Automatic Demographic Classification of Indonesian Twitter Users

Hashfi Rasis, Alva Erwin, James Purnama, Maulahikmah Galinium
Department of Information Technology, Faculty of Engineering and Information Technology,
Swiss German University

Keywords:

Twitter
Demographics
Classification
Indonesia
Users

ABSTRACT

Demographic Classification is a method to classify people by its demographic. Twitter has become one of largest social media which there are millions of tweets posted every day. Indonesia also becomes one of major country who uses Twitter. In Twitter there is no way to know how to classify each user into its demographic attributes. Because in Twitter profile there's no demographic attributes like gender or age. This research will focus on how to classify Indonesian Twitter users based on their gender, age and occupation. This research will use Naïve Bayes and K-Nearest Neighbor as its classifier algorithm. According to the result, Naïve Bayes performs well only in gender classification while K-Nearest Neighbor does not perform well in any demographic classification. Testing set successfully classified gender but failed with age and occupation

Corresponding Author:

Hashfi Rasis,
Departement of Information Technology, Faculty of Engineering and Information Technology,
Swiss German University,
Edu Town BSD City, Tangerang Selatan, Banten, Indonesia.
Email: hashfi.rasis@student.sgu.ac.id

1. INTRODUCTION

Nowadays social media has become trend not only for youngsters but elderly have begun to aware about social media. There is only small amount of people who don't have social media such as Facebook and Twitter. Social media has become a place not only for chat or meeting up with friends in the internet. But it can lead people to find people who have same interests to meet up with each other and talk about it together. Social media has been a part of our life nowadays. At least once a day people are checking their social media or even they post something there. With huge amount of posts (in the twitter case, tweets) per day, this will make social media as source of big data that can be analyzed for research purpose.

Demographic Classification is classification among people around the world according to the class that has been defined such as age, gender, etc. People may be classified more than two classes at the same time according to their personality and habits. In social media to classify the users based on their age and gender, we can find them right in their profile. But not all social media contains these important information only several like Facebook and Blogs that have the age and gender attributes that are mandatory to have in their profile. However with the features of social media to hide the profile information, there are some people that use the feature to hide their information from others. This is not the main problem as we can crawl the data with a simple crawling program. The main problem with the classification is the accuracy. We have to accurately classify each user into its correct demographics.

The rapid growth of social media like Facebook and Twitter has led to mass volume of user-generated informal text [1]. Indonesia has become a country with many twitter users and it will keep increasing over time. Over 91% of Twitter users choose to make their profile and communication history publicly visible [4].With so many data around social media, people began researching of demographic classification. There are classifications based on age, gender, occupation. With these classifications we can know what the people need based on its classification. However, unlike Facebook, Twitter has limited metadata available about its users [7]. Important attributes of the user such as age and gender that are directly useful for providing personalized services are not typically available. Because twitter doesn't provide this

information as twitter only showing the total of followers, total of whom you follows, date joined and total number of tweets. This make the demographic classification on twitter is more difficult than Facebook. The classifications that we want to analyze are Gender, Age and Occupation.

[1] The approach to distinguish the gender of Twitter users that we are using is comparing its full name and screen-name with the database that we have. We have already scraping the baby names from website and put it in our database. Gender probably is the most simplistic classification because we only distinguish between male and female and we can simply get it by comparing the names with the names that we have in database

[2] The approach to distinguish the age of Twitter users that we are using is analyzing the tweets of certain user and finding samples of users with age that we have already known. Whenever an user tweets about a certain topic we can give scores to them and if we found the user tweets aout the topic over and over again then we can put them in age range which currently discussing about the topic

[3] The approach to distinguish the occupation of Twitter users that we are using is almost the same as age but we just only cover the topics of school, university and office. Because we divide the occupation only as "Pelajar", "Mahasiswa", "Karyawan"

2. DATA

Twitter is a micro-blogging platform whose users post a short message called tweets. In late 2010, it was estimated that Twitter had 175 million registered users worldwide, producing 65 million tweets per day. Because of its massive volume the twitter become an interest of social media researcher, especially in data mining. In addition to their tweets, each twitter users also has profile with these attributes:

- Screen name (e.g., HalidaEdaNH, jaztinbibur)
- Full name (e.g., Halida Eda, Jaztin Bibur)
- Location (e.g., Indonesia, Jakarta)
- URL (e.g., user's website, user's Facebook, user's blog)
- Description (e.g., A happy executive which run several business)

Screen name is mandatory to have but the others are optional and can be empty. All of these profile elements can be changed. As seen here there is no demographic attributes that we are interested in such as age and gender. The profile elements above are not directly useful when we want to classify them into demographic attributes that we have defined earlier.

3. DEMOGRAPHIC CLASSIFICATION

Demographic analysis [8] includes the sets of methods that allow us to measure the dimensions and dynamics of populations. These methods have primarily been developed to study human populations, but are extended to a variety of areas where researchers want to know how populations of social actors can change across time through processes of birth, death, and migration. However, demographic analysis in social media is quite different because not all of social media display information about the user information. In Twitter we cannot determine user's gender and age without further inspection of their timeline. However, in other social media like Facebook, we can automatically determine user's gender and age just by inspecting their profile.

Accurate prediction of demographic classification has proven useful in marketing, personalization and investigation [1]. The main interests of demographic attributes are age and gender. Because knowing age and gender makes the marketing, personalization and investigation become easier.

The goal of Demographic Classification is to accurately classify each user based on the pre-defined attributes. The main point in the demographic classification is accuracy, the more accurate the better. In this research we classify the attributes as these three attributes:

1) Gender only consists of male and female
2) Occupation consists of 3 attributes: "Pelajar", "Maha-siswa" and "Karyawan".
3) Age consists of less than 25 years old and more than 25 years old.

These attributes will be a base to categorize users based on the demographic aspects.

4. TWITTER CRAWLER

This research will require numbers of automations process of information extracting from Twitter. The purpose of the crawler is to extract information about tweets, which tweets them, when they tweet them and where they tweets. This twitter crawler acts as an information gathering in order to find the demographic information about the user.

The crawler is built in Java by using Twitter4J API. Twit-ter4J API is one of best Twitter API for

Java. They also have Twitter4J Stream API, the difference between them is Twitter4J Stream API retrieving the data real-time, while the standard Twitter4J API doesn't.

In order to retrieve accurate information, the Twitter4J Search API is used in this research. Twitter4J Search API is the same as Twitter Search API but it's specifically designed for Java by Twitter4J. The advantage of Twitter4J Search API than Twitter4J Stream API is it can retrieve the data since the user start using twitter while Twitter4J Stream API can only retrieve the data real time. The other is in search API we can define the language of tweet we want to search while in stream API we cannot do that. This is important because we only want to search tweets in Indonesian language and not the other language.

The crawler can also be used to see the timeline of the user. Timeline is a list of tweets from the specific user. With inspecting user's timeline, we can know what kind of tweet that has been posted by the user. This is the main point of investigating user's tweets to be classified. Firstly we have to determine the keywords in each classification then if we find those keywords in the user's timeline. If the keywords repeated many times in user's timeline then we can classify the user to be in one of classification attributes.

The problem with the crawler is there are users whose timeline cannot be retrieved because of several reasons. Either they have not tweet anything, their accounts is protected or their accounts has been suspended. It is a problem that often arise when getting the data for training set.

Table 1. Testing Set and Training Set

	Total of Users
Training Set	500
Testing Set	410

5. ANALYSIS

To get the information of gender and age in twitter we have to search deep into the user's timeline. Timeline is a list of tweets that is posted by the particular user. With inspecting each user's post in their timeline we can see that they will subconsciously talk about a topic over and over again. This indicates that the user really has interest in the topic. With finding more person who has interest in the same topic then we can find out the majority of users who have interest in the topic. After that we can conclude that if there is another user who has tweeted about a certain topic over and over, we can list the user to be the same as majority of user who has interest in that topic.

For gender we use 500 training set which consists of 250 male and female each. Firstly we find users who are already identified as male and female and classify them as male and female. Next, we find users which gender have not been identified for the testing set. In Table 1 we can see the total of users in training set and testing set.

We use the classification of gender with Rapidminer. We use the Naive Bayes classification and K-Nearest Neighbor [3] classification. Based on the result of Rapidminer the Naive Bayes algorithm has better accuracy than K-Nearest Neighbor result. We differentiate gender of each user using username and full name [4] of each user. However different testing set provides different result. The Table 2 shows the Testing set which has the highest accuracy of all training sets.

Table 2. Comparison of Naïve Bayes and K-Nearest Neighbor Accuracy Using Rapidminer

	Accuracy
Naïve Bayes	85.12%
K-Nearest Neighbor	47.80%

Based on Naive Bayes calculation of Zhang[9]:

$$fnb(user) = \frac{p(Male=+)\prod_{i=1}^{n} p(x_i|Male=+)}{p(Male=)\,p(x_i|Male=)}$$

We define user into classifications. In this calculation we define user to be classified as male or female. Where x_i is the attribute of user. It includes username/screen name, full name and tweets. p is the probability of user being in the class or not.

In other occasion it is not really recommended to use K-Nearest Neighbor[5] to classify each user. Because K-Nearest Neighbor find only a similar word in the user attributes rather than the contents of attributes itself. K-Nearest Neighbor may be effective if there are users with similar username or full name. Other than that it will be pretty much not accurate.

Based on the data that we have collecting so far, male and female has different style when they are tweeting [6]. A tweet can be differentiated as male and female based on their username, full name and tweets [7]. In Table 4 we can see that we can see that male talk about something specific while in table 3 female

chat to their peers and no specific topic that is discussed. This is true in most cases but sometimes there is also male who chat around like female. In that case we have to look further to their timeline and find a certain keyword. Female tweets can easily be differentiated if they use hash tag [2] #KamusCewek or something that describe about feminine things because male accounts do not follow those accounts. However we cannot claim the user who doesn't mention these keywords is a male.

Table 3. Sample of Female Tweets

Username	Tweets
@sarahdira	RT @KamusCewek: When you love someone, you just do. There are no but's, no maybe's, & no why's #KamusCewek
@syuhada_pertiwi	morning ;) goodluck yang ujian hari ini :D
@armitalinda	Astagaaa barusan ka liat org pacaran begitu xD *lirik @andinur_azizah
@ninaadiana	9ie istirahat nic de-„„-"@raniwdnt: Ka nina lagiapa "@ninaadiana: Takdirnya begitu wkw"@raniwdnt: Salah mele yaaa "
@faizahnrlaini	@metaaaaP lo simpen fotonya dimna teh?.-.
@devimeyzahra	Panas ya ciiin-_- capek² ke smp2 eh gk tauny pengumuman jam 4'-' sakit hati langsung. :|
@shaniaalvnta	Justin bieberRT @AhSpeakDoang: #NewLangitMusik Lagu yg bisa bikin lo jadi semangat itu lagunya siapa ?
@saputri_98	Yuk lahh :p ka bsok aku kllusn doa'in yaa LULUS :-) "@TantiTatan: Di hatinya ade juga boleh :D :p RT @saputri_98: Yuk dimna ka? :p "
@DestiNurAnisa	Besok ngrjain madingnya yok. Siti cari aja dlu trus simpan di flash. Bsok kita print berdua,bisa? @sitinkerbell
@shafanrf	@ifahnp aaaaaaaa„„jangaan sampeee:""""(aku aja sma loh baru kesana(?)*gaknanya.-.

Table 4 Sample of Male Tweets

Username	Tweets
@hndra_adhy	Cc: @officialJKT48 =)) @negativisme: Para WOTAlay rela bayar mahal untuk sekedar salaman sama (cont) http://t.co/zZamKMKiqJ
@harrynopriangga	<- cowok yang abis nyuci satu ember beee... Kecil :D
@abrahamfirdaus	di dahsyat? RT"@JKT48FCB: Are you ready? ARE YOU READYYY????!!!! J-K-T-48!!! J-K-T-48!!!!! ARE YOU READY?!!!!!!!! #koke"
@ABCDEsky	komeng RT @Aa_Y2K: Si Doel Anak Sekolahan at RCTI
@phoenixpies	Woi woi woi, nem berapa zwar ? Dm ae "@azwarashari27: @MaulidiRenaldi Hahaha, buset gua masih pengen pacaran wkwkwkw"
@ajmdaz	RT @officialJKT48: [INFO] Mulai hari ini, "LOVE JKT48 2013 ~ The 2nd Official Guide Book" telah tersedia di Gramedia. Harga: Rp150.000,-

6. CONCLUSION

There are many kinds of users in twitter and they can be classified into certain demographics. However in the user profile in twitter there is no information about its demographics like their gender and their occupation. Based on the data that we have been collecting so far there is some keywords that could differentiate the classification in the demographics attribute. For gender there are keywords that represent female like #KamusCewek which users who mention that hash tag in their tweets are mostly female. This could help a lot of time when deciding if the user is female or not. Certain keywords from user's tweets can be effective to find out the demographics of the user. Whenever the certain keyword that represent one of demographic attribute, we can classify them into the correct demographic attribute.

ACKNOWLEDGEMENTS

I would like to thank all of people who contributed in this research. Especially people who assist me in gathering Twitter data used in this study.

REFERENCES

[1] JB. Burger, J. Henderson, G. Kim, and G. Zarrella. "Discriminating gender on twitter."
[2] J. Huang, KM. Thornton, and EN. Efthimiadis. "Con-versational tagging in twitter".
[3] E.O. J. Laaksonen, "Classification with learning k-nearest neighbors", 1996.
[4] A. Mislove, S. Lehmann, Y.Y. Ahn, J.P Onnela, and J. N. Rosenquist. "Understanding the demographics of twitter users".
[5] Li Xiong Multiple Private Databases. k nearest neighbor classification across, 2006.
[6] David Yarowsky Nikesh Garera. Modeling latent biographic attributes in conversational genres.
[7] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter, 2010.
[8] Yves Charbit Véronique Petit. Demographic analysis.
[9] Huajie Zhang, Charles X. Ling, and Zhiduo Zhao. The learnability of naive bayes, 2005.

BIBLIOGRAPHY OF AUTHORS

Hashfi Rasis
hashfi.rasis@student.sgu.ac.id
Education: Swiss German University
Department: Information Technology
Internship Experience:
Semester 3: PT. Astragraphia
Semester 6: German Healthcare Services GMBH
Skill: Programming (Java, PHP, HTML, CSS, Javascript)

Alva Erwin, M. Sc.
alva_erwin@yahoo.com
Education:
Bachelor Degree: Trisakti University, Jakarta, Indonesia
Master Degree: Pelita Harapan University, Tangerang, Indonesia
Doctor Candidate: Curtin University of Technology, Australia
Occupation: Lecturer of Information Technology Department at SGU

James Purnama, S.Kom, M.Kom
james.purnama@sgu.ac.id
Education:
Bachelor Degree: Budi Luhur University, Tangerang, Indonesia
Master Degree: Swiss German University, Tangerang, Indonesia
Occupation: Lecturer of Information Technology Department and Head of SPQA Department at SGU

Dr. Maulahikmah Galinium, S.Kom, M.Sc
maulahikmah.galinium@sgu.ac.id
Education:
Bachelor Degree: Swiss German University, Tangerang, Indonesia
Master Degree: Lunds Universitet, Lund, Sweden
Doctoral Degree: University of Rome Tor Vergata, Rome, Italy
Occupation: Lecturer and Deputy Head of Information Technology Department at SGU