Neighbor Weighted K-Nearest Neighbor for Sambat Online Classification

Annisya Aprilia Prasanti, M. Ali Fauzi, M. Tanzil Furqon
Faculty of Computer Science, Brawijaya University, Malang, Indonesia

Article Info

Article history:

Received May 6, 2018
Revised May 29, 2018
Accepted Jun 16, 2018

Keywords:

Text Classification
Sambat Online
N-Gram
NW-KNN
Neighbor Weighted K-Nearest Neighbor.

ABSTRACT

Sambat Online is one of the implementation of E-Government for complaints management provided by Malang City Government. All of the complaints will be classified into its intended department. In this study, automatic complaint classification system using Neighbor Weighted K-Nearest Neighbor (NW-KNN) is poposed because Sambat Online has imbalanced data. The system developed is composed of three major phases including preprocessing, N-Gram feature extraction, and classification using NW-KNN. Based on the experiment results, it can be resumed that the NW-KNN algorithm is able to classify the imbalanced data well with the most optimal k-neighbor value is 3 and unigram as the best features by 77.85% precision, 74.18% recall, and 75.25% f-measure value. Compared to the conventional KNN, NW-KNN algorithm also proved to be better for imbalanced data problems with very slight differences.

Corresponding Author:

M. Ali Fauzi
Faculty of Computer Science,
Brawijaya University,
Malang, Indonesia.
E-mail: moch.ali.fauzi@ub.ac.id

1. INTRODUCTION

Electronic government (e-government) has become an emerging trend for the past two decades. Nowadays, e-government is not limited to the developed countries. There are some innovative e-government application in the developing countries, as ICTs are being growingly used by government and connect it more closely with their citizens. With the application of e-government, two-way communication between citizens and government can be developed easily. Citizens can convey their aspiration, critics, or opinion to the government without any difficulties [1]. SAMBAT Online is one of the implementation of e-government provided by Diskominfo (Communication and Information Department) of Malang city government. SAMBAT Online is an application for complaint system that enable people of Malang city to express their opinions, suggestions, criticisms, questions or complaints about the performance of public facilities or services held by the government. Furthermore, Diskominfo will verify and accept all incoming complaints. They also have to sort and classify the complaints based on the intended department manually. Obviously, with the large number of incoming complaints, this process is expensive and takes a lot of time. Hence, an automatic complaints classification is required.

Sambat Online classification can be considered as topical text classification. Various traditional machine learning methods have been applied to solve this problem such as Naïve Bayes [2-6], Support Vector Machines [7-8]. K-Nearest Neighbors [9-12], Neural Network [13-14]. These methods have been shown to provide excellent performance in text classification. However, Sambat online dataset is an imbalanced data. The performance of these methods has encountered a significant drawback when dealing with imbalanced data [15-16]. The imbalance data issue rises frequently in clustering and classification scenarios when the amount of data with a particular class is much more than the data in the other classes [17].

Traditional machine learning methods tend to be flooded by the major class and neglect the minor ones as they are applied to such skewed data [18].

One of the improved machine learning methods devoted to tackle the issue of imbalanced data is Neighbor Weighted K-Nearest Neighbor (NW-KNN). NW-KNN is an improved K-Nearest Neighbor (KNN) method proposed by Tan [19] that adding a weighting stage to solve imbalanced data problems. This method assigns a small weight value to the neighbors coming from the majority class and assigns a larger weight value to the neighbors from minority classes. This method proven to obtain significant improved performance on imbalanced data.

In this study, we implement the NW-KNN method for Sambat Online classification. We use cosine similarity for measuring text proximity to determine neighbors in NW-KNN. We also use N-gram features to improve the performance of this classification method due to its promising performance as combined with cosine similarity [20]. By applying the NW-KNN method supported by N-gram feature extraction, it is expected that the classification system can handle the imbalance data classification problem well.

2. RESEARCH METHOD

As depicted in Figure 1, Sambat Online classification in this study is compsed of three majir phases: 1) preprocessing; 2) N-gram feature extraction; and 3) classification using NW-KNN.

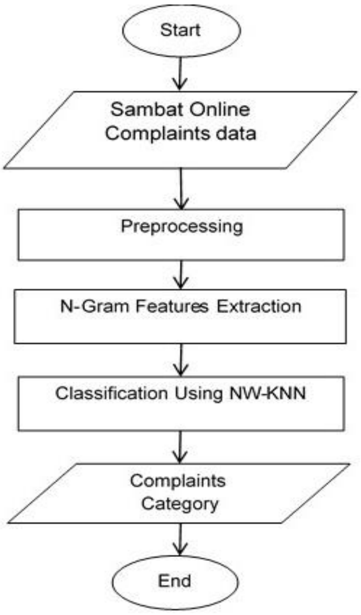

Start
Sambat Online Complaints data
Preprocessing
N-Gram Features Extraction
Classification Using NW-KNN
Complaints Category
End

Figure 1. Sambat Online Classification System Main Flowchart

2.1. Document Preprocessing

Preprocessing is a process that aims to prepare raw documents before being processed, either from training documents or test documents. There are some steps included in document preprocessing stage incuding tokenization, filtering, and stemming. In the first step, the document is splitted into smaller units called tokens or terms [21-22]. In this step, all of characters are converted into lowercase and punctuation, numbers, html tag and characters outside of the alphabet are also removed. The next step is filtering or removing uninformative words called stoplist based on an existing stoplist dictionary by Tala [23]. The fourth step is stemming. In stemming, every words is converted to its root [24-25]. For example, the words ‘jalan’, ‘dijalankan’, and ‘perjalanan’ will be converted to the same word ‘jalan’.

2.2. N-Gram Features Extraction

N-Gram is a slice of n-word obtained from a document [26]. The n can varies from 1 (unigram), 2 (bigram), 3 (tigram), 4, and so on. In this work, we use unigram, bigram, and combination of them. For example, if we have a document that contain a sentence: “we eat rice”, then the N-gram features of this document is presented in Table 1.

Table 1. N-Gram Feature Extraction Result

Unigram	Bigram	Combination of Unigram and Bigram
we	we eat	we
eat	eat rice	eat
rice		rice
		we eat
		eat rice

Furthermore, we represent the features with TF.IDF weighting. TF.IDF is the most highly employed term weighting algorithm in document classification [27]. TF.IDF incorporate term frequency (TF) and inverse document frequency (IDF). The TF.IDF weight of term feature t in document d is formulated as follows:

$$TF \cdot IDF(t,d) = (1+\log\left(f_{t,d}\right)) \cdot (1+\log\left(\frac{N_d}{df_t}\right))$$

Where $f_{t,d}$ is the number of occurrences of feature t in document d and N_d is the number of document in dataset .

and df_t is the number of document in dataset that contains feature t. This feature representation will be used in the classification stage.

2.3. Classification using NW-KNN

The last stage is document classification using Neighbor Weighted K-Nearest Neighbor (NW-KNN). Each complaint will be classified based on the intended department. NW-KNN is a modification of KNN algorithm to solve the problem of imbalanced data. The initial stage is finding k nearest neighbors by calculating the distance or similarity between the testing and training data. Cosine similarity is used in this study for those task. The application of NW-KNN algorithm is not much different from traditional KNN algorithm. The only difference between the two algorithms lies in the calculation class weight. In traditional KNN, each class has the same weight. On the other hand, NW-KNN give the minority class a greater weight, while the majority class will be given smaller weight. The weight of each class is calculated as follows:

$$\mathrm{Weight_i} = \frac{1}{\left(\frac{\mathrm{Num}\left(C_i^d\right)}{\left(\min\left\{\mathrm{Num}\left(C_j^d\right) \mid j = 1,2,\dots,k\right\}\right)}\right)^{\frac{1}{\exp}}}$$

Where $\mathrm{Weight_i}$ is the weight of class i, $\mathrm{Num}\left(C_i^d\right)$ is the number of training data in class i, $\min\left\{\mathrm{Num}\left(C_j^d\right) \mid j = 1,2,\dots,k\right\}$ is the least number of data training in each class, and exp is a constant magic number that its value usually more than 1. In this study, we use 2 as the exp value.

This weight, alongside with the k nearest neighbors, will be used to calculate the score for each class. The class with highest score will be the class of the test data. The calculation of the scores of each class can be calculated as follows:

$$Score(q, C_i) = Weight_i \left(\sum_{dj \in KNN(q)} Sim(q, dj)\delta(dj, C_i) \right)$$

where $\mathrm{Score}(q, C_i)$ is the score of class i for testing data q, $\mathrm{Weight_i}$ is the weight of class i, $\mathrm{dj} \in \mathrm{KNN}(q)$ is a set of training data that located the k nearest neighbor of the test data q, and $\mathrm{Sim}(q, dj)$ is the similarity between training data dj and testing data q. We employ cosine similarity for this measure. Meanwhile, $\delta(dj, C_i)$ is the binary weight that has value of 1 if training data dj is belong to class i. Otherwhise, its weight will be 0. By using this formula, NW-KNN can handle majority class dominance in imbalanced data because it give lower weight for majority class and higher class for the minority one.

3. RESULTS AND ANALYSIS

The data used in this study is taken from SAMBAT Online. The text of the complaint is taken from 3 departments including Department of Transportation or Dinas Perhubungan (DISHUB), Department of Sanitation and Parks or Dinas Kebersihan dan Pertamanan (DKP), and Department of Public Works, Housing and Building Supervision or Dinas Pekerjaan Umum, Perumahan dan Pengawasan Bangunan (DPUPPB). Total data used is 310 divided into 237 training data and 73 test data. The training data consist of 27 data form DKP class, 49 data form DPUPPB class and 161 data from DISHUB. Meanwhile, the test data used consist of 13 data froma DKP class, 21 data from DPUPPB and 39 data from DISHUB class.

There are three experiment scenarios performed on this study. Firstly, the experiment is focused on the effect of k values of NW-KNN and finding the most optimal value of k. he following experiment is is focused on the effect of N-Gram as features for classficition using NW-KNN. In the last one, we will compare the performance of NW-KNN and conventional KNN method. We use precision, recall, and f-measure for evaluation in all of these experiments.

3.1. K Value Variation Experiment

In this experiment, we performed a comparison of k values variations of 1, 3, 5, 7 and 15. Unigram (Bag of Word) is used for this experiment. Table 2 shows the result of this experiment. The results depicts that generally the performance of this classification system is decreasing as the value of k is getting higher. This is because the higher the value of k, the higher the probability of neighbors that have further distances are also considerably taken into consideration. This far neighbors can be the irrelevant for choosing the right class. The value of k=3 has the most optimal performance with 77.85% precision, 74.18% recall, and 75.25% f-measure value. However, the value of k=1 has the most inferior performance with f-measure value only 65.51% because it only consider one neighbor that can be very biased.

Table 2. K Value Variation Experiment Result.

K Value	Precison	Recall	F-Measure
1	69.60%	63.51%	65.51%
3	**77.85%**	**74.18%**	**75.25%**
5	75.13%	68.31%	70.60%
7	76.51%	68.31%	70.95%
15	74.02%	64.50%	67.02%

3.2. N-Gram Variation Experiment

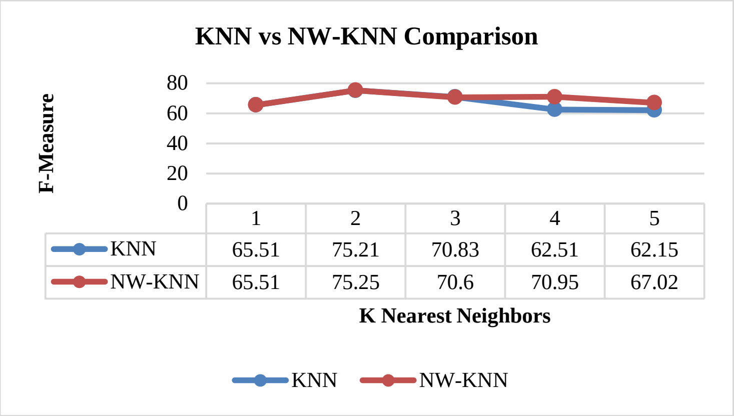
In this experiment, the variety of N-Gram used were unigram, bigram and combination of both as feaures. This experiment is conduceted using k=3 as Table 3 shows the result. As seen on Table 3, unigram feature shows the best performance compared to the others with 77.85% precision, 74.18% recall, and 75.25% f-measure value. Meanwhile, the worst performance is obtained when bigram is employed with f-measure value only 48.51%. This is because many of Bigram's terms, which is a combination of two words, rarely appear on more than one document. It is often only occurs in the document where the term is located. It is very different from unigram feature that only consist one word. It makes this fetaures can be occurs in a lot of documents.

Table 3. N-Gram Variation Experiment Result.

K Value	Precison	Recall	F-Measure
Unigram	**77.85%**	**74.18%**	**75.25%**
Bigram	55.85%	46.44%	48.51%
Combination of Unigram and Bigram	70.51%	69.57%	69.57%

3.3. NW-KNN and KNN Comparison Experiment

A comparison of KNN and NW-KNN algorithm is performed in this experiment. The unigram feature is used in this experiment with variations of k neighboring values used include 1, 3, 5, 7, and 15 as Figure 2 shows the result. The result depicts that generally NW-KNN algorithm shows a better performance than conventional KNN algorithms as the k value getting bigger. This is because the distribution of the amount of training data in each class is imbalanced. As the neighboring value of k grows bigger, KNN algorithmtend to consider far neighbors that often belong to the class that has the highest amount of training data. As the result, by using KNN, will be a lot of testing data that classified into majority class even though it should not. Meanwhile, this problem can be avoided by NW-KNN algorithm because it gives lower weights for majority class and higher weight for the minority one. The best performance is showed by NW-KNN when using k value of 2 with 75.25% f-measure value, while KNN also shows its best performance at the same k value with slight difference value of f-meausre of 75.21%.



	1	2	3	4	5
KNN	65.51	75.21	70.83	62.51	62.15
NW-KNN	65.51	75.25	70.6	70.95	67.02

Figure 2. KNN and NW-KNN Comparison Result

4. CONCLUSION

In this work, Neighbor Weighted K-Nearest Neighbor (NW-KNN) was used for imbalanced Sambat Online classification. The system developed is composed of three major phases including preprocessing, N-Gram feature extraction, and classification using NW-KNN. Based on the experiment results, it can be resumed that the NW-KNN algorithm is able to classify the imbalanced data well with the most optimal k-neighbor value is 3 and unigram as the best features by 77.85% precision, 74.18% recall, and 75.25% f-measure value. This study show that greater value of k decrease the f-measure value of classification system. This study also depict that the bigram and combination of both unigram and bigram fail to improve the system performance. Compared to the conventional KNN, NW-KNN algorithm proved to be better for imbalanced data problems as the value of k neighbors getting greater because it gives lower weights for majority class and higher weight for the minority one. Some future works that can be conducted is the detection of abbreviated words and slang words because many complaints in Sambat Online are written using that kind of words.

REFERENCES

[1] Anandita N. “Elemen Sukses E – Government: Studi Kasus Layanan Aspirasi Dan Pengaduan Online Rakyat (Lapor!) Kota Bandung”. Universitas Katolik Parahyangan, Bandung. 2016.

[2] Fauzi MA, Arifin AZ, Gosaria SC. “Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model. Indonesian”. *Journal of Electrical Engineering and Computer Science (IJEECS)*. 2017 Dec 1;8(3):610-5.

[3] Antinasari P, Perdana RS, Fauzi MA. “Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 2017; 1(12):1733-41.

[4] Gunawan F, Fauzi MA, Adikara PP. “Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan Naive Bayes Dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi BCA Mobile)”. *Systemic: Information System and Informatics Journal*. 2017 Des 31; 3(2):1-6.

[5] Fauzi MA, Afirianto T. “Improving Sentiment Analysis of Short Informal Indonesian Product Reviews using Synonym Based Feature Expansion”. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. 2018 Jun 1;16(3).

[6] Fanissa S, Fauzi MA, Adinugroho S. “Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*.2018; 2(8):2766-70.

[7] Rofiqoh U, Perdana RS, Fauzi MA. “Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexicon Based Features”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 2017; 1(12):1725-32.

[8] Joachims T. “Text categorization with support vector machines: Learning with many relevant features”. *In European conference on machine learning* 1998 Apr 21 (pp. 137-142). Springer, Berlin, Heidelberg.

[9] Nurjanah WE, Perdana RS, Fauzi MA. “Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 2017; 1 (12), 1750-57.

[10] Suharno CF, Fauzi MA, Perdana RS. “Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors dan Chi-Square”. *Systemic: Information System and Informatics Journal*. 2017 Dec 7;3(1):25-32.

[11] Mentari ND, Fauzi MA, Muflikhah L. “Analisis Sentimen Kurikulum 2013 Pada Sosial Media Twitter Menggunakan Metode K-Nearest Neighbor dan Feature Selection Query Expansion Ranking”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 2018; 2 (8):2739-43.

[12] Claudy YI, Perdana RS, Fauzi MA. “Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN)”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 2018; 2(8):2761-65.

[13] Munir MM, Fauzi MA, Perdana RS. “Implementasi Metode Backpropagation Neural Network berbasis Lexicon Based Features dan Bag of Words Untuk Identifikasi Ujaran Kebencian Pada Twitter”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* e-ISSN. 2017;2548:964X.

[14] Lam SL, Lee DL. “Feature reduction for neural network based text categorization”. *InDatabase Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on 1999* (pp. 195-202). IEEE.

[15] Sun Y, Wong AK, Kamel MS. “Classification of imbalanced data: A review”. *International Journal of Pattern Recognition and Artificial Intelligence*. 2009 Jun;23(04):687-719.

[16] Frank E, Bouckaert RR. “Naive bayes for text classification with unbalanced classes”. *InEuropean Conference on Principles of Data Mining and Knowledge Discovery* 2006 Sep 18 (pp. 503-510). Springer, Berlin, Heidelberg.

[17] Liu Y, Loh HT, Sun A. “Imbalanced text classification: A term weighting approach”. *Expert systems with Applications*. 2009 Jan 1;36(1):690-701.

[18] Chawla NV, Japkowicz N, Kotcz A. “Special issue on learning from imbalanced data sets”. *ACM Sigkdd Explorations Newsletter*. 2004 Jun 1;6(1):1-6.

[19] Tan S. “Neighbor-weighted k-nearest neighbor for unbalanced text corpus”. *Expert Systems with Applications*. 2005 May 1;28(4):667-71.

[20] Rosi F, Fauzi MA, Perdana RS. “Prediksi Rating Pada Review Produk Kecantikan Menggunakan Metode Naïve Bayes dan Categorical Proportional Difference (CPD)”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 2018; 2(5):1991-97.

[21] Lestari AR, Perdana RS, Fauzi MA. “Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji”. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. 2017; 1(12):1718-24.

[22] Fauzi MA, Arifin A, Yuniarti A. “Term Weighting Berbasis Indeks Buku dan Kelas untuk Perangkingan Dokumen Berbahasa Arab”. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*. 2013.

[23] Tala FZ. “A study of stemming effects on information retrieval in Bahasa Indonesia”. *Institute for Logic, Language and Computation, Universiteit van Amsterdam*, The Netherlands. 2003 Jul.

[24] Pramukantoro ES, Fauzi MA. “Comparative analysis of string similarity and corpus-based similarity for automatic essay scoring system on e-learning gamification”. *InAdvanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on 2016* Oct 15 (pp. 149-155). IEEE.

[25] Fauzi MA, Yuniarti A. “Ensemble Method for Indonesian Twitter Hate Speech Detection”. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS).* 2018 Jul 1;11(1).

[26] Fauzi MA, Utomo DC, Setiawan BD, Pramukantoro ES. “Automatic Essay Scoring System Using N-Gram and Cosine Similarity for Gamification Based E-Learning”. *InProceedings of the International Conference on Advances in Image Processing 2017* Aug 25 (pp. 151-155). ACM.

[27] Fauzi MA, Arifin AZ, Yuniarti A. “Arabic Book Retrieval using Class and Book Index Based Term Weighting”. *International Journal of Electrical and Computer Engineering (IJECE)*. 2017 Dec 1;7(6):3705-10.