimating poverty at the district level with social media data, and (iv) estimating poverty at the household level with social media data and household survey results.

[1] https://www.brookings.edu/blog/future-development/2017/11/07/global-poverty-is-declining-but-not-fast-enough/

[2] A. Suryahadi et al. 2010. Review of the Government's Poverty Reduction Strategies, Policies, and Programs in Indonesia. The SMERU Research Institute. Jakarta

[3] National Team for The Acceleration of Poverty Reduction. 2014. Reaching Indonesia's Poor and Vulnerable and Reducing Inequality : Improving Programme Targeting, Design, and Process

[4] http://www.worldbank.org/en/country/indonesia/overview

[5] http://www.jblumenstock.com/files/papers/jblumenstock$_2$016$_s$cience.pdf

[6] http://www.jblumenstock.com/files/papers/jblumenstock$_2$016$_s$cience.pdf

2 DATASETS

In this section, we explain briefly about the types of data used by the participants during the Research Dive.

2.1 Satellite Imagery Data

2.1.1 Indonesian National Institute of Aeronautics and Space (LAPAN). In partnership with LAPAN, PLJ provided nighttime satellite data from 2015. The data comes from Visible Infrared Imaging Radiometer Suite (VIIRS) on Suomi National Polar-orbiting Partnership (NPP) and is in .tif format.

2.1.2 Imagery of Asia - Australia. The imagery data is from 2012 and 2016 and can be accessed through NASA website. It comes from Visible Infrared Imaging Radiometer Suite (VIIRS) on Suomi NPP Satellite.

2.1.3 World Imagery. PLJ provided the world imagery data 2010-2013 from Operational Line Scan (OLS) imaging systems on Defense Meteorological Satellite Program (DMSP) spacecraft version 4.

2.2 E-Commerce Data

In partnership with OLX Indonesia, PLJ provided e-commerce data of Java Island from 2016. Specifically, e-commerce data was provided on property, car and motorcycle. Under property, we focused on house, apartment and land area. We aggregated the data to add information by calculating the average price, average sold price, standard deviation of price and trimmed price (5% percentile of the lowest and the highest), total viewer (number of people who viewed advertisements) and total contact (number of people who tried to contact sellers).

2.3 Social Media Data

PLJ shared Twitter data from 2014 with the participants. The data provided are on 30 minutes interval, and are geo-tagged Twitter data. The data provided are all aggregated at the district level.

2.4 Household Survey Data

PLJ provided data of Pemutakhiran Basis Data Terpadu (PBDT) in 2015 from the National Team for Alleviating Poverty (TNP2K), including household and individual data. The shared data are anonymised. PLJ also provided the BDT from Ministry of Social Affairs from 2015, 2017, 2018, such as aggregated data at the village level. Additionally, Indonesian Family Life Survey year 2015 from Surveymeter also shared. Data from Indonesian Family Life Survey, which took the sample of about 83% of the Indonesian population, contains more than 30.000 individuals living in 13 of the 27 provinces in the country. The survey consisted of two questionnaires: household and community or facility. The household survey includes data on consumption, household characteristics, education, employment, and health. The community or facility data includes village facility, access to social welfare program, health facility, education facility, and facility for economy activity.

3 DATA AND TASK MAPPING

We defined four research questions with a different dataset for each question. Participants were allowed to use other data sources that were not provided by Pulse Lab Jakarta to support and elaborate on their proposed solutions. The satellite imagery data was given to team one to estimate poverty at the provincial level. The e-commerce data was shared with the second team to estimate poverty at the city level. The third group was assigned the social media data. In this Research Dive, Pulse Lab Jakarta shared Twitter Data from 2014 to estimate poverty at the district level. Team four estimated poverty at the household level and was given access to Twitter data from 2014 and household survey results.

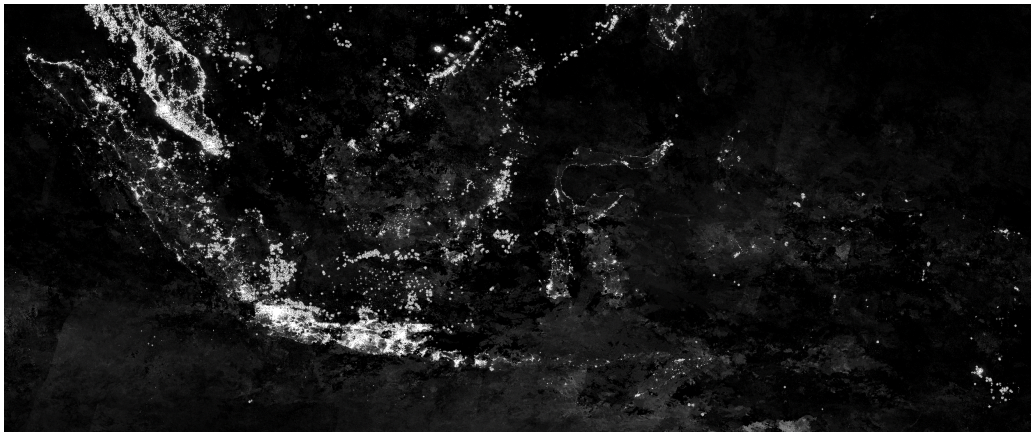

Figure 1: Nighttime Satellite Data

Table 1: Example of tweets per-district

Column name	Column description	Example of value
cat_l3_id	Category level 3 id	4695
cat_l3_name	Category level 3 name	Mobil CBU
cat_l2_id	Category level 2 id	198
cat_l2_name	Category level 2 name	Mobil Bekas
cat_l1_id	Category level 1 id	86
cat_l1_name	Category level 1 name	Mobil
cat_group	Category type (Inner / Outer) - Inner category: selain jasa dan property	Inner

Table 2: Example of tweets per-district

Column Name	Type	Sample	Description
TIMESTAMP	Datetime (yyyy-mm-dd HH:MM:SS)	2014-01-01 05:00:00	Timestamp of tweet post activity
PROV	int	51	Code of province
KAB	int	5102	Code of region
KEC	int	5102011	Code of district
PROV_NAME	string	Jawa Barat	Name of province
KAB_NAME	string	BEKASI	Name of region
KEC_NAME	string	BEKASI BARAT	Name of dsitrict
LAT	float	106.47743	Coordinate for latitude
LONG	float	-6.23672	Coordinate for longitude
GENDER	string	female	Gender
SOURCE	string	Others	Source of tweet (web/ others)
CONTENT	string		Tweet posted by others

Table 3: Example of Household Survey Data

Variable data	Coloumn
Jumlah keluarga	jml_keluarga
Jumlah anggota rumah tangga	jml_art
Keterangan perumahan	
Milik sendiri	sta_bangunan1
Milik orang lain	sta_bangunan2
Luas bangunan	luas_bangunan (rata-rata)
Jenis lantai terluas	
Marmer/granit	jns_lantai1
Kepemilikan aset dan keikutsertaan program	
Rumah tangga dan aset bergerak	
Tabung gas 5,5 kg atau lebih	gas_55
Kepemilikan aset tidak bergerak	
Lahan	lahan (luas lahan = rata-rata)
Rumah di tempat lain	rumah_lain
Kepemilikan kartu pintar, sejahtera, dan lainnya	
kartu keluarga sejahtera (KKS)/ kartu perlindungan sosial (KPS)	kartu_bansos1
kartu indonesia pintar (KIP)/ bantuan siswa miskin (BSM)	kartu_bansos2

Estimating Poverty at the Provincial Level with Satellite Data

Nursida Arif
Universitas Muhammadiyah Gorontalo
Gorontalo, Indonesia
nursida.arif@gmail.com

Pamungkas Jutta Prahara
Pulse Lab Jakarta
Jakarta, Indonesia
pamungkas.prahara@un.or.id

ABSTRACT

Research suggest that satellite data could be used as a proxy for a different parameters, including urbanisation, density, and economic growth. One of the parameters that could be extracted from satellite images is land use. In the research, land use classification has been done using Landsat 7 image and topographic map with visual interpretation method. The research took Yogyakarta Province for study case for its diverse historical poverty line and regions. In Kulon Progo, the region with high poverty and vulnerability is generally an area with a small population and physically land, including disaster prone areas.

KEYWORDS

satellite imagery, land use, poverty line, visual interpretation, vulnerability, poverty indicators, proxy indicators

1 INTRODUCTION

Satellite imagery is increasingly available for free at certain resolution for global scale and contains a lot of information at pixel level that could be associated with economic activity. Many research suggest that satellite data can be used as a proxy for a number of variables, including urbanization, density, and economic growth. The environment can be a parameter of poverty but the complex nature of the population making it difficult to measure its influence. Spatially the effect of environment on poverty can be done with remote sensing approach. Although there is no evidence that satellite prediction is better in poverty estimation than in conventional censuses but this approach can be used as support.

One of the parameters that can be extracted from satellite images is land use. Some previous researchers used the land-use approach as an indicator of poverty [3] [5]. An increase in population will trigger land clearing for various interests such as settlement, farmland or industry. [6] describes one perspective of land use as a functional space intended to accommodate diverse uses. In this perspective the land accommodates the growth of the area driven by population growth and economic expansion. Rural areas have different characteristics with urban areas. According to Law No. 26 of 2007 and Minister of Public Works Regulation No. 41 of 2007, rural areas are areas that have major agricultural activities including natural resource management with the arrangement of regional functions as a place of rural settlements, government services, social services and economic activities.

In contrast to urban areas dominated by non-agricultural activities. [4] in more detail defines the pattern of cities can be seen from the existence of built areas such as settlements and infrastructure, while the pattern of villages is dominated by agricultural land, forest land and settlement patterns are small and not centralized. Rural land is mostly used for mining and agrarian activities, such as agriculture, plantation, animal husbandry and fisheries. In accordance with the characteristics of its activities, land use in rural areas tends to use large land units with low intensity of use, which means it tends not to be built land. Therefore, in this research land use classification is explained into two (2) i.e built up area and not built up as an indicator of vulnerability to poverty.

2 METHODS

Land use classification has been done using Landsat 7 image and topographic map with visual interpretation method. Land use is explained into five (5) i.e. forest/garden, rice field, moor/field, open field, settlement. Each class is grouped into two classes, i.e., the land is built and non-built as shown in figure 1.

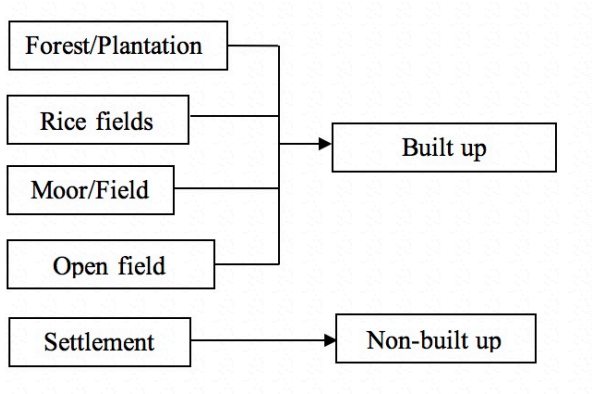

Figure 1: Land Use Class Groups

Yogyakarta Province is choosen for the case study in the research. The rationale are because of the historical population living below poverty line has diverse conditions (low to high) and regions (urban, rural, coastal).

3 DISCUSSION

Geographically, Yogyakarta is a region complete with various forms of land. In the northern region is the form of volcanic soil, the eastern and western regions are the plateau, the middle and the south are the lowlands. Different physiographic conditions have an impact on population distribution and economic progress. So that the level of poverty will be more vulnerable in areas with hilly topography and mountains. This is evidenced by the land use shown in figure 2

Figure 2 shows the residential area of Sleman Regency is dominantly distributed in the southern Sleman region bordering Yogyakarta and Central Sleman. In Kulonprogo District dominant settlements in the south and east. Bantul District, the dominant

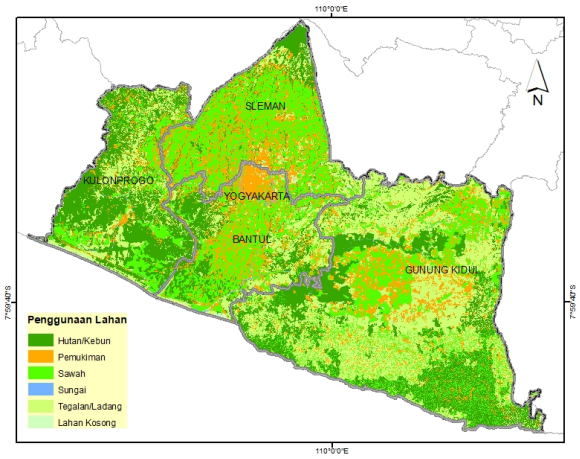

Figure 2: Yogyakarta Land Use Map

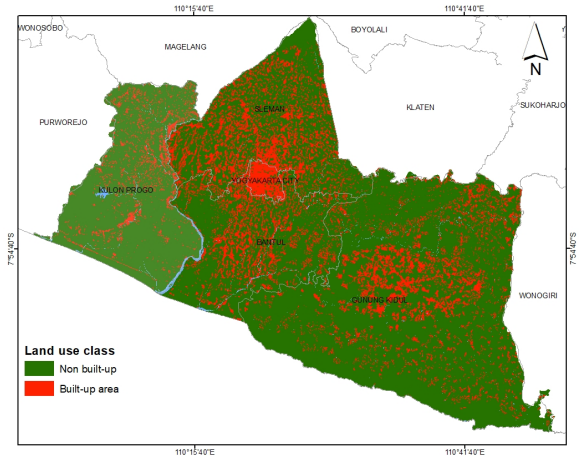

Figure 3: Land Use Maps with Built Up Area

settlement in the center is similar to the pattern in Gunung Kidul Regency. Percentage of settlements or built areas compared to the area of each based on the image used is presented in table 1.

Table 1: Area of each region vs. the area built

	Area (Km^2)	Built up area (%)
Sleman	573.9	27.83
Kulonprogo	581.73	9.79
Yogyakarta City	32.97	84.6
Gunung Kidul	1,476	13.01
Bantul	514.3	21.6

Analysis, 2018

Table 1 shows the city of Yogyakarta is an area with a percentage of settlements / built areas greater than other regions because of its location as the capital of the Province to become the center of the economy. While the area with the percentage of smaller built up area that is Kulonprogo (9.7%) and Gunung Kidul (13%). The results of these analyzes indicate that Kulonprogo and Gunung Kidul are vulnerable to poverty compared to other regions. Based on this, further analysis will be specified to Kulon Progo Regency to test the correlation of image interpretation with the condition of the population. Land use maps with built-up areas are shown in figure 3.

Table 2 shows that 12 sub-districts in Kulonprogo regency which have the smallest percentage of built up area that is Lendah, Panjatan, Kokap, Galur, Temon and Sentolo. While the region with the most populations of Pengasih, Sentolo and Wates. Graphs of population and built-up areas are shown in Table 2.

Based on Figure 5, the results of the analysis show that the areas that are vulnerable to poverty are the sub-districts of Galur, Kokap and Panjatan where the number of areas is slightly built and the number of small populations compared to other sub-districts.While District Pengasih, Wates and Sentolo have a smaller degree of vulnerability. Kokap as one of the districts that are vulnerable to

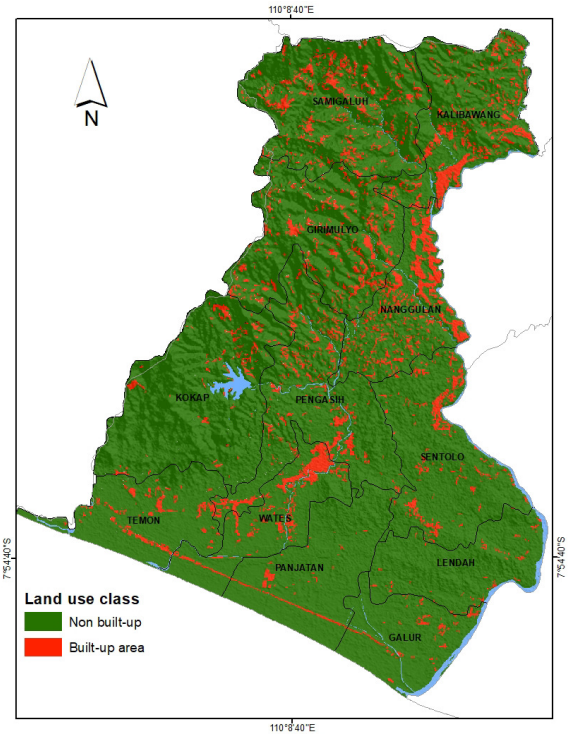

Figure 4: Land use map with built area in Kulonprogo

poverty is supported by the physical condition of the land that is vulnerable to disaster. [2] assessed the landslide in some areas of Kokap Sub-district explained that the actual landslide process that

Table 2: Area of each region vs. built area

	Area (Km^2)	Built up area (Km^2)	(%)	Population
Girimulyo	58.31	7.67	13.16	12,417
Temon	38.86	2.50	6.43	13,452
Galur	31.51	1.26	4.00	16,182
Samigaluh	65.76	8.36	12.72	16,729
Kalibawang	48.87	8.33	17.05	17,277
Nanggulan	38.86	8.71	22.42	17,360
Panjatan	43.95	1.22	2.77	19,020
Kokap	72.42	2.60	3.59	19,430
Wates	32.43	4.88	15.04	19,582
Lendah	36.94	0.77	2.09	22,544
Pengasih	60.55	6.07	10.02	25,528
Sentolo	53.27	4.20	7.89	27,812

Analysis, 2018

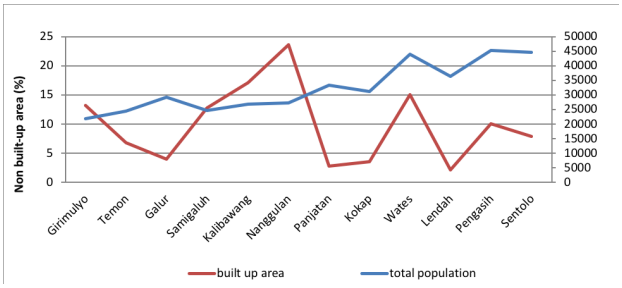

Figure 5: Graphs of population and built-up areas in Kulonprogo

occurs in the hills denudasional. The same is mentioned by Panjatan and Kokap sub-districts included in the class of very serious erosion susceptibility [1].

4 CONCLUSIONS

The results of land use analysis, poverty-prone areas in the Province of Yogyakarta Special Region of Kulon Progo and Gunung Kidul since the number of built areas smaller than other regions. In Kulon Progo, the region with high poverty and vulnerability is generally an area with a small population and physically land, including disaster prone areas such as Kecamatan Kokap.

5 LIMITATION AND RECOMMENDATION

This study uses medium-resolution imagery with analysis at the Provincial level so that land use classification can not be done in more detail. Poverty is very complex if only done based on land use. An assessment of poverty estimates using satellite imagery should be used for smaller areas at sub-district and village levels with high-resolution imagery. So a more detailed analysis can show the type of roof, the type of settlement patterns, access roads and other attributes that support the identification of poverty levels. Future research can be estimated poverty through the integration of remote sensing images with other parameters such as hazard vulnerability and other social factors such as income and nutrient intake per household.

REFERENCES

[1] Nursida Arif, Projo Danoedoro, and Hartono Hartono. 2017. Remote Sensing and GIS Approaches to A Qualitative Assessment of Soul Erosion Risk in Serang Watershed, Kulonprogo, Indonesia. *Geoplanning: Journal of Geomatics and Planning* 4, 2 (2017), 131–142.

[2] Suprapto Dibyosaputro. 1992. *Longsorlahan di Daerah Kecamatan Samigaluh, Kabupaten Kulon Progo, Daerah Istimewa Yogyakarta.* Technical Report. Yogyakarta.

[3] Gilvan R Guedes, Leah K VanWey, James R Hull, Mariangela Antigo, and Alisson F Barbieri. 2014. Poverty dynamics, ecological endowments, and land use among smallholders in the Brazilian Amazon. *Social science research* 43 (2014), 74–91.

[4] Fransiscus Xaferius Herwirawan, Cecep Kusmana, Endang Suhendang, and Widiatmaka Widiatmaka. 2017. Changes in Land Use/Land Cover Patterns in Indonesia's Border and their Relation to Population and Poverty. *Jurnal Manajemen Hutan Tropika* 23, 2 (2017), 90–101.

[5] Eddie CM Hui, Jiawei Zhong, and Kahung Yu. 2016. Land use, housing preferences and income poverty: In the context of a fast rising market. *Land Use Policy* 58 (2016), 289–301.

[6] Edward J Kaiser, David R Godschalk, and F Stuart Chapin. 1995. *Urban land use planning*. Vol. 4. University of Illinois Press Urbana, IL.

Estimating City-Level Poverty Rate based on e-Commerce Data with Machine Learning

Dedy Rahman Wijaya
Telkom University and Institut Teknologi Sepuluh Nopember
dedyrw@tass.telkomuniversity.ac.id

Ni Luh Putu Satyaning Pradnya Paramita
Institut Teknologi Sepuluh Nopember
pradnya@statistika.its.ac.id

Ana Uluwiyah
Education and Training Center, Statistics Indonesia
auluwiyah@bps.go.id

Dwi Rani Puspita
Institute for Economic and Social Research, Universitas Indonesia
dwirani.puspa@ui.ac.id

Muhammad Rheza
Pulse Lab Jakarta
Jakarta Pusat, Jakarta
muhammad.rheza@un.or.id

Annisa Zahara
Pulse Lab Jakarta
Jakarta Pusat, Jakarta
anissa.zahara@un.or.id

ABSTRACT

Indonesia abundantly produces big data from various resources, e.g. social media, financial transaction, transportation, call detail records, e-commerce. These types of data have been considered as potential resources to complement periodic survey, even census, to monitor development indicators in which poverty rate is included. This research aims to estimate poverty rate at city-level based on e-commerce data using machine learning methods i.e. Support Vector Regression (SVR) and Artificial Neural Network (ANN). Feature selection has been performed with Fast Correlation-Based Filter (FCBF). The result shows that ANN-based model predicts the city-level poverty rate very well, with high accuracy, low error and low bias. This research suggests that e-commerce is potential to be used as proxy for city-level poverty rate.

KEYWORDS

TBC

1 INTRODUCTION

The poverty rate is the ratio of the number of people whose income falls below the poverty line; taken as half the median household income of the total population [1]. In Indonesia, poverty rate is produced yearly by Statistics Indonesia by conducting a National Social and Economic Survey [4]. Despite all the benefits conducting this survey regularly every year, there are a couple of limitations such as (i) inability to gather information on poverty rate in between the surveys and (ii) requirement of certain resources in order to conduct the surveys.

Along with that, Indonesia abundantly produces big data from various resources, e.g. social media, financial transaction, transportation, call detail records, e-commerce, etc. These types of data have been considered as potential resources to complement periodic survey, even census, to monitor development indicators in which poverty rate is included. As a dimension of economics, poverty rate could be highly correlated with the data related to consumption and purchasing power, which can be potentially represented by e-commerce data. According to [3], revenue in the e-commerce market in Indonesia amounts to USD 9,138m in 2018, with user penetration is at 40% in 2018 and is expected to hit 48,3% in 2022.

This research aims to estimate poverty rate at city-level based on e-commerce data using machine learning methods i.e. Support Vector Regression (SVR) and Artificial Neural Network (ANN). Considering the representation of e-commerce in certain areas in Indonesia, the scope of this research is 118 cities in Java island.

2 DATASET

The main dataset used in this research is the advertisements of goods posted in one of the big e-commerce platforms in Indonesia, OLX. The following goods are included in the analysis.

This study utilizes two main data sources that complements each other.

(1) Car
(2) Motorbike
(3) House to sell
(4) House to rent
(5) Apartment to sell
(6) Apartment to rent
(7) Land to sell
(8) Land to rent

For each of those goods, the information of number of items sold, price sold, number of viewers, and number of buyers were extracted. Then, the aggregation by city has been done for each of those information per goods to calculate statistics measurements i.e. sum, average, and standard deviation, to capture both central tendency and variation of the data. In total, there are 96 initial features extracted for this research.

As the ground-truth, the poverty rate at city-level published by Statistics Indonesia (see Table 1). For both e-commerce and official data, the data in 2016 has been used for this research.

Table 1: Train-and-test splitting procedure

	Split 1	**Split 2**
Odd observation	Train	Test
Even observation	Test	Train

3 METHODOLOGY

3.1 Pre-processing

Given 96 features and city $i = i,..,118$, normalisation has been done for each feature with the following formula:

City ID 1	Poverty Rate 1	City ID 2	Poverty Rate 2	City ID 3	Poverty Rate 3	City ID 4	Poverty Rate 4	City ID 5	Poverty Rate 5
3101	11,4	3278	16,28	3329	19,79	3519	12,54	3521	15,61
3171	3,41	3279	7,41	3371	9,05	3520	11,35	3522	15,71
3172	3,24	3301	14,39	3372	10,89	3521	15,61	3523	17,08
3173	4,16	3302	17,52	3373	5,8	3522	15,71	3524	15,38
3174	3,64	3303	19,7	3374	4,97	3523	17,08	3525	13,63
3175	5,91	3304	18,37	3375	8,09	3524	15,38	3526	22,57
3201	8,96	3305	20,44	3376	8,26	3525	13,63	3527	25,69
3202	8,96	3306	14,27	3401	21,4	3526	22,57	3528	17,41
3203	12,21	3307	21,45	3402	16,33	3527	25,69	3529	20,2
3204	8	3308	13,07	3403	21,73	3528	17,41	3571	8,51
3205	12,81	3309	12,45	3404	9,46	3529	20,2	3572	7,29
3206	11,99	3310	14,89	3405	8,75	3571	8,51	3573	4,6
3207	8,98	3311	9,26	3501	16,68	3572	7,29	3574	8,17
3208	13,97	3312	12,98	3502	11,91	3573	4,6	3575	7,47
3209	14,77	3313	12,46	3503	13,39	3574	8,17	3576	6,16
3210	14,19	3314	14,86	3504	8,57	3575	7,47	3577	4,89
3211	11,36	3315	13,68	3505	9,97	3576	6,16	3578	5,82
3212	14,98	3316	13,52	3506	12,91	3577	4,89	3579	4,71
3213	12,27	3317	19,28	3507	11,53	3578	5,82	3601	10,43
3214	9,14	3318	11,95	3508	11,52	3579	4,71	3602	9,97
3215	10,37	3319	7,73	3509	11,22	3601	10,43	3603	5,71
3216	5,27	3320	8,5	3510	9,17	3602	9,97	3604	5,09
3217	12,67	3321	14,44	3511	14,96	3603	5,71	3671	5,04
3271	7,6	3322	8,15	3512	13,63	3604	5,09	3672	4,1
3272	8,79	3323	11,76	3513	20,82	3671	5,04	3673	6,28
3273	4,61	3324	11,62	3514	10,72	3672	4,1	3674	1,69
3274	10,36	3325	11,27	3515	6,44	3673	6,28		
3275	5,46	3326	12,84	3516	10,57	3674	1,69		
3276	2,4	3327	18,3	3517	10,79	3519	12,54		
3277	5,84	3328	10,09	3518	12,69	3520	11,35		

Figure 1: Poverty rate (%) at city-level, 2016 [5]

$$z_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)'} \quad (1)$$

z is the normalised value of the respected features, x is the original value of the respected features, min(x) and max(x) are the minimum and maximum value of .To perform the analysis, the data of 118 cities has been split into training and testing data with the procedure shown in Table 2.

3.2 Feature Selection

In this research, feature selection is done with Fast Correlation-Based Feature (FCBF). FCBF includes two aspects such as (i) decide whether a feature is relevant to the class or not and (ii) decide whether such a relevant feature is redundant, i.e. correlated with other features, or not.

The main principle of the FCBF algorithm is to maximize C-Correlation and minimize F-Correlation. Maximizing C-Correlation aims to find out which features are most closely related to the class label which means it is possible to predict the class label well. While minimizing F-Correlation means reducing the number of features/predictor redundant. The values of C-Correlation and F-Correlation are measured by Symmetrical Uncertainty (SU) to calculate non-linear correlations between two discrete random variables V and W. SU can be expressed as follows:

$$SU(V, W) = 2\left[\frac{IG(V|W)}{H(V) + H(W)}\right] \quad (2)$$

where,

$$IG(V|W) = H(V) - H(V|W) \quad (3)$$

$$H(V) = -\sum_i =_1 P(v_i) log_2(P(v_i)) \quad (4)$$

$$H(W) = -\sum_i = 1P(w) log_2(P(w_i)) \quad (5)$$

$$H(V|W) = -\sum_j =_1 P(W_j) \sum_i =_1 P(v_i|w_i) log_2(P(v_i|w_i)) \quad (6)$$

where $P(v_i)$ is the prior probabilities for elements V and $P(v_i|w_i)$ is the posterior probabilities of V to the W value. The range of SU values ranges between 0 and 1. The higher the SU value the higher the correlation between the two variable [7]. In this experiment,

we use SU threshold = 0 to provide feature opportunities with a small SU.

With FCBF, 29 features are successfully selected from 96 initial features.

3.3 Support Vector Regression (SVR)

SVR has the same concept as Support Vector Machine (SVM). Considering a set of data training, $\{(X_1, Z_1), ..., (X_1, Z_1)\}$ corresponds to the response of sensors where $\{X_i, Z_i\}$ are feature vector and target output, respectively. In ε-SVR, the primal optimization problem can be expressed as follows [6]:

$$\min_{w,b,\xi,xi*} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i + C \sum_{i=1}^{l} \xi*_i \tag{7}$$

$$subject\ to\ w^T \phi(X_i) + b - Z_i \leq \epsilon + \xi\ *_i\ Z_i - w^T \phi(X_i) - b$$

$$math \leq \epsilon + \xi*_i, \xi, \xi* \geq 0,\ i = 1, ..., l$$

while w, C, ξ, ϵ, b denote slope matrix, regularization parameter, slack variable for soft margin, the margin of tolerance, and the intercept/bias, respectively. The symbol $\phi(X_i)$ indicates mapping X_i into higher dimensional space. The dual problem optimization is given by

$$\min_{\alpha,\alpha*} \frac{1}{2}(\alpha - \alpha*)^T Q(\alpha - \alpha*) + \epsilon \sum_{i=1}^{l}(\alpha + \alpha*) + \sum_{i=1}^{l} Z_i(\alpha - \alpha*) \tag{8}$$

$$subject\ to\ e^T(\alpha - \alpha*) = 0,\ 0 \leq \alpha,\ \alpha* \leq C, i = 1,\ \ldots,\ l$$

where α and α^* denotes Lagrangian multipliers. $Q_{i,j} = K(X_i, X_j) \equiv \phi(X_i)^T \phi(X_j)$ *and* $e = [1, ..., 1]^T$. In linear SVR, the decision function is expressed by

$$Y = \sum_{i=1}^{l}(\alpha_i - \alpha_i*)\langle X_i, X\rangle + b \tag{9}$$

In non-linear SVR, the kernel function e.g., RBF transforms the data input into a higher dimensional feature space to perform the linear separation. The decision function is computed by

$$Y = \sum_{i=1}^{l}(\alpha_i - \alpha_i*)\langle \phi(X_i), \phi(X)\rangle + b \tag{10}$$

$$Y = \sum_{i=1}^{l}(\alpha_i - \alpha_i*)\langle K(X_i, X)\rangle + b \tag{11}$$

The RBF kernel is used to deal with non-linear data that can be computed by the following equation.

$$K(X_i, X) = exp(-\gamma ||X_i - X||^2) \tag{12}$$

In this experiment, grid search is performed to determined parameter from [0.01, 0.1, 1, 10, 100, 1000] and parameter from [0.01, 0.1, 1, 10, 100].

3.4 Artificial Neural Network (ANN)

ANN contains three layers including input, hidden, and output layer. These layers are built from several neurons that convert the signals based on the connection weight, bias, and activation function. Figure 2 shows the neural network architecture constructed for this research. The input layer contains 29 neurons correspond to the input features. Moreover, the hidden layer has 200 neurons with *tanh* activation function. Finally, the output layer only has one neuron to accommodate the continuous outputs in regression tasks.

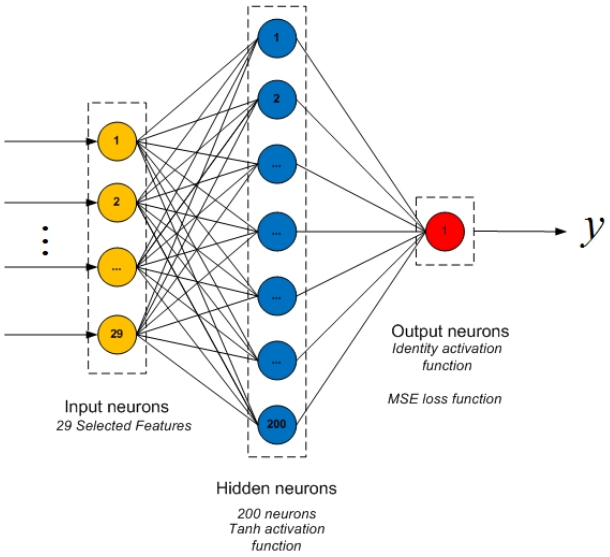

Figure 2: Architecture of ANN

3.5 Performance Measurement

The performance of the models measures by three metrics, i.e. root mean squared error (RMSE), bias factor, and accuracy factor. The equation and description of those measurements detailed in the Table 3.

4 RESULT AND CONCLUSION

Three following models have been built in this research.

(1) SVR-based model with 96 features
(2) SVR-based model with 29 features selected through FCBF procedure
(3) ANN-based model with 29 features selected through FCBF procedure

As discussed in the methodology section, the models performance is assessed by three metrics, i.e. RMSE, accuracy factor, and bias factor. The performance of the three models are shown in Table 4. RMSE of SVR-based model with FCBF (4.3363) is only slightly lower than SVR-based model without feature selection (4.9037), showing that with or without feature selection, SVR-based model produces almost the similar error from the estimation of poverty rate at city-level.

Table 2: Performance measurements

Metric	Equation	Description
RMSE	$RMSE(y,\hat{y}) = \sqrt{\frac{\sum_{i=1}^{L}(y_i-\hat{y}_i)^2}{L}}$	RMSE is used to measure the difference/error between actual and prediction vector. The lower RMSE value means the less difference between actual and prediction values.
B_f	$B_f(y,\hat{y}) = exp\left[\frac{\sum_{i=1}^{L}(ln(y_i)-ln(\hat{y}))}{L}\right]$	Bias factor denotes whether the predictions are "under" or "over" estimate against actual values. The unbiased prediction is indicated by B_f equals to 1. xxxx means that the prediction result is lower than actual values (underestimate) and vice versa [2]
A_f	$A_f(y,\hat{y}) = exp\left[\sqrt{\frac{\sum_{i=1}^{L}(ln(y_i)-ln(\hat{y}))^2}{L}}\right]$	Accuracy factor measures the average accuracy of the prediction model. The value of is equal or greater than one. The larger value than one indicates less accurate prediction results [2]

Meanwhile, the ANN-based model with FCBF produces much smaller RMSE, i.e. 0.2725, than the two precedent SVM-based models. This number indicates that estimating poverty rate at city-level with ANN-FCBF gives very good accuracy, since the RMSE value almost reaches zero value. Although all three models predict the poverty rate lower than the actual (underestimate), indicated by the value of bias factor that is less than 1, the bias factor for ANN-FCBF model is only slightly less from 1 (0.9981). Moreover, the ANN-FCBF model gives almost the accurate prediction since the value of accuracy factor is very close to 1 (1.0007).

Figure 2 and Figure 3 show predicted poverty rate based on SVR-FCBF and ANN-FCBF, respectively, compared with the actual poverty rate. The actual poverty rate is sorted for better visualization. The prediction of poverty rate produced by ANN-FCBF follows the actual poverty rate, with around one-third of city-level poverty rate are predicted exactly and precisely the same as the actual one. Cities with high actual poverty rate are relatively difficult to predict its poverty rate by SVM-FCBF model.

Table 3: Example of tweets per-district

Measurement	SVR	SVR-FCBF	ANN-FCBF
RMSE	4.9037	4.3363	0.2725
Accuracy factor	1.2853	1.1772	1.0007
Bias factor	0.9277	0.9883	0.9981

If the city-level poverty rate represented by ranges, as shown by the maps in Figure 4, Figure 5, and Figure 6, ANN-FCBF model predicts the category of poverty rate exactly the same as the category of actual poverty rate. Based on the results discussed above, this research concludes and suggests that:

(1) e-commerce data is potential to be used as proxy for city-level poverty rate,
(2) e-commerce data can be used to complement official data to poverty rate in between surveys and censuses,
(3) in the future, the method presented in this research is potential to be replicated and scaled-up for all cities in Indonesia or other administrative levels (i.e. province and sub-district), time series or panel data, and data from different e-commerce platforms.

REFERENCES

[1] Organisation for Economic Co-operation and Development. 2018. Poverty Rate. (2018). https://data.oecd.org/inequality/poverty-rate.htm.
[2] C. P. T. R. Baranyi J. 1999. Validating and comparing predictive model. *Journal of Food Microbiology* 3 (1999).
[3] Statistia. [n. d.]. ([n. d.]). https://www.statista.com/outlook/243/120/ecommerce/indonesia
[4] Badan Pusat Statistik. [n. d.]. ([n. d.]). https://microdata.bps.go.id/mikrodata/index.php/catalog/SUSENAS/about
[5] Badan Pusat Statistik. [n. d.]. Persentase Penduduk Miskin Menurut Kabupaten/Kota, 2015 - 2017. ([n. d.]). https://www.bps.go.id/dynamictable/2017/08/03/1261/persentase-penduduk-miskin-menurut-kabupaten-kota-2015---2017.html
[6] V.N. Vapnik. 1998. *Statistical Learning Theory.*
[7] H. L. L. Yu. 2003. Feature Selection for High-Dimensional Data: Fast Correlation-Based Filter Solution. *Twentieth International Conference on Machine Learning (ICML)* (2003).

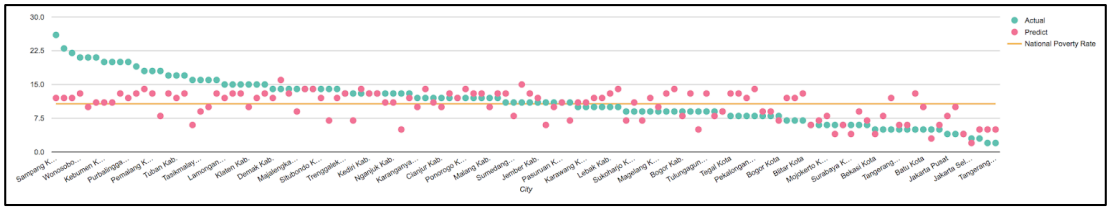

Figure 3: Predicted city-level poverty rate based on SVR-FCBF

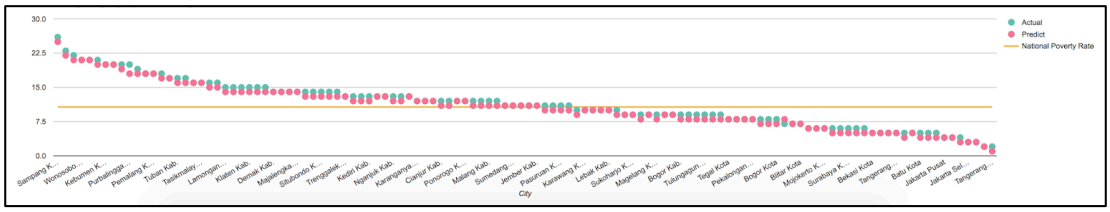

Figure 4: Predicted city-level poverty rate based on ANN-FCBF

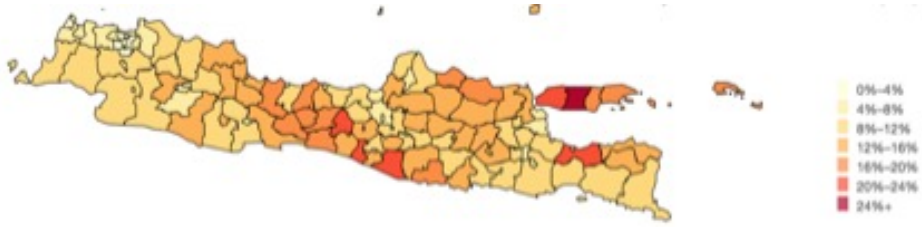

Figure 5: Actual city-level poverty rate map

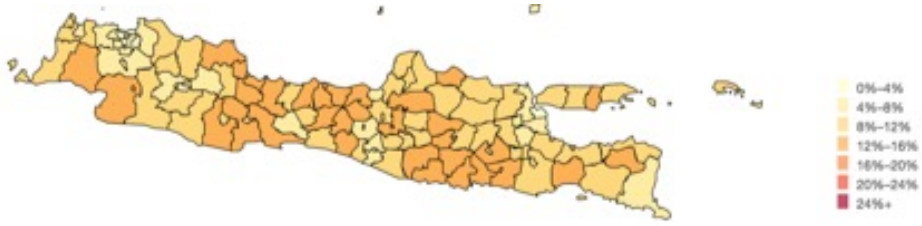

Figure 6: Predicted city-level poverty rate based on SVR-FCBF map

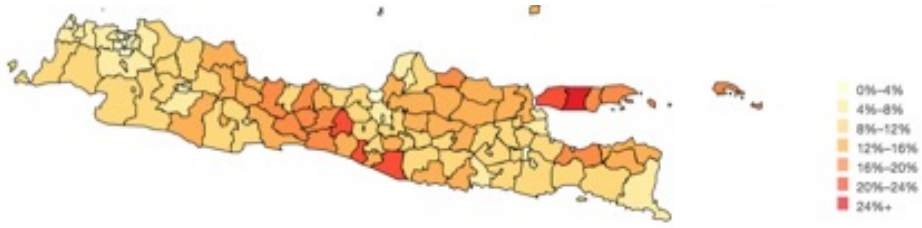

Figure 7: Predicted city-level poverty rate based on ANN-FCBF map

Estimating Poverty at the District Level with Social Media

Extended Abstract

Lili Ayu Wulandhari
Bina Nusantara University
Jakarta, Indonesia
lili.wulandhari@gmail.com

Sri Redjeki
STMIK AKAKOM Yogyakarta
Yogyakarta, Indonesia
dzeky@akakom.ac.id

Yunita Sari
Universitas Gadjah Mada
Yogyakarta, Indonesia
yunita.sari@ugm.ac.id

Widaryatmo
Bappenas
Jakarta, Indonesia
yunita.sari@ugm.ac.id

M. Rizal Khaefi
Pulse Lab Jakarta
Jakarta, Indonesia
muhammad.khaefi@un.or.id

ABSTRACT

The rise of social media has encouraged data driven research in many disciplines. Previous studies have utilised Twitter, a popular micro-blogging service as a valuable information resource for various applications including demographic identifications and event detections. This study explores how information from Twitter can be used as a leading poverty indicator which supports data obtained from survey-based method. We define an array of poverty-related keywords based on Socio Economic Survey (SUSENAS) conducted by National Statistic Agency (BPS) and apply a semi-supervised machine learning algorithm called Pseudo Labeling to identify tweets with poverty indicator. Results of our experiments on subset of Twitter data in JABODETABEK area show our pre-defined keywords are effective by obtaining 65% on accuracy. This study acts as a preliminary benchmark for further work on identifying poverty from social media especially in Bahasa Indonesia.

KEYWORDS

Poverty indicator, Twitter, Semi-supervised Learning

1 INTRODUCTION

According to a survey conducted by Indonesian Internet Service Providers Association (APJII) in 2017, Indonesia had 143.26 million internet users which covered nearly 54.68% of the total population of the country [2]. This made Indonesia in the top five countries with the highest internet users after China, India, USA, and Brazil. The survey revealed that 87.13% of the users utilised the internet to access social media. APJII reported that Twitter, a popular micro-blogging service, is one of the most visited social media [1]. Previous studies have indicated that user demographic profiles such as age, gender, education background, income, and occupation can be predicted from their posts in Twitter [6, 8, 11, 13]. In addition, Twitter has been widely used as a valuable information resource for various applications including earthquake detection [14], stock market prediction [5] [4], crime prediction [9] and analysis of politic power distribution in election [7]. Among the numerous potential applications, this study addresses the issue of identifying poverty indicator which presents outstanding advantages over traditional survey-based methods. Currently, poverty-related data (i.e. amount of income, source of income, number of dependent family member, and living condition) in Indonesia is obtained through survey-based method conducted by National Statistics Agency (BPS). This method may provide precise results, albeit expensive. Compared to survey-based method, Twitter provides a huge data volume which can be obtained at no-cost. In addition, Twitter enables real-time and direct surveillance which make it useful for detecting economic dynamic in the society.

Nevertheless, extracting poverty-related information from Twitter is challenging. Thus, the aim of this study is not to replace the survey-based method as the main source of the poverty data, but to provide additional information to support the survey-based data. This study explores how information from Twitter content can be used as leading poverty indicator. Experiments are carried out using subset of Twitter data in JABODETABEK (Jakarta Bogor Depok Tangerang Bekasi) area within the year of 2014. We first define an array of poverty-related keywords based on data from BPS. We then filter tweets that contain those keywords and perform human annotation to assign *poor* and *non-poor* labels to the selected tweets. We assume that Twitter users are middle-upper class society with good financial condition. Thus, the labels assigned do not represent the poverty level of an individual user but rather indicate the poverty condition in an area where the users reside/post their tweets. Labelling Tweets to generate training and test datasets is a labour-intensive process. Thus, we apply Pseudo-Labeling [10], a semi-supervised machine learning algorithm to classify tweets into the pre-defined classes. In this way, the model is not only learning from the labeled data but also from the large size of unlabeled data. The experiment results show that the pre-defined poverty-related keywords are effective to identify tweets with poverty indicator. Results from this study act as a benchmark for further work on identifying poverty from social media especially in Bahasa Indonesia.

The rest of the paper is organised as follows: in Section 2 we describe related work which use Twitter as the main source of data. Section 3.1 presents the details of the data and how the human annotation is conducted. Section 4 describes our methodology to identify poverty indicator on Twitter. Results and analysis are presented in Section 4. Finally we summarize our findings, and describe possible work to be explored further in Section 5.

2 RELATED WORK

The vast availability of social media data has encouraged data driven research in many disciplines. Twitter is one of the social media

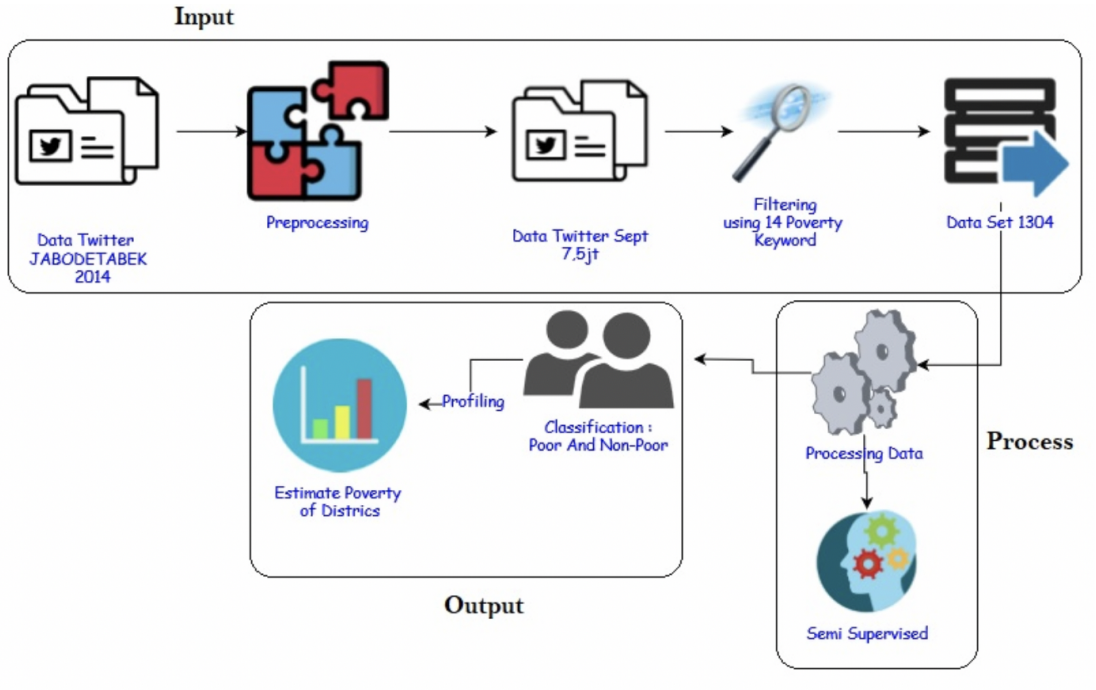

Figure 1: Flowchart of Research Methodology

which serves as a potential resource for many applications. Sakaki et al. [14] investigated the real-time interaction of events such as earthquakes, in Twitter and proposed an algorithm to monitor tweets and to detect a target event. They used Support Vector Machine (SVM) and three feature groups for event detection and applied *Kalman* and particle filters for location estimation. By considering each Twitter user as a sensor, Sakaki et al. constructed a reporting system which detects earthquake promptly and sends notification e-mails to registered users. They reported that their system is able to deliver notification faster than the announcements broadcasted by the Japan Meteorological Agency (JMA).

Another work by Aramaki et al. [3] addressed the issue of detecting influenza epidemic using Twitter. They constructed an influenza corpus consists of 0.4 million tweets which are divided into training and testing parts. Using two pre-defined conditions, human annotation was conducted to assign positive and negative labels to the tweets. Several machine learning classifiers with Bag Of Words (BOW) features were used to identify whether a given tweet is positive or negative. Their experiment results outperformed the state-of-the-art Google method by obtaining high correlation (correlation ratio=0.89).

Previous work has also tried to predict demographic attributes from Twitter which are useful for marketing, personalization, and legal investigation. A work by Burger et al. [6] constructed a large, multilingual dataset labeled with gender and explored several statistical approaches for identifying gender of Twitter users. In order to assign gender labels to the Twitter accounts, the authors sampled the corresponding user profiles which were obtained by following the Twitter URL links to several of the most represented blog sites in their dataset. Burger et al. used the content of the tweets and three user profiles including full name, screen name, and description to discriminate the gender of the users. Their approaches successfully obtained the best accuracy of 92%. In addition, by using only the content of the tweets, the model gained 76% accuracy.

Similar to Burger et al., Flekova et al. [8] explored stylistic variation with age and income on Twitter. By using variety of features such as word and character lengths, readability measures (i.e. the Automatic Readability Index, the Flesch Kincaid Grade Level, the Flesch Reading Ease), part-of-speech (POS), and contextuality measure combined with linear regression, they successfully predicted age and income groups of the Twitter users. Some interesting findings are Flesch Reading Ease–previously reported to correlate with education levels at a community level–is highly indicative for income. In addition, the increased use of nouns, determiners and adjectives is correlated higher with age as opposed to income.

3 METHODOLOGY

This research is conducted in four steps, namely data preparation, data preprocessing and annotation, pseudo labeling and evaluation. Data preparation aims to comprehend and analyze information contained in the data. This step becomes the justification which techniques are chosen for preprocessing. Preprocessing step is conducted to extract keyword as one of leading indicator from raw data. Result from preprocessing step is used in pseudo labeling algorithm. Detail explanation of data and pseudo labeling is presented in subsection 3.1 and subsection 3.3

Table 1: Example of Twitter Data

LAT	LON	PROV	PROV_ NAME	KAB	KAB_ NAME	KEC	KEC_ NAME	GENDER	CONTENT	TIME STAMP	SOURCE
106. 980. 765	-624. 022	32	Jawa Barat	3275	BEKASI	3275050	BEKASI SELATAN	Female	Apapn statusnya miskin atupun kaya tua muda sehat sakit sibuk lelah kewajiban seorang hamba harus tetap dilaksanakan.	140 950 4693	Twitter for Android
106. 826. 796	-6.168. 961	31	Daerah Khusus Ibukota Jakarta	3173	JAKARTA PUSAT	3173080	GAMBIR	Male	Orang miskin jangan melawan orang kaya dan orang kaya jangan melawan pejabat.Kalau mau mengubah bangsa ini jadilah PEJABAT.	140 973 8003	Twitter Web Client
106. 675. 791	-6.194. 607	36	Banten	3671	TANGERANG	3671020	CIPONDOH	Male	gua kan jelek miskin bego ga akan selamanya gua begitu	141 042 0787	Twitter for Android

3.1 Dataset Details

In this research we analyze Twitter data for JABODETABEK area for twelve months in 2014. From these twelve months, we exclude January, June, July and December due to special occasions such as new year and religious holidays happened in those months. Each tweet contains information as follows:

- TIMESTAMP : Timestamp of tweet post activity
- PROV : Code of province
- KAB : Code of region
- KEC : Code of district
- PROV_NAME : Name of province
- KAB_NAME : Name of region
- KEC_NAME : Name of district
- LAT : Coordinate for latitude
- LONG : Coordinate for longitude
- GENDER : Gender of Twitter user
- SOURCE : Source of tweet (web/ others)
- CONTENT : Tweet posted by user

Sample of Twitter data are presented in Table 1. In addition to that, we perform an analysis to the Twitter user's behaviour. Based on data profiling, we obtain that most users post tweets on the weekend, with the highest peak is on Saturday followed by Sunday and Friday (see Figure 2). While in a day, Twitter users tend to post in the span of 9.00 am - 4.00 pm around Jakarta Selatan and Jakarta Pusat area (see Figure 3). Moreover, the analysis provide an information that Twitter is used actively around business center, downtown and entertainment center area which spreads in Jakarta Selatan and Jakarta Pusat. This is equivalent to number of poverty content which is identified from these area.

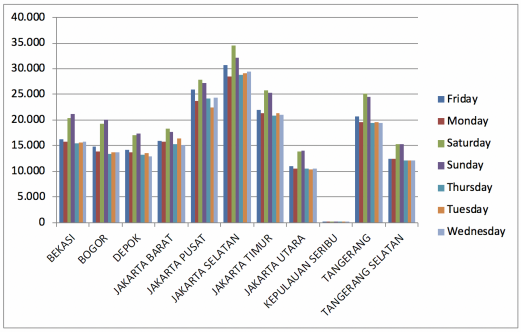

Figure 2: Twitter users behavior in a week within JABODETABEK area

From this data, poverty indicator keywords are extracted based on Socio Economic National Survey (SUSENAS) in 2018 (See Figure 2). This survey presents that things related to food such as *"beras"* (rice) dominates society needs. Therefore, we extract food and poverty related words as the keywords. Overall, fourteen keywords and tweets metadata are used as the input for identifying poverty indicator. Each non numeric features; PROV NAME, KAB NAME and GENDER is transformed into numeric form with following definition:

PROV_NAME:

- Daerah Khusus Ibukota Jakarta:31
- Jawa Barat: 32
- Banten: 36

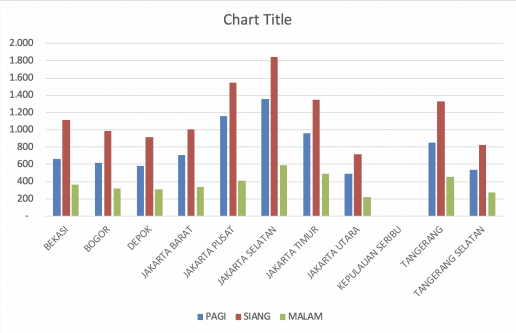

Figure 3: Twitter users behavior in a day within JABODETABEK area

KAB_NAME:

- Jakarta Selatan: 1
- Jakarta Timur: 2
- Jakarta Pusat: 3
- Jakarta Barat: 4
- Jakarta Utara: 5
- Kepulauan Seribu: 6
- Bogor: 7
- Depok: 8
- Bekasi: 9
- Tangerang: 10
- Tangerang Selatan: 11

GENDER:

- Male : 1
- Female: 2

This data is annotated manually into poor and non-poor classes to be the input data in pseudo labeling approach

Table 2: List of commodities contribute to poverty (%) (source: Socio Economic National Survey, March 2018)

Jenis komoditi	Perkotaan	Jenis komoditi	Perdesaan
(1)	(2)	(3)	(4)
Makanan:	**71,04**	**Makanan:**	**76,66**
Beras	20,95	Beras	26,79
Rokok kretek filter	11,07	Rokok kretek filter	10,21
Telur ayam ras	4,09	Telur ayam ras	3,28
Daging ayam ras	3,55	Gula pasir	3,07
Mie instan	2,43	Mie instan	2,21
Gula pasir	2,24	Daging ayam ras	2,08
Kopi bubuk & kopi instan (sachet)	1,88	Kopi bubuk & kopi instan (sachet)	1,93
Kue basah	1,78	Bawang merah	1,81
Tempe	1,74	Roti	1,80
Tahu	1,70	Kue basah	1,77
Roti	1,65	Cabe rawit	1,64
Bawang merah	1,50	Tempe	1,63
Lainnya	16,45	Lainnya	18,45
Bukan Makanan:	**28,96**	**Bukan Makanan:**	**23,34**
Perumahan	8,30	Perumahan	6,91
Bensin	4,36	Bensin	3,69
Listrik	3,89	Listrik	2,01
Pendidikan	1,99	Pendidikan	1,23
Perlengkapan mandi	1,30	Perlengkapan mandi	1,11
Angkutan	0,95	Kayu bakar	0,83
Kesehatan	0,85	Kesehatan	0,81
Lainnya	7,31	Lainnya	6,76

3.2 Manual Annotation

Manual annotation aims to assign poor and non-poor labels to each tweet. Each content of tweets is identified manually to provide training data for pseudo labeling. In total, there are 153 tweets have been labeled. Example of tweets with poor and non-poor indicator is presented in Table 3. This annotation also to determine whether tweets contain pre-defined keywords, with value 1 if the keywords are exist and value 0 otherwise.

Table 3: Example tweets with *poor* and *non-poor* indicator

class	tweets
poor	Jumlah uang yg beredar sama, namun penduduk meningkat, anak2 itu tidak dapat lapangan kerja, kemiskinan meningkat
poor	Dan anyway, karena pengontrolan yg kurang itu, negara dipuyengkan oleh beban generasi anak yg tidak terkontrol, efeknya,kemiskinan meningkat
poor	Awal bulan padahal, udah miskin ajayaa
poor	Negara miskin karena penduduknya sendiri. Yg miskin ngga berusaha maju yg kaya berlagak miskin punya mobil pribadi masih pake bbm subsidi
poor	Orang miskin jgn sakit biaya rumah sakit mahal. Orang miskin kerja lebih keras lg bayar uang kuliah anakmu dua kali lipat penghasilanmu.
non-poor	Ngedengerin ecen ngomong suaranya berat bgt kqya bawa beras sekarung -_-
non-poor	Kaya materi tapi miskin hati. Kasian
non-poor	makanan buat brino... bubur kacang merah + beras merah + daging ayam cincang.. udh kyk baby ni ci binyo
non-poor	Cendol terbuat dari tepung beras trus di kasih pandan gituu...??? Pantesss konyang wak dek nyoo... :)
non-poor	Minum beras kencur biar sixpack

3.3 Pseudo Labeling

We approach the task as a classification problem by applying Pseudo Labeling, a semi-supervised machine learning method to the Twitter corpus described in Section 3.1. Figure 4 shows the illustration of Pseudo Labeling. First, the model is trained on a small size of labeled data. Then, using the trained model, we predict labels on the unlabeled data creating pseudo-labels. The model is re-trained on the combination of the labeled and newly pseudo-labeled data. These steps are repeated until there is no more unlabeled data left. It is expected that the model performance will improve when more pseudo-label data added. In our experiment, we use the implementation of Pseudo Labeling from Scikit Learn [12]

Our model uses 17 numerical features include three tweets information: PROV, KEC, GENDER and 14 pre-defined keyword frequencies: *beras, sembako, miskin, paket sembako, beras murah, harga beras, beras naik, harga naik, antri sembako, hidup susah, #berasnaik, #bbmnaik, #sembako, #berasmahal.* These keywords were defined

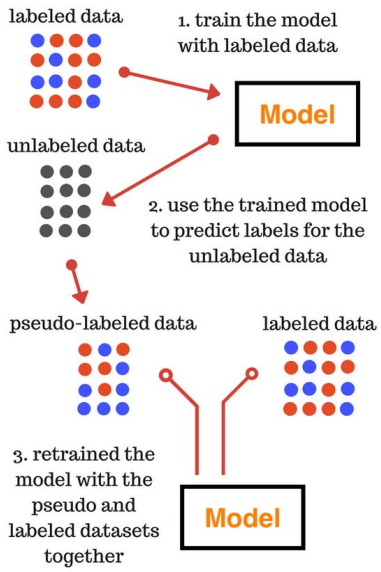

Figure 4: Pseudo labeling

based on Socio Economic National Survey (SUSENAS) in 2018 (see details in Section 3.1).

4 RESULTS AND ANALYSIS

Table 4 presents the results of our experiment. Using only 10 labeled data as initial seeds, our model obtained 65.73% on accuracy with precision and recall of 64% and 66% respectively. Our results demonstrate that the chosen features are effective for this task. Moreover, the results provide evidences that the pre-defined keywords are suitable for identifying tweets with poverty-related information. The model's performance can be improved by adding more pseudo labeled data. However, in our experiment we did not remove fake tweets which are generated automatically using bot software. Thus, our results may not describe the real poverty condition in a particular area.

Table 4: Experiment results

precision	recall	F1-score	accuracy
64%	66%	65%	65.73%

5 CONCLUSIONS AND FUTURE WORKS

This paper describes experiments on identifying poverty indicator on Twitter. We find that the tweets with poverty-related information can be classified using several pre-defined poverty-related keywords. Pseudo labeling is proved to be suitable method for this task since the available training data is limited. A possible extension of this work is to apply pre-processing steps to remove fake twitter accounts. This is to assure that only tweets from real account are included. In addition to that it is interesting to investigate the correlation between the authors' writing style and their poverty/income level. Thus, it can validate our initial assumption that most Twitter users are from middle-upper class society.

ACKNOWLEDGMENT

We would like to acknowledge Pulse Lab Jakarta for organising Research Dive event and providing the data. We also wish to thank Prof. Arief Anshory Yusuf and Prof. Dedi Rosadi for their insightful feedbacks.

REFERENCES

[1] APJI APJII. 2016. Penetrasi dan Perilaku Pengguna Internet Indonesia. *Infografis Hasil Survey* (2016), 1–35.
[2] APJI APJII. 2017. Penetrasi dan Perilaku Pengguna Internet Indonesia. *Infografis Hasil Survey* (2017), 1–39.
[3] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1568–1576. http://dl.acm.org/citation.cfm?id=2145432.2145600
[4] Pablo D Azar and Andrew W Lo. 2016. Practical Applications of The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds. *Practical Applications* 4, 2 (2016), 1–4.
[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.
[6] John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1301–1309. http://dl.acm.org/citation.cfm?id=2145432.2145568
[7] Nugroho Dwi Prasetyo and Claudia Hauff. 2015. Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 149–158.
[8] Lucie Flekova, Daniel PreoŢiuc-Pietro, and Lyle Ungar. 2016. Exploring Stylistic Variation with Age and Income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 313–319. https://doi.org/10.18653/v1/P16-2051
[9] Matthew S Gerber. 2014. Predicting crime using Twitter and kernel density estimation. , 115–125 pages.
[10] Dong-Hyun Lee. 2013. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. (07 2013).
[11] Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2017. Language-independent Gender Prediction on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, 1–6. http://aclweb.org/anthology/W17-2901
[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[13] Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. 2014. Twitter User Gender Inference Using Combined Analysis of Text and Image Processing. In *Proceedings of the Third Workshop on Vision and Language*. Dublin City University and the Association for Computational Linguistics, 54–61. https://doi.org/10.3115/v1/W14-5408
[14] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.

Estimating Poverty at the Household Level with Social Media Data and Household Survey Results

Eka Puspitawati
Universitas Pertamina
Jakarta, Indonesia
ekapuspitawati@gmail.com

Hizkia H. D. Tasik
Sam Ratulangi University
Manado, Indonesia
hizkiatasik1@gmail.com

Nurlatifah
Statistics Indonesia
Bogor, Indonesia
ifah@bps.go.id

M. Eko Fahillah
TNP2K
Jakarta, Indonesia
eko.fadhillah@tnp2k.go.id

Rajius Idzalika
Pulse Lab Jakarta
Jakarta, Indonesia
rajius.idzalika@un.or.id

ABSTRACT

Social media data, collected automatically through the interaction of individuals, can provide insights on many emerging issues such as from social life to politics. This extended abstract will explore how social media data correlate to poverty measurement on both regional and household levels using community/village level poverty mapping and household poverty measurement survey. Discovering ways to measure poverty through social media data offers a more rapid and inexpensive measure of poverty compared to completing poverty mapping or household surveys. We describe the statistical techniques that allow us to evaluate the potency of poverty estimation using social media data, particularly Twitter. We also discuss follow-ups that can contribute to better estimations.

KEYWORDS

Poverty, Twitter, Household Survey

1 INTRODUCTION

Household/individual poverty in Indonesia is usually measured through household/individual expenditure level. This measurement requires a survey on a representative sample of the household population in Indonesia. The household survey is generally expensive and time-consuming, and certain time-frame constrains the data generated from the survey.

To overcome that restriction, we investigate the available connection between social media (Twitter) data and household survey data. Twitter users have significantly increased in numbers for the past six years (2012-2018) especially in Asia-Pacific.

Since increased access to internet services boosts economic growth and improves the well-being of the poor [2], we are interested to see how the data from Twitter users (with internet access) can explain the poverty in both regional and household levels.

We are using Twitter data of Jabodetabek region in 2014 available as part of Research Dive 7 [3] initiative by UN Pulse Lab. For the poverty measurement data, we use Poverty Map of Indonesia 2015 representing poverty (headcount ratio and GINI ratio) at community/village level available from SMERU research institute, and the Indonesian Family Life Survey (IFLS) [1] available from RAND Corporation. The poverty map data is used to investigate the correlation between social media data and regional poverty level measurement, while the IFLS data is used to investigate the correlation between social media and household poverty level measurement.

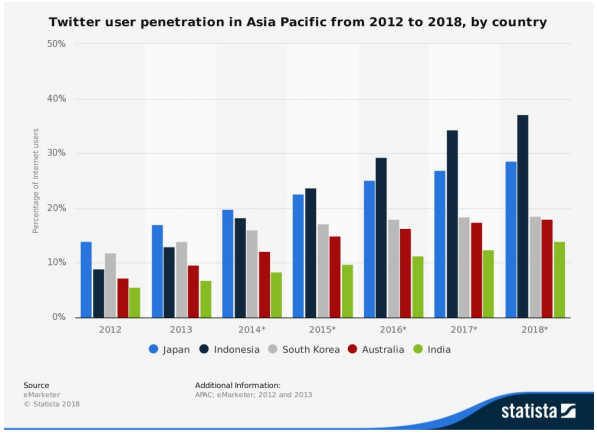

Figure 1: Twitter user penetration from 2012-2018, by country

2 TWITTER METADATA PERFORMANCE

Expanding on available Twitter data, we find variation between districts in Jabotabek regarding total Twitter User IDs, some messages posted, some mentioned in messages and total hashtag in messages. Districts observed in Jabodetabek region are:

3 REGIONAL LEVEL CORRELATION WITH POVERTY

SMERU Poverty map of Indonesia 2015 calculates various poverty measures based on several surveys conducted by Statistics Indonesia. We use two poverty measures to explore against social media data: poverty headcount index and Gini coefficient.

Poverty headcount index (P0) is defined as:

$$P_0 = \frac{1}{N}\sum_{i=1}^{q}\left(\frac{z - y_1}{z}^0\right) \quad (1)$$

Where:
P_0 : Headcount index

3101	Kepulauan Seribu
3171	Kota Jakarta Selatan
3172	Kota Jakarta Timur
3173	Kota Jakarta Pusat
3174	Kota Jakarta Barat
3175	Kota Jakarta Utara
3201	Kabupaten Bogor
3216	Kabupaten Bekasi
3271	Kota Bogor
3275	Kota Bekasi
3276	Kota Depok
3603	Kabupaten Tangerang
3671	Kota Tangerang
3674	Kota Tangerang Selatan

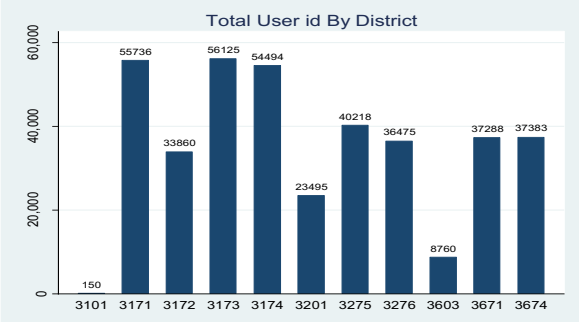

Figure 2: Total Twitter User IDs in Jabotabek region 2014, at district level

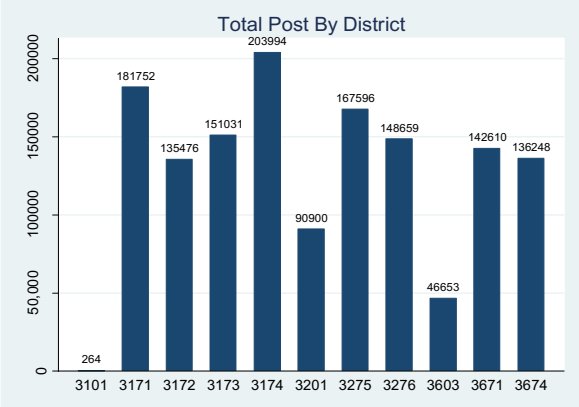

Figure 3: Total Twitter message posted in Jabotabek region 2014, at district level

z : Poverty line

y_1 : Average monthly per-capita expenditure of household below the poverty line

q : Number of the population below the poverty line

N : Number of population

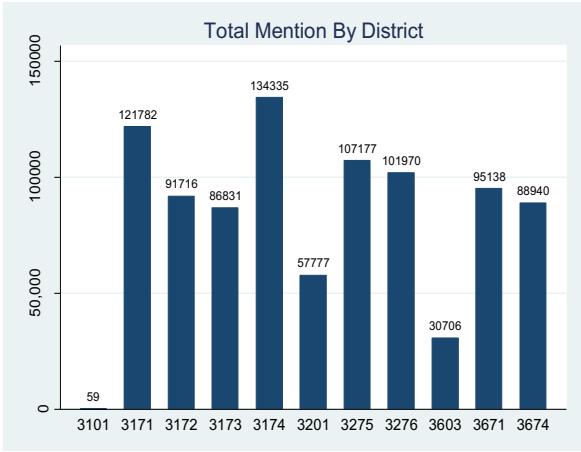

Figure 4: Total mentions in Twitter message in Jabotabek region 2014, at district level

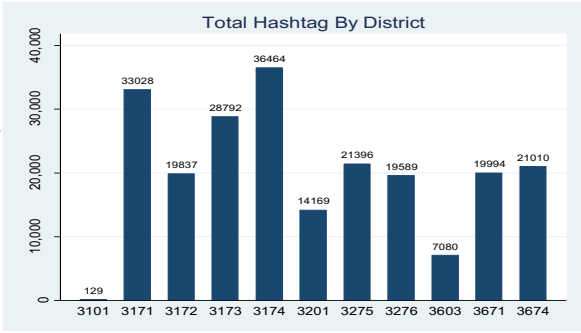

Figure 5: Total hashtag in Twitter's message in Jabotabek region 2014, at district level

Poverty headcount index directly indicates the ratio of people below the poverty line in a region. A higher index indicates a higher level of poverty.

Gini coefficient (G1) is defined as:

$$G_1 = 1 - \frac{1}{N}\sum_{i=1}^{N}(Y_i + Y_{i-1}) \tag{2}$$

Where:

G_1 : Gini coefficient

N : Number of population

y_i : Average monthly per-capita expenditure of household below the poverty line

Gini coefficient indicates the equality of distribution, Gini coefficient of zero (0) represents perfect equality while the Gini coefficient of one (1) represents maximum inequality.

3.1 Correlation between Twitter's aggregated data and poverty headcount index

To explore the correlation between Twitter and regional poverty headcount index at the community/village level, we need to aggregate the Twitter data into the community/village level and compared the results. We use Ordinary Least Squares (OLS) and observe the resulted significance of Twitter aggregated indicators towards estimating poverty headcount:

$$y_i = \beta_0 + \beta_1 ln(x_{i1}) + \beta_2 ln(x_{i2}) + \beta_3 ln(x_{i3}) + \beta_4 ln(x_{i4}) + \beta_5 ln(x_{i5}) + \epsilon \quad (3)$$

Where:
y_i: headcount index
x_{i1}: total number of unique Twitter User IDs
x_{i2}: total number of Twitter's message posted
x_{i3}: total number of mentions (@) in twitter message
x_{i4}: total number of link (http://) in Twitter message
x_{i5}: total number unique location (latitude/longitude) of Twitter posts
β: estimated regression coefficients
ϵ: errors

The results suggest that the headcount index has a stronger correlation with unique Twitter User IDs and unique Twitter posts (mobility). The findings show that a higher number of Twitter users and mobility of users can indicate communities/villages with lower poverty headcount.

Variables	Y Poor head
ln_total_user_id	-0.11495**
	(0.05778)
ln_total_post	-0.05048
	(0.10393)
ln_mention	0.02122
	(0.06475)
ln_hashtag	-0.00730
	(0.03431)
ln_links	-0.09113
	(0.05894)
ln_locations	-0.21993***
	(0.06801)
Constant	-1.41331***
	(0.12818)
Observations	1,008
R-squared	0,44891

***p<0.01, **p<0.05, p<0.1

Table 1. Estimation results for correlation between Twitter data and poverty headcount index

3.2 Correlation between Twitters aggregated data and Gini ratio quintile

To explore the correlation between Twitter and regional Gini Ratio Quintile at community/village level, we take the same aggregation process to prepare comparable Twitter datasets. We use Ordered Logit method and observe the resulted significance of Twitter aggregated indicators towards estimating Gini quantile:

$$Pr(Y = 1|X_1, X_2, X_3, X_4, X_5, X_6) \quad (4)$$

$$= \frac{1}{1 + e^{-}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6)} \quad (5)$$

Where:

Y: Quintile Gini Ratio (percent)
X1: Total User IDs of Twitter (number)
X2: Total post in Twitter (number)
X3: Average mention in Twitter per user (number)
X4: Average hashtag in Twitter per user (number)
X5: Average link for Twitter per user (number)
X6: Average locations for Twitter per user (number)
β(1,...6): Ordered log-odds (logit) regression coefficients
e : exponential

The results suggest that the headcount index has a stronger correlation with unique Twitter User IDs and average links posted (sharing internet content resource). This outcome shows that a higher number of Twitter users and the higher level of the resource sharing can indicate communities/villages with lower Gini quintile.

Variables	Y (Gini Ratio Outline)
total_user_id	-0.00199***
	(0.00036)
total_post	-0.00005
	(0.00010)
avg_mention_per_user	-0.01490
	(0.01311)
avg_hashtag_per_user	-0.01452
	(0.03309)
avg_links_per_user	0.04284*
	(0.02405)
avg_location_per_user	-0.04441
	(0.03525)
Constant cut1	-0.90845***
	(0.09990)
Constant cut2	-0.17323*
	(0.09498)
Constant cut3	-1.12987***
	(0.10124)
Constant cut4	2.35951***
	(0.12148)
Constant	-1.41331***
	(0.12818)
Observations	1,199
Pseudo R2	0.0729

***p<0.01, **p<0.05, p<0.1

Table 2. Estimation results for the correlation between Twitter data and Gini ratio quintile

4 HOUSEHOLD LEVEL CORRELATION WITH POVERTY

4.1 Data Preparation

Twitter does not have information on users income or expenditure. To provide such information, we borrow the information of household expenditure from the IFLS. The assumption involved is that observable attributes on both data source are the matching factors, so that Twitter user with specific matching factors is assumed to be identical with IFLS individuals with similar characteristics. This individual later is connected to their household ID to obtain information on household expenditure.

The first step of this analysis is to seek the common factors on both sides. We selected November 2014 as the time frame for Twitter data, to match it with the IFLS survey wave 5 that only started in October 2014, and to avoid anomalies during the year-end. We also restricted the age of the Twitter user and IFLS individuals to be between 16-60 years old. Further, the common factors identified are gender (F/M), province, district and sub-district code of home location, and the type of device used (notebook/cellphone). The matching factors are so limited, and we expect that upcoming surveys include more of internet related information for this proof of concept to be more reliable.

From Twitter extraction, we obtain 128,761 users to match with 3,204 of IFLS individuals and their respective households.

4.2 Matching based imputation and significant proxy from Twitter indicators

The second step is to do imputation of Twitter user expenditure based on the similarity with IFLS individuals. Among the imputation methods, we simply utilize univariate method of linear regression to estimate the expenditure because our expenditure variable is continuous. Imputation is originally a method to fill in missing values using a credible and scientific way. According to Rubin[4], inferences from multiple imputation when done properly is statistically valid. However, for experiment purpose we only produce one imputed data set as the showcase.

After obtaining the imputed expenditure for Twitter user, we regress by OLS expenditure on several common variables as well metadata attributes such as type of device, gender, number of tweets, number of mentions, number of hashtags, number of links, number of unique locations for tweet posts, the maximum and minimum of latitude and longitude of tweet posts, and the width of mobility. The latter variable is derived from maximum and minimum latitude and longitude.

A preliminary insight provided in Table 3 suggests that users who access Twitter using cellphone have a significantly higher expenditure compared to a user who accesses Twitter via notebook. From the spatial perspective, lower latitude and lower longitude of tweet posts in Jabodetabek area are negatively correlated with expenditure.

Variables	Y (Expenditure)
hp	3.425e-02***
	(<2e-16)
male	3.679e-04
	(0.1729)
tweets	1552e-05
	(0.1752)
mentions	7.793e-06
	(0.3196)
hashtags	-2.200e-05
	(0.0837)
links	-3.024e-05
	(0.0516)
location_count	-2.510e-05
	(0.1006)
min_lat	-5.122e-01***
	(<2e-16)
max_lat	1.216e-01***
	(<2e-16)
min_lon	-6.750e-02***
	(<2e-16)
max_lon	1.404e-02***
	(1.68e-09)
distance_box	1.438e-02
	(0.2667)
Adjusted R-sq	0.5287
p-value	<2.2e-16

***p<0.01, **p<0.05, p<0.1

REFERENCES

[1] RAND Corporation. 2015. The Indonesian Family Life Survey. (2015). https://www.rand.org/labor/FLS/IFLS.html

[2] Hernan Galperin and M. Fernanda Viecens. [n. d.]. Connected for Development? Theory and evidence about the impact of Internet technologies on poverty alleviation. *Development Policy Review* 35, 3 ([n. d.]), 315–336. https://doi.org/10.1111/dpr.12210 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/dpr.12210

[3] Pulse Lab Jakarta. 2018. Artificial Intelligence and Machine Learning for Estimating Poverty. (2018). http://rd.pulselabjakarta.id/research_dive/getdetail/21

http://rd.pulselabjakarta.id/

Pulse Lab Jakarta is grateful for the generous support from
the Government of Australia