RESEARCH QUESTION

Using topics in Twitter to predict cigarette consumption behavior among 7 provinces in Indonesia

Christanto Maulana Adityanugraha, S. Psi.

Master Applicant at Digital Humanities, University of Regensburg

Tobacco has been world's most important trading commodity since 1920. The world's favorite tobacco product is cigarette with 80 per cent of world tobacco is used, the rest products are pipes, cigars, roll-your-own, bidis, and kretek cigarettes [1]. The peak of cigarette consumption happened in year 1975 with around 3000 cigarettes was consumed by adult annually [2].

World Health Organization estimated that the number of smoking populations above 15 years old in Indonesia would be increasing about 6.25 per cent in 2020-2025 [3]. Back in 2010, smoking trend among men was dominated by the age-group 55-69 and among women by the age-group above 70. Ranked in 5th world's most cigarette consumption, over 96 million population in Indonesian in 2025 is estimated to buy cigarette and kretek cigarette products type.

For sure, Indonesian Government will raise the tobacco product taxation in coming years to discourage people from smoking. But no further policy will be issued or revised due to socio-economy factors in tobacco. Tobacco industry has provided jobs, tax revenue and foreign exchange earnings. Indonesian cigarette market is estimated being stable and on high demand for few next decades.

Speaking about how to measure and predict cigarettes consumption in upcoming years, there is always a question of which methods best used in last decades. Questionnaire about participant's cigarette consumption and cigarette sale and shipment data has been used for so long, but they are expensive, time-consuming, and vulnerable to participant recall bias, socially desirable responded, and unintentionally biased by researcher attendance [4]. Nowadays, using big data as a way to gather human behavior data becomes more popular because it is less expensive, efficient, and flexible.

The using of Twitter to access millions of tweets in real time is a big solution to find a population-level data without presenting any research or response biases because it can be operated just in a few lines of computer codes and remoted [5]. Many psychologists nowadays use big data analysis to get pure information about population's personalities, perspectival cognition, and market behavior, because it doesn't require any observer attendance during observation. A deeper understanding and a predictive modelling are key advantages of big data research to tackle the complexity of real-world human behavior. As an example, Twitter data can be used for sentiment analysis and an automated tobacco surveillance application in United States of America, which uses Naïve Bayes' predictive modelling. Myslin et al.'s study about tobacco related Tweets could depict a sentiment analysis of tobacco products. Once a dataset of sentiment analysis be founded, an operation of multinomial Naïve Bias model using N-grams text representation is possible to make a machine learning algorithm of predicting smoking behavior [6]. With using those above-mentioned methods in psychological data analysis, I can control some important data to gain information for such marketing or public policy knowledge.

This study was purposed to discover the usefulness of Tweets data resource for a deeper understanding of cigarette consumption in the seven provinces in Java and Bali islands, Indonesia. My research questions are: 1) How to build a predictive model of the most occurred Tweets topics as a behavior trend of cigarette consumption? 2) Which Tweets topics are most positively and negatively correlated with cigarette consumption behavior in the seven provinces? 3) How far the socio-economy and demography index play a role as mediator/ covariance between smoking behavior and Tweets topics?

My reason to choose Twitter as the main source for data collecting is because Twitter has the most active daily users who expressing their current opinion, statement and interaction in a written text form, which is possible to analyze using Natural Language Processing method in Python. Twitter is a free social media and microblogging service which everyone who's already signed up can post, like and retweet short messages called Tweets. With Twitter Developers, it is possible to stream Tweets in real time for research and study purpose using Twitter Application Programming Interface. In 2020 Twitter has 11.8 Million users in Indonesia with approximately 47 Million dailies Tweets [7]. This great number of Tweets and various regional languages made me to consider for collecting Tweets in just seven Province in Indonesia: Banten, Special Capital Region of Jakarta, Jawa Barat, Jawa Tengah, Special Region of Yogyakarta, Jawa Timur and Bali. These seven provinces are Indonesian's most populated region and Jakarta itself was the world's busiest and most active city in Tweets with a total 2% of Tweets worlds volume [8].

References

1. Liemt, Gijsbert van. 2002. The world tobacco industry: Trends and prospects. Working paper. International Labour Office, Geneva.
2. World Health Organization (WHO). 1997. Tobacco or health: A global status report (WHO, Geneva).
3. World Health Organization (WHO). 2015. WHO global report on trends in Prevalence on tobacco smoking 2015. WHO Library Cataloguing-in Publication Data.
4. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. Journal of Clinical Epidemiology 2014; 67 (3): 267–277. https://doi.org/10. 1016/j.jclinepi.2013.08.015 PMID: 24275499
5. American Psychological Association. 2018. Big Data Gets Bigger. In: monitor on psychology: a publication of the American Psychological Association; 49 (10), p. 68-72. Washington, DC: American Psychological Association Publication.
6. Myslin M, Zhu SH, Chapman W, Conway M. 2013. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. Journal of Medical Internet Research (vol. 15, iss. 8, e.174, p.1-16). San Diego: Department of Medicine University of California.
7. Clement J. 2020. Countries with the Most Twitter Users 2020. Digital 2020: April Global Statshot Report, p.84. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries
8. Carley KM, Malik M, Kowalchuk M, Pfeffer J, Landwehr P. 2015. Twitter Usage in Indonesia. Technical Report December 2015. Pittsburgh: Center for the Computational Analysis of Social and Organizational Systems, Institute for Software Research, School of Computer Science, Carnegie Mellon University.