Application of Multi Agent System for Information Extraction Needs Blood on Twitter using Naive Bayes Classifier

[1]Afwan Anggara, [2]Widya Setia Findari

[1]*Information System, Universitas Teknologi Yogyakarta, Indonesia*
[2]*Industrial Engineering, Universitas Teknologi Yogyakarta, Indonesia*
[1]*angga_afw@uty.ac.id, [2] widyasetia@uty.ac.id*

ARTICLE INFO

ABSTRACT

Increased demand sometimes not accompanied by the increasing number of donors, the UTD PMI shortage of blood supply, thus requiring the donor's blood is searching needs. Along the development era of social media used Twitter to share information needs blood. The sentence contained in the tweet is not a sentence is structured, so it must use a computational approach to word processing on a tweet. Text processing is used to change the unstructured text into structured text, which includes stages case folding, filtering, tokenizing, hose replacement and stop word removal. Naïve Bayes classifier method is applied using an agent technology to perform tweet data classification that relevant to the information of blood needs, so that the information about blood needs from tweet can spread in real time and also be able to work autonomous. We use Prometheus methodology to designing an agent. Agents that built consisting of a collection agent, preprocessing agent, classification agents and matching agent. From the test results to classify the agent system shows that the value of the smallest accuracy of 89.1% is generated in the test with the first sample, and for the highest accuracy value of 96.1% generated in the third test using 920 training data.

I. Introduction

Indonesian Red Cross Society (PMI) is one of the agencies that supply blood other than the institution designated by the Minister of Health [1]. If a blood need increases, UTD PMI should be able to provide blood as required by the patient. In this case social media is expected to be used to inform the needs of blood, one example is Twitter.

Twitter can be utilized for sharing information, as is done by the blood of the donor community BFL Indonesia whose account is @Blood4LifeID (https://twitter.com/Blood4LifeID) to share information about the need for blood. The initial concept of "Blood For Life" (BFL) is as a bridge of communication and information between patients who need blood donors [2].

Multi-agent systems can be used to help inform the blood needs through tweet, because the agent system can used in information processing of blood needs from tweets real time and also able to work autonomous. Classification done in information processing is to determine information categories from tweet that has relevancy to blood needs information, while extraction is a classification of each token or word in the unstructured text into a structured form [3].

A suitable classification method for this case is naïve Bayes classifier. Therefore, a system that can help an information shared quickly is needed. In this research, we use Bayes classifier method and multi agent system with Prometheus methodology to develop software in BDI (Belief Desire Intention) architecture. JADE platform used for the development of multi agent system implementation.

II. Research Method

The use of multi-agent system for communication between two components using Secure Socket Layer protocol (SSL), as an example conducted by [4] implementing JADE-LEAP platform that runs on the android operating system. In a few years before the implementation of JADE-LEAP is also done by [5]. Research conducted by them is implementation of agent system that run on a PDA mobile device. [6] also conducted a research about multi-agent systems for data processing in twitter which is expected to play a role in processing event information on twitter. In classifying these categories into the category of entertainment, sport and educational, he used Naïve Bayes classification method. Data classification in twitter with naïve Bayes classification is also conducted by [7] to inform the traffic congestion in Bandung. Extraction of data of traffic condition information from twitter is also conducted by [8]. Steps they are doing in the information extraction process are fetch tweet, tokenizing, heading tags, sentence analysis, template filling, and visualization. [9] conducted research about online transaction information drawn from the data content on twitter. Experiments tweet classification model use algorithm C4.5 with an accuracy of 85.09% and use learning IBK algorithm with an accuracy of 81.49%. [10] performed a search and detection breaking news on twitter. They built Hotstream.com used to find breaking news updates in real time. In the following year [11] also conducted research to find similarity on the micro blogging content that is data from twitter for the detection and tracking of news. The goal is to find a collection of messages that have similarities. Research about classification of prospective blood donors has been done by [12], they conducted research to determine the person is worthy to be a donor or not can be. The method used in the classification of donors is Naïve Bayes classifier algorithm. Selection of blood donors has also been done by [13], in their study, they made a Decision Support System (DSS), which can generate a list of potential blood donors. They use the data mining algorithm C4.5 and Tahani Fuzzy database.

A. Requirement analysis

The system we built is a multi-agent system that can perform data retrieval process from tweet based on blood needs information. Tweet will then be processed by the agent to produce the value of each word, then it will be processed with naïve bayes classification to determine model which will be a reference in predicting a class on the new tweet. In this case the multi-agent systems applied to collecting data, then it is processed and used as training and testing data by naïve bayes classifier models, so that it can generate class predictions for new tweets. Agents also used to fit the blood needs information criteria that runs on mobile devices.

B. System Designing

System main phase consists of four parts, that is the download tweet (tweet data retrieval), the tweet data preprocessing, tweet classification and tweet data matching relevant to information about blood needs with donors are eligible to donate. General overview of the system to be built can be seen in Fig. 1.
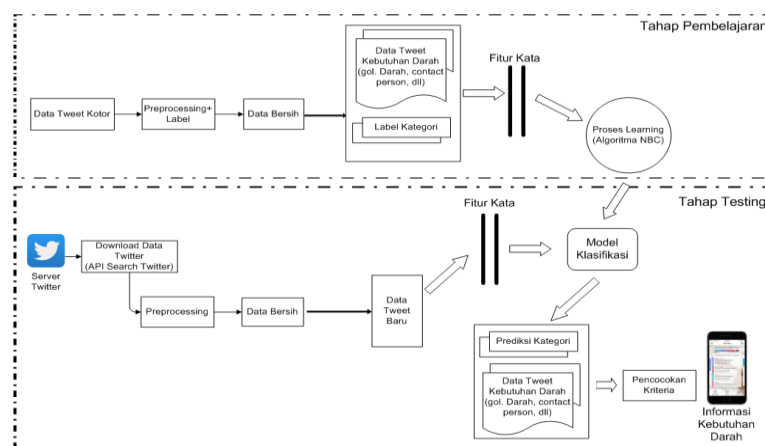

Fig. 1. General overview of the system

- **Tweet Download Design**

  In this study, we design a system that has ability to extract data tweet which drawn in real time. According to issue scope in this study, tweet that will be collected are from account BFL INDONESIA (@ Blood4LifeID).

- **Preprocessing Phase**

  The process of determining the labels or categories of blood needs information performed on training data for each tweet content, then data will go to preprocessing phase. Flowchart for preprocessing phase can be seen in Fig. 2.
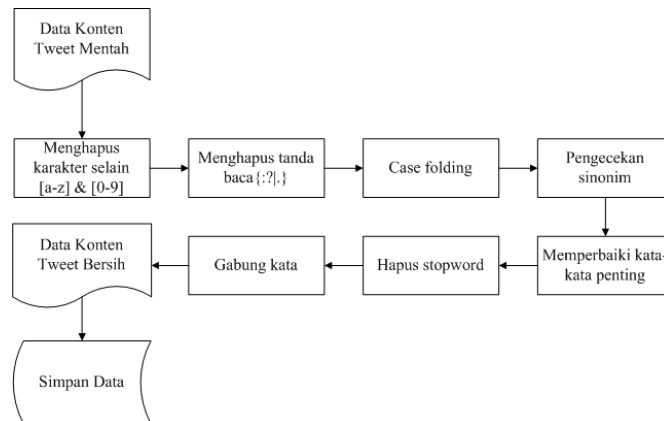

Fig. 2. Flow chart preprocessing

- **Naïve Bayes Classifier Model**

  In this phase, we classified data tweet using naïve Bayes classification method. This classification phase begin with calculating value of each term of collected tweet data using tf-idf technique. There are two classes, that is relevant and not relevant category, then probability value of every word will be calculated for each category.

- **Information Extraction Process**

  The extraction process is the process to obtain information on blood type, rhesus type and location from tweet.

C. Multi-Agent Model

    The process that occurs on systems is done automatically which uses an intelligent agent. Schematic model of multi-agent can be seen in Fig. 3.
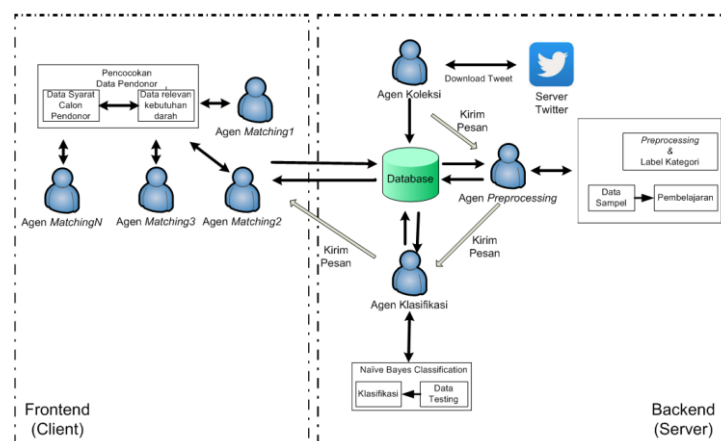

Fig. 3. Schema model of multi agent

III. Results and Discussion

A. Testing with N-fold Cross Validation

Testing the classification process is testing the accuracy of the learning outcomes (classification model naïve bayes classifier) that use n-fold cross validation method. Tests carried out three (3) times, so that the value of n in this study is three (3). Data testing used in the classification process is divided into three groups drawn from 20% to the total data interchangeably. The composition of each data presented in Table 1.

Table 1 Composition of testing data division

Testing Data	Number of Testing Data
1	20% random test data the first
2	20% random test data the second
3	20% random test data the third

The results of the testing process of the classification model for testing the data shown in Table 2.

Table 2 Testing of classification process

No	User ID	Posting date	Account name	Tweet	Probabilistic value	Category
1	Hendartr	2016-07-27	Hendar Toha Rumapea	#red_alert 27(14)-nd #urgent jakarta golongan darah a positif 4ktg ibu elis aryani rs.fatmawati cp.087771758584 via @june3gi cc @info…	1.64378895807769e-047	Relevant
2	Novi_cha niago	2016-07-27	Novianti chaniago	jadwal kegiatan mobil unit donor darah (25-31 juli 2016) @palangmerah	2.2394548074916e-031	Not-Relevant

From total training data that will be tested, there are two classes, namely relevant and not relevant category. Total training data is 920 which is 637 is relevant category and 283 is not-relevant category. It is using equation (1).

$$P(v_j) = \frac{dok_j}{training} \tag{1}$$

It can be seen that the weight of each category as follow:

$$P(\text{Relevant}) = \frac{637}{920} = 0{,}69$$

$$P(\text{Not} - \text{Relevant}) = \frac{283}{920} = 0{,}31$$

In the test data, to generate the category of a tweet calculated using VMAP taken out of the equation (2).

$$V_{MAP} = argmax\, P(v_j)\, \text{X}\, \pi_i P(z_i|v_j) \tag{2}$$

Where to calculate folder for each category of weight that is multiplied by the total number of categories probability tweet categories tested. VMAP calculation can be seen in Table 3.

Table 3 Calculation of vmap of each tweet

Category	P(Vj)	P(Zi|Vj)	Vmap
Relevant	0,69	Total number of the probability of relevant words	P(Vj)xP(Zi|Vj)
Not-Relevant	0,31	Total number of the probability of not-relevant words	P(Vj)xP(Zi|Vj)

We conducted a test conceptually (manual) before we can calculate the accuracy of the classification model. The data used is data of new testing pre-determined category. The data is compared with the classification results for all testing data. We determine the accuracy of the classification model testing using equation (3).

$$Akurasi = \frac{Jumlah\ Klasifikasi\ Benar}{Jumlah\ Data\ Uji} \times 100\% \tag{3}$$

Testing results are shown in Table 4.

Table 4 Accuracy value of classification model testing

No.	Number of Training Data (80% total sample data)	Number of Data Testing	Accuracy (%)
1	920	20% random test data the first	91,7
2	920	20% random test data the second	89,1
3	920	20% random test data the third	96,1

Based on the result we can calculate average of accuracy value of 92.3%, whereas the highest accuracy is obtained from training data in the third test data that is 96.1%.

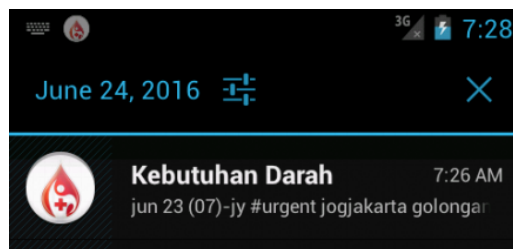The System will display a notification to inform the user's blood needs as shown in Fig. 4.

Fig. 4. Notification needs blood

B. *System Evaluation*

- **Evaluation System of information extraction needs blood from the tweet data**
  We can evaluate from each agent that conducted his function as follow:
  1. Data collecting from Twitter server is carried out with query that inputted in API twitter.
  2. Filtering is used for data checking in order to avoid duplication of tweet data in database.
  3. In preprocessing phase, it can do some training process so that it can produce probabilistic value of each word for each category.
  4. Classification process result can be shown on the web application.

Classification process with *naïve bayes classifier* method can separate blood need information into categories with accuracy of 92,3%.

- **Evaluation of the characteristics of the system agent**
  Some characteristic that owned by agent in this study as follow:
  1. Independence or autonomy (Autonomy): the agent can perform tasks independently with a given trigger event button to activate the agent assignment.
  2. Delegation (Delegation): running errands for data extraction to Collection Agent to delegate to Preprocessing Agent and to determine the category to Classification Agent. Matching agent perform tasks based on data information needs new blood.
  3. Coordination and communication (coordination and communication): in performing their duties, Collection agent can communicate with preprocessing agent, and preprocessing agent can communicate with classification agent, and classification agent can also communicate with matching agent.

4. Reasoning and Learning (Reasoning and Learning): the agent has intelligence to do preprocessing with preprocessing algorithm in text and the learning process and classification with naive Bayes classifier algorithm.

Reactive (Reactivity): the agent has the ability to quickly adapt to the changes of information that occur in its environment. When collection agent finished collecting the data, the next task to be done is to store the data and send messages to Preprocessing Agent to perform preprocessing tasks. Likewise Preprocessing Agent, when it finished preprocessing, the next task is saving the data and sends a message to Classification Agent to perform the task of data categorization.

IV. Conclusion

Fourth agents are implemented, namely a collection agent, preprocessing agent, agent classification and matching agent is a multi-agent system that is able to inform the needs of blood drawn from the data twitter. Multi-agent system capable of applying the method of naïve Bayes classifier in determining the classification categories of tweets that are relevant to the needs of blood. The result is an average accuracy of 92.3% classification model.

Acknowledgment

References

[1] Akhdemila, W., 2009, *Analisis Pengendalian Persediaan Darah Palang Merah Indonesia (PMI) Unit Transfusi Darah Cabang (UTDC) Kota Depok*, Skripsi, Program Sarjana Manajemen Penyelenggaraan Khusus, Fakultas Ekonomi dan Manajemen, Institut Pertanian Bogor, Bogor.

[2] Wirawan, N. P. S., 2011, *"Blood For Life", Menyelamatkan Jiwa Melalui Donor Darah: Penerapan Teknik Mass Interpersonal Persuassion oleh Valencia Mieke Randa,* Magister Psikologi Terapan - Intervensi Sosial, Fakultas Psikologi, Universitas Indonesia, Depok.

[3] Manning, C. D., 2012, *Information Extraction and Named Entity Recognition*, Stanford University, California.

[4] Abidar, R., Moummadi, K. And Medromi, H., 2015, Multi-Agent System for Work Orders Management Based on Android Operating System, *International Journal of Engineering Research & Technology (IJERT),* Vol. 4, Issue 01, 799-805.

[5] Moreno, A., Valls, A. and Viejo, A., 2003, *Using JADE-LEAP to implement agents in mobile devices*, Grup de Sistemes Multi-Agent, Department d'Enginyeria Informàticai Matemàtiques, Escola Tècnica Superior d'Enginyeria, Universitat Rovira i Virgili.

[6] Putra, F. P., 2014, *Sistem Multi Agen untuk Ekstraksi Waktu dan Tempat Event dari Data Twitter untuk Web Kalender*, Tesis, Program Studi S2 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada, Yogyakarta.

[7] Rodiyansyah, S. F., 2012, *Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification*, Tesis, Program Studi S2 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada, Yogyakarta.

[8] Endarnoto, S. K., Pradipta, S., Nugroho, A. S., and Purnama, J., 2011, Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application, *2011 International Conference on Electrical Engineering and Informatics*, 17-19 July 2011, Bandung, Indonesia.

[9] Khodra, M. L. dan Purwarianti, A., 2013, Ekstraksi Informasi Transaksi Online pada Twitter, *Jurnal Cybermatika,* Vol. 1, Issue 1, Artikel 4.

[10] Phuvipadawat, S. and Murata, T., 2010, Breaking News Detection and Tracking in Twitter, *International Conference on Web Intelligence and Intelligence Agent Technology*, pp. 120-123.

[11] Phuvipadawat, S. and Murata, T., 2011, Detecting a Multi-Level Content Similarity from Microblogs Based on Community Structures and Named Entities, *Journal of Emerging Technologies in Web Intelligence*, Vol. 3, No. 1, pp. 11-19.

[12] Kusworo, W. A. dan Nurhindarto, A., ,*Klasifikasi Penentuan Calon Pendonor Darah Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus PMI Kabupaten Blora),* Program Studi Sistem Informasi-S1, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang.

[13] Yunus, M., Dahlan, H. S. dan Santoso, P. B., 2014, SPK Pemilihan Calon Pendonor Darah Potensial dengan Algoritma C4.5 dan Fuzzy Tahani, *Jurnal EECCIS (Electrics-Electronics-Communications-Controls-Informatics-System),* Vol. 8, No. 1, pp. 47-54.