Implementation of Data Mining in Analyzing Social Media Users Personality with Naïve Bayes Classifier: A Case Study of Instagram Social Media

Drs. R. Sudrajat, M.Si.[1], Rudi Rosadi, S.Si., M.Kom.[2], and Harits Muhammad[3]

[1] Department of Computer Science, Padjadjaran University
Sumedang, 45363, Indonesia

[2] Department of Computer Science, Padjadjaran University
Sumedang, 45363, Indonesia

[3] Department of Computer Science, Padjadjaran University
Sumedang, 45363, Indonesia

Abstract

Instagram is a social networking application where the users reveal a lot about themselves. This data gives contribution to big data, so the authors wanted to know what information can be retrieved on the user's personality. Data mining plays an important role which aims to transform raw data into a structure that can be understood to be used furthermore. Text mining refers to the process of taking high-quality information from text, one of the classification method that can be used is Naïve Bayes Classifier. In this research will be performed a desktop-based application creation using Visual Studio 2015, C# programming language, and Microsoft Access 2010. This application could classify Instagram user’s personality with a .csv formatted data source. Based on five factor model theory, research results concluded that 24.59% is classified as Openness to New Experiences personality, 21.5% as Conscientiousness personality, 16.22% as Extraversion personality, 21.73% as a Agreeableness personality, and 15.85% as Neuroticsm personality.

Keywords: *Data Mining, Five Factor Model, Instagram, Naïve Bayes Classifier.*

1. Introduction

Instagram is a popular mobile photo-sharing, and social networking application with currently over 400 million active users. In the process of creating a social network profile, users reveal a lot about themselves both in what they share and how they say it. Through self-description, status updates, photos, and videos along with its description, many personalities of the users can be identified through their profiles. The description appears with the question: what information can be retrieved via Instagram uploads about the user’s personality? [2]

Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets patterns discovery. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use [6].

Text mining refers to the process of taking high-quality information from texts. High quality information is usually obtained through forecasting patterns and trends through means such as learning statistical patterns [8].

One of the classification methods that can be used is Naïve Bayes Classifier. The advantages of Naïve Bayes classifier algorithm is simple and high speed in the training process and classification process that makes it interesting to be used as one of the classification method [3].

This research will explore the analysis of Instagram social media users personality using Naïve Bayes classification method which will show information about the Instagram user's personality classification results.

2. Literature Reviews

2.1 Social Media

Social Media is an online media where users can easily participate in meaning someone would easily share information, create content or the content to be conveyed to others, make comments on the feedback received, and so on. All of that can be done quickly and boundless [4].

2.2 Instagram

Arranged from two words namely "Insta" and "Gram", the first meaning of the word taken from the term "Instant" or fast-paced/easy. But in the history of the use of cameras, the term "Instant" is another name for a Polaroid camera, the type of camera that can instantly print photos moments after the shot. While the word "Gram" is taken from the "Telegram" which meaning is attributed as the media to send information very quick. Then Instagram is as a medium to create a photo and send it in a very quick time. The purpose is very possible by Internet technology that became the basis of this social media activity.

More people are realizing that Instagram is a very powerful promotional tool. The tendency of Internet users are more interested in the visual language. Compared to other social media, Instagram is to maximize its features to communicate through pictures or photos. When the visual language dominates the world of the internet, that is where businessmen can take advantage of the opportunities that lay in plain sight [5].

2.3 Data Mining

Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets patterns discovery. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use.

Data mining tasks can be classified into two categories - descriptive and predictive. Descriptive task is to characterize the general nature of the data in the database. Predictive task is to show conclusions from the data at this time which aims to make predictions [6].

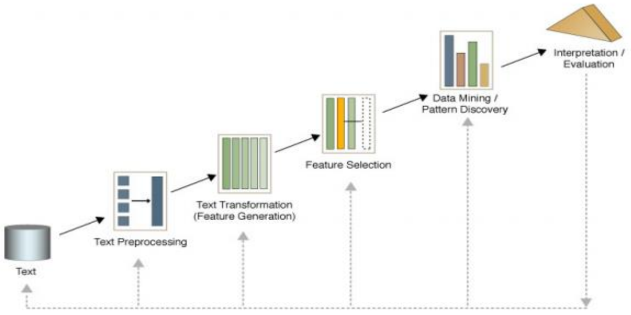


Fig. 1 Text mining phases

Text mining is one of the specialized field of data mining. It could be referred to as data mining to input data in the form of text. Text mining can be defined as a process of digging up information which a user interacts with a set of documents using the analysis tools are components in data mining, which one of them is the classification [7].

2.4 Document Classification

The classification process can be done in two phases, those are the process of learning from training data and classification of new cases. In the process of learning, classification algorithms process the data to produce a training model. Once the model has been tested and accepted, at the stage of classification, the model is used to predict the class of new cases to help the decision making process. Classes that can be predicted are classes that have been defined on the training data [9].

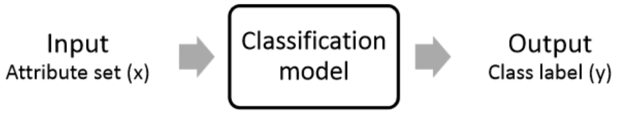


Fig. 2 The classification process by entering the attribute value X into a class label Y

2.5 Naïve Bayes Classifier

Naïve Bayes is one of machine learning methods that uses the calculation of opportunities. Here is an overview process of Naïve Bayes classification algorithm [10]:

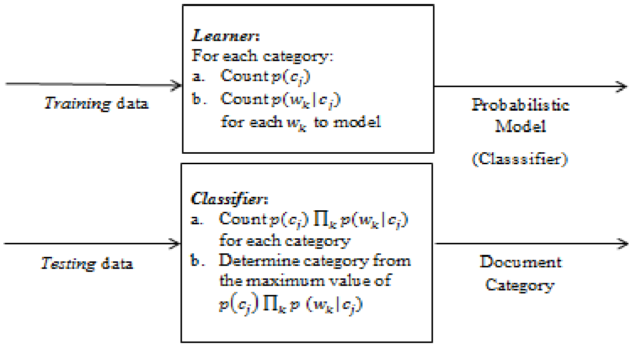


Fig. 3 Document classification phases with Naïve Bayes algorithm

2.6 Five Factor Model

In [1], big five personality or five factor model consists of five types or factors. There are several terms to describe the five factors with following terms:

1. *Neuroticism* (N)
2. *Extraversion* (E)

3. *Openness to New Experience* (O)
4. *Agreeableness* (A)
5. *Conscientiousness* (C)

From Costa and McCrae research in [1], Neuroticism mentioned that contrary to the emotional stability that include negative feelings, such as anxiety, sadness, irritability, and tension. Openness to Experience describes the breadth, depth, and complexity of the mental aspects and life experiences. Extraversion and Agreeableness summarizes interpersonal traits, what a person does with and to others. Conscientiousness describes the behavior to achieve goals and the ability to control the required boost in social life.

2.7 Previous Research

Research related to text mining using Naïve Bayes method with research object as follows:

The study entitled Personality Types Classification for Indonesian Text in Partners Searching Website Using Naïve Bayes Methods. This experiment has successfully obtained the type of personality and finds a mate based on personality types by using the text mining with Naïve Bayes method for personality classification. The success rate of the classification depends on the amount of learning document used. Personality classification process is done by the determination of the biggest VMap from each category. For matching couple output, the programs use Personality compatibilities theory, where the matching couples are the couples who have opposite personalities. [11]

3. Research Methodology

3.1 Research Data

The data used as training data about personality assessments on Facebook users who upload their status, obtained through http://mypersonality.org. Each document representing each personality category. Each dataset contains a description of the upload status of the users dataset using English. For this study the authors also translated into Indonesian to expand vocabulary in Naive Bayes classifier training process.

The data used as testing data is Instagram users dataset who upload their photos with their captions located in Jakarta, obtained through the website https://netlytic.org. The dataset used in Indonesian and English.

Source of data used as stopwords and basic words dictionary comes from http://hikaruyuuki.lecture.ub.ac.id/kamus-kata-dasar-dan-stopword-list-bahasa-indonesia/ basis.

3.2 System Analysis and Data

The systematic stages in the system illustrated in the following flowchart:

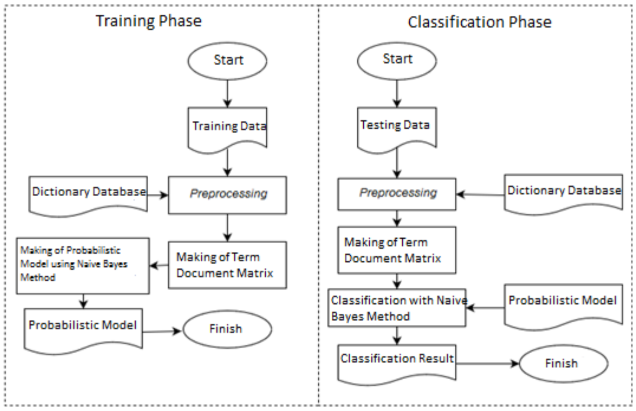


Fig. 4 Flowchart of system phases

3.2.1 Preprocessing

Preprocessing stages includes data preprocessing, adding contents on missing or empty attributes, remove data duplicates, checking data consistency, and correcting some errors on data. Next is the data selection, select the attributes from dataset columns that should be stored and columns that should be removed. The next step is data transformation, data is transformed into a form suitable for further data mining process. This study uses a file with the extension comma separated values (.csv).

The next preprocessing stage is text preprocessing that aims to unify the form of words, eliminating characters other than letters, reducing the volume of vocabulary and produce data that will be used in the next phase of data mining. The text preprocessing steps consists of case folding, tokenization, stopwords removal, and stemming.

3.2.2 Word Selection

This phase is done by removing the dimension of words, which is removing words that are not important. This step also using frequency feature, by calculating how many occurrences of a word in a document.

3.2.3 Term Document Matrix

Term document matrix represents a set of documents that will be used to process text document classification. In the term document matrix, a document is represented as a collection of data and can be illustrated as d_j = [w_1j, w_2j, ..., w_kj] with d_j represents at document-j and w_kj represents a value k means an occurrence of the word in the document d_j. This matrix will contain values of occurrence of the word. The lines in the term document matrix is a document data, while columns of the term document matrix is the word used.

3.2.4 Naïve Bayes Classifier Analysis

First things to do in this classification method is to determine the category c∈C = (c_1, c_2, c_3, ..., c_n) of a document d∈D = (D_1, d_2, d_3, ..., d_n) based on the words contained in the document. Training documents and testing documents used represented in the term document matrix. In forming the term document matrix, there are also forming a dictionary of keywords for each category, calculated by determining the largest value of w_k then compared among the five categories, so that each category has different keywords.

After the term document matrix has being formed, the next step is make the probabilistic model with the following calculation:

$$p(c_i) = \frac{f_d(c_i)}{|D|} \qquad (1)$$

f_d (c_i) is the number of documents that have c_i category. | D | is the total number of training documents. In this research, c_i is for c_Openness to New Experiences, c_Conscientiousness, c_Extraversion, c_Agreeableness, and c_Neuroticsm. For value f_d (c_i) is 100 for each training data in each category c_i, while the value of | D | is 500 from all rows in those five training data. So the value of p (c_i) for each category is 0.2. Further calculations is:

$$p(w_k|c_i) = \frac{f(w_k|c_i)+1}{f(c_i)+|W|} \qquad (2)$$

f (w_k | c_i) is the value of word occurence w_k on c_i category. f (c_i) is the total number of words on c_i category. | W | is the total number of words used in the entire document of training data.

When the probabilistic model has been made, the final step is the determination by perform the calculation using the following equation:

$$c^* = \arg max_{c_i \in C}\, p(c_i|\, d_j)$$
$$= \arg max_{c_i \in C} \prod\nolimits_k p(w_k|c_i) \times p(c_i) \qquad (3)$$

Where w_k is the word of the document you want to know the class. The value p (w_k | c_i) learned from the training data owned by using the information of different word types, which refer to the keyword.

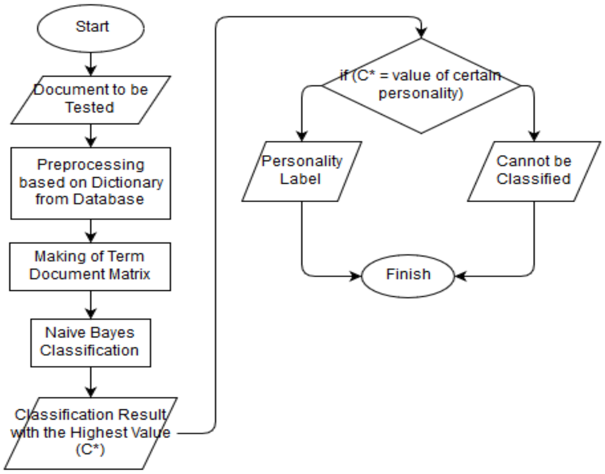

Start
Document to be Tested
Preprocessing based on Dictionary from Database
Making of Term Document Matrix
Naive Bayes Classification
Classification Result with the Highest Value (C*)
if (C* = value of certain personality)
Personality Label
Cannot be Classified
Finish

Fig. 5 NBC flowchart on research program

4. Implementation and Results

4.1 Implementation Constraints

Limitations in this research are:

- Application can only handle data in comma separated values (.csv) for preprocessing, the training process, and the classification process.

- Both training data as well as testing data are a mixture of Indonesian and English, therefore the preparation of stopwords and basic words dictionary use both languages.

4.2 Implementation System

In the following discussion will only be discussed the forms that related to the flowchart as follows:

- mainMenu.cs.
  This form is the display of the main menu screen. In this form there are five options, which is text preprocessing, training process, classification process, guidance, and about me.

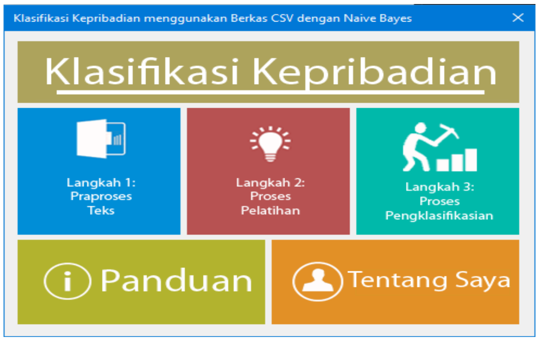
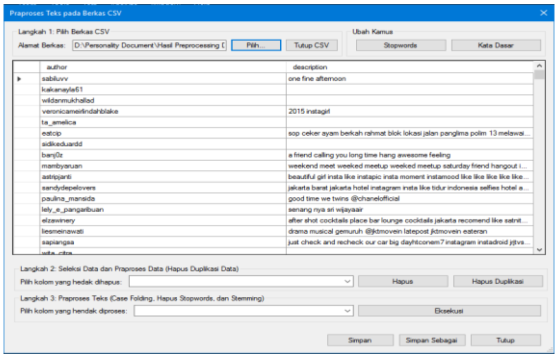

Klasifikasi Kepribadian menggunakan Berkas CSV dengan Naive Bayes
Klasifikasi Kepribadian
Langkah 1: Praproses Teks
Langkah 2: Proses Pelatihan
Langkah 3: Pengklasifikasian
Panduan
Tentang Saya

Fig. 6 Main menu form

- textPreprocessing.cs.
  This form is the display screen to perform text preprocessing functions by entering the training data and testing data in .csv format associated with this research. This form also has two forms as a reference in performing the text preprocessing functions, which is stopwords.cs form that used to insert, delete, and change the stopwords, and basicWords.cs form that used to insert, delete, and change the basic words.


Praproses Teks pada Berkas CSV

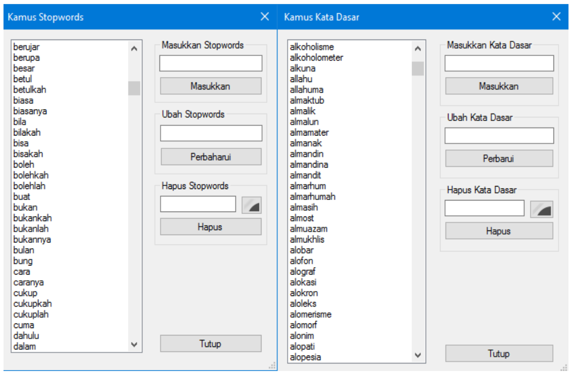
Langkah 1: Pilih Berkas CSV
Alamat Berkas: D:\Personality Document\Hasil Preprocessing [

author	description
sabiluvv	one fine afternoon
kakanayla61	
wildanmukhallad	
veronicameirlindahblake	2015 instagirl
ta_amelica	
eatcip	sop ceker ayam berkah rahmat blok lokasi jalan panglima polim 13 melawai...
sidikeduardd	
banj0z	a friend calling you long time hang awesome feeling
mambyaruan	weekend meet weeked meetup weeked meetup saturday friend hangout i...
astripjanti	beautiful girl insta like instapic insta moment instamood like like like like like...
sandydepelovers	jakarta barat jakarta hotel instagram insta like tidur indonesia selfies hotel a...
paulina_mansida	good time we twins @chaneloffcial
lely_e_pangaribuan	senang nya sri wijayaair
elzawinery	after shot cocktails place bar lounge cocktails jakarta recomend like satrit...
liesmeinawati	drama musical gemuruh @jktmovein latepost jktmovein eateran
sapiangsa	just check and recheck our car big dayhtconem7 instagram instadroid jrjtvs...
wita_citra	

Langkah 2: Seleksi Data dan Praproses Data (Hapus Duplikasi Data)
Pilih kolom yang hedak dihapus:

Langkah 3: Praproses Teks (Case Folding, Hapus Stopwords, dan Stemming)
Pilih kolom yang hendak diproses:

Fig. 7 Text preprocessing form


Kamus Stopwords

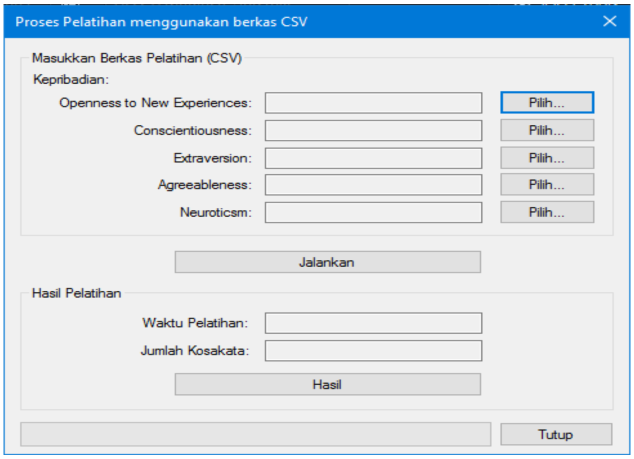
berujar
berupa
besar
betul
betulkah
biasa
biasanya
bila
bilakah
bisa
bisakah
boleh
bolehkah
bolehlah
buat
bukan
bukankah
bukanlah
bukannya
bulan
bung
cara
caranya
cukup
cukupkah
cukuplah
cuma
dahulu
dalam

Kamus Kata Dasar

alkoholisme
alkoholometer
alkuna
allahu
allahuma
almaktub
almalik
almalun
almamater
almanak
almandin
almandina
almandit
almarhum
almarhumah
almasih
almost
almuazam
almukhlis
alobar
alofon
alograf
alokasi
alokron
aloleks
alomerisme
alomorf
alonim
alopati
alopesia

Fig. 8 Stopwords and basic words dictionary forms

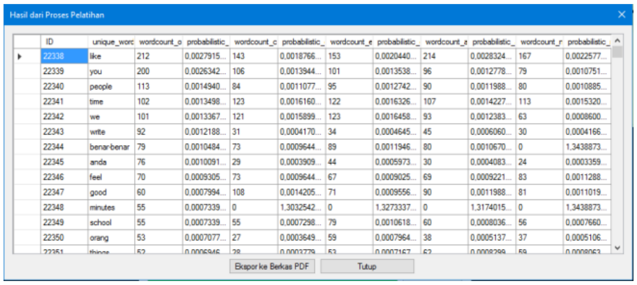
- learningProcess.cs.
  This form is the display screen to perform the training function by entering the .csv format training data which is related to this study. This form also contained a form as a result of the training process for displaying a probabilistic model based on the results of the training process, which has the function to export the probabilistic model into .pdf format


Proses Pelatihan menggunakan berkas CSV

Masukkan Berkas Pelatihan (CSV)
Kepribadian:
Openness to New Experiences:
Conscientiousness:
Extraversion:
Agreeableness:
Neuroticsm:

Hasil Pelatihan
Waktu Pelatihan:
Jumlah Kosakata:

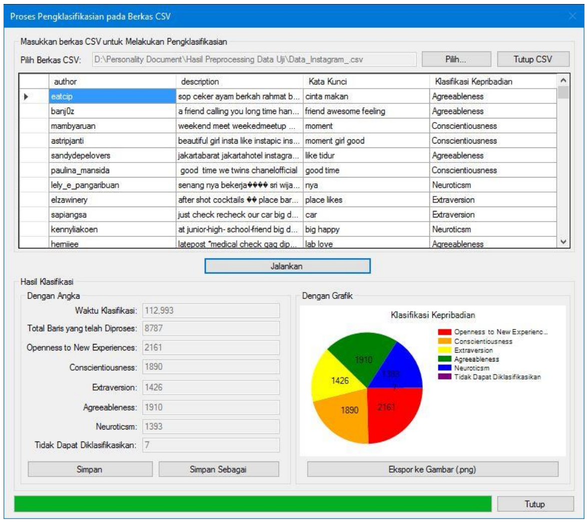
Fig. 9 Training form


Hasil dari Proses Pelatihan

ID	unique_word	wordcount_o	probabilistic_	wordcount_c	probabilistic_	wordcount_e	probabilistic_	wordcount_a	probabilistic_	wordcount_n	probabilistic_
22338	like	212	0,0027915...	143	0,0018766...	153	0,0020440...	214	0,0028324...	167	0,0022577...
22339	you	200	0,0026342...	106	0,0013944...	101	0,0013538...	96	0,0012778...	79	0,0010751...
22340	people	113	0,0014940...	84	0,0011077...	95	0,0012742...	90	0,0011988...	80	0,0010885...
22341	time	102	0,0013498...	123	0,0016160...	122	0,0016326...	107	0,0014227...	113	0,0015320...
22342	we	101	0,0013367...	121	0,0015899...	123	0,0016458...	93	0,0012383...	63	0,0008600...
22343	write	92	0,0012188...	31	0,0004170...	34	0,0004645...	45	0,0006060...	30	0,0004166...
22344	benar-benar	79	0,0010484...	73	0,0009644...	89	0,0011946...	80	0,0010670...	0	1,3438873...
22345	anda	76	0,0010091...	29	0,0003909...	44	0,0005973...	30	0,0004083...	24	0,0003359...
22346	feel	70	0,0009305...	73	0,0009644...	67	0,0009025...	69	0,0009221...	83	0,0011288...
22347	good	60	0,0007994...	108	0,0014205...	71	0,0009556...	90	0,0011988...	81	0,0011019...
22348	minutes	55	0,0007339...	0	1,3032542...	0	1,3273337...	0	1,3174015...	0	1,3438873...
22349	school	55	0,0007339...	55	0,0007298...	79	0,0010618...	60	0,0008036...	56	0,0007660...
22350	orang	53	0,0007077...	27	0,0003649...	59	0,0007964...	38	0,0005137...	37	0,0005106...
22351	things	52	0,0006946	28	0,0003779	53	0,0007167	62	0,0008299	59	0,0008063

Fig. 10 Training results form

- classifyingProcess.cs.
  This form is the display screen to perform a classification process by entering a .csv formatted test ing data which is related to this study.



Proses Pengklasifikasian pada Berkas CSV

Masukkan berkas CSV untuk Melakukan Pengklasifikasian

Pilih Berkas CSV: D:\Personality Document\Hasil Preprocessing Data Uji\Data_Instagram_.csv

author	description	Kata Kunci	Klasifikasi Kepribadian
eatcip	sop ceker ayam berkah rahmat b...	cinta makan	Agreeableness
banj0z	a friend calling you long time han...	friend awesome feeling	Agreeableness
mambyaruan	weekend meet weekedmeetup ...	moment	Conscientiousness
astripjanti	beautiful girl insta like instapic ins...	moment girl good	Conscientiousness
sandydepelovers	jakartabarat jakartahotel instagra...	like tidur	Agreeableness
paulina_mansida	good time we twins chanelofficial	good time	Conscientiousness
lely_e_pangaribuan	senang nya bekerja���� sri wija...	nya	Neuroticsm
elzawinery	after shot cocktails �� place bar...	place likes	Extraversion
sapiangsa	just check recheck our car big d...	car	Extraversion
kennyliakoen	at junior-high- school-friend big d...	big happy	Neuroticsm
hemiiee	latepost *medical check gag dip...	lab love	Agreeableness

Hasil Klasifikasi

Dengan Angka

Waktu Klasifikasi: 112,993
Total Baris yang telah Diproses: 8787
Openness to New Experiences: 2161
Conscientiousness: 1890
Extraversion: 1426
Agreeableness: 1910
Neuroticsm: 1393
Tidak Dapat Diklasifikasikan: 7

Fig. 11 Classification form

4.3 Test Results

Tests performed on the Instagram users testing data that has 8787 lines. As training data, included 100 rows of status uploads data with the label of openness to new experiences personality, 100 rows of status uploads data with the label of conscientiousness personality, 100 rows of status uploads data with the label of extraversion personality, 100 rows of status uploads data with the label of agreeableness personality, and 100 rows of status uploads data with the label of neuroticsm personality. To determine the classification accuracy of each label's personality, the training data is also used as the testing data after a probabilistic model has been formed from the overall training data.

Calculation of the classification accuracy percentage in this research are as follows:

$$\mathit{accuracy\ percentage}\ (\alpha) = \frac{\mathit{amount\ of\ correct\ classifications}}{\mathit{amount\ of\ testing\ data}} \times 100\% \quad (4)$$

Table 1: Classification accuracy of each personality label based in training data

Label	***Sum of Training Data***	***Sum of Correct Classifications***	(α)
Openness to New Experiences	100	86	86%
Conscientiousnes	100	74	74%
Extraversion	100	88	88%
Agreeableness	100	88	88%
Neuroticsm	100	80	80%

After the classification accuracy has been known, then do the classification process to Instagram testing data, in this research it is named “January_2016_Jakarta.*csv*”.

Table 2: Sum of classification in persentage on each personality based on testing data

Label	***Sum***	***Percentage***
Openness to New Experiences	584	24.59%
Conscientiousnes	267	21.5%
Extraversion	275	16.22%
Agreeableness	427	21.73%
Neuroticsm	241	15.85%

5. Conclusions and Suggestions

5.1 Conclusions

The application has been successfully developed using C # programming language in Visual Studio 2015 and Microsoft Access 2010. Results of the classification personalities of Instagram users in Jakarta concluded that 21.5% of Instagram users into the classification of Conscientiousness personality which describes the behavior goal achievements and the ability to control the required boost in life. 24.59% of Instagram users into the classification of Openness to New Experiences personality that explains the behavior of the breadth, depth, and complexity of the mental aspects and life experiences. 21.73% of Instagram users into the classification of Agreeableness personality and 16.22% to the classification of Extraversion personality that explains interpersonal traits, namely what one does with and to others. 15.85% of Instagram users into the classification of Neuroticsm personality contrary to the emotional stability that include negative feelings such as anxiety, sadness, irritability, and tension.

5.2 Suggestions

Based on research that has been done, the suggestions for the development of this study are as follows:

- For further research may involve collaboration of researchers with a background in psychology, with the goal of understanding the psychological theory thoroughly, also more detailed and accurate from the data processing results.
- Applications can be developed using data source from pictures in analyzing, conducting training and classification of Instagram users personality.

Acknowledgments

Our thank goes to Department of Computer Science, also Faculty of Mathematics and Science at Padjadjaran University, who has helped organize this research in Indonesia.

References

[1] D. A. Risti, "Pengaruh Kepribadian Berdasarkan Five Factor Model Terhadap Intensitas Penggunaan Jejaring Sosial Instagram pada Mahasiswa," 2015. [Online]. Available: http://psychology.binus.ac.id/2015/09/22/pengaruh-kepribadian-berdasarkan-five-factor-model-terhadap-intensitas-penggunaan-jejaring-sosial-instagram-pada-mahasiswa/.

[2] B. Ferwerda, M. Schedl and M. Tkalcic, "Predicting Personality Traits with Instagram Pictures," *ACM,* p. 4, 2015.

[3] Y. Wibisono, Klasifikasi Berita Berbahasa Indonesia menggunakan Naive Bayes Classifier, Bandung: Universitas Pendidikan Indonesia, 2005.

[4] P. Utari, Media Sosial, New Media dan Gender dalam Pusaran Teori Komunikasi, Bab Buku Komunikasi 2.0: Teoritisasi dan Implikasi., Yogyakarta: Aspikom, 2011.

[5] A. Febian, "Pengertian Instagram dan Keistimewaannya," 21 September 2015. [Online]. Available: http://www.dumetdevelopment.com/blog/pengertian-instagram-dan-keistimewaannya.

[6] N. Jain and V. Srivastava, "Data Mining Techniques: A Survey Paper," *International Journal of Research in Engineering and Technology,* 2013.

[7] R. Feldman and J. Sanger, The Text Mining Handbook, New York: Cambridge University, 2007.

[8] N. W. S. Saraswati, Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis, Denpasar: Universitas Udayana, 2011.

[9] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Pearson, 2006.

[10] T. M. Mitchell, Naive Bayes and Logistic Regression, Carnegie Mellon University, 2005.

[11] N. M. A. Lestari, I. K. G. D. Putra and A. K. A. Cahyawan, "Personality Types Classification for Indonesian Text in Partners Searching Website Using Naive Bayes Methods," *International Journal of Computer Science Issues,* vol. 10, no. 1, p. 8, 2013.

Drs. R. Sudrajat, M.Si., achieved his bachelor of science degree in Padjadjaran University at 1986, and his master of computer science degree in Bogor Agricultural Institute at 2007. Currently works as a lecturer at Computer Science Department, with the subjects of teaching in information technology, algorithm analysis, entrepreneurship, and introduction of computer and informatics technology in Padjadjaran University.

Rudi Rosadi, S.Si., M.Kom,. achieved his bachelor of science degree in Padjadjaran University at 2001, and his master of computer science degree in Gadjah Mada University at 2007. Currently works as a lecturer at Computer Science Department, with the subjects of teaching in web programming, languages and automata theory, introduction of information technology, and computer network in Padjadjaran University.

Harits Muhammad studied in Informatics Engineering, Department of Computer Science Padjadjaran University since August 2012 to August 2016, and has already achieved his S.Kom degree (bachelor of informatics engineering). Currently works as a data analyst. Plans to continue his master degree concentrates in data mining.