IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

The Automation System Censor Speech for the Indonesian Rude Swear Words Based on Support Vector Machine and Pitch Analysis

To cite this article: S N Endah *et al* 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **190** 012039

IOP Conf. Series: Materials Science and Engineering **190** (2017) 012039 doi:10.1088/1757-899X/190/1/012039

The Automation System Censor Speech for the Indonesian Rude Swear Words Based on Support Vector Machine and Pitch Analysis

S N Endah[1], D M K Nugraheni[1], S Adhy[1] and Sutikno[1]

[1] Department of Informatics, Universitas Diponegoro, Jl. Prof Soedharto, Kampus UNDIP Tembalang, Semarang, Indonesia

sukma_ne@yahoo.co.id

Abstract. According to Law No. 32 of 2002 and the Indonesian Broadcasting Commission Regulation No. 02/P/KPI/12/2009 & No. 03/P/KPI/12/2009, stated that broadcast programs should not scold with harsh words, not harass, insult or demean minorities and marginalized groups. However, there are no suitable tools to censor those words automatically. Therefore, researches to develop a system of intelligent software to censor the words automatically are needed. To conduct censor, the system must be able to recognize the words in question. This research proposes the classification of speech divide into two classes using Support Vector Machine (SVM), first class is set of rude words and the second class is set of properly words. The speech pitch values as an input in SVM, it used for the development of the system for the Indonesian rude swear word. The results of the experiment show that SVM is good for this system.

1. Introduction

One of the powerful tools for pattern recognition that uses a discriminative approach is a Support Vector Machine (SVM) [1]. SVMs use linear and nonlinear separating hyper-planes for data classification [2][3]. It is not only for speech recognition [4], but it can be applied for another application like image processing [5][6] and bioinformatics [7]. From this research, the classification using support vector machine has a high degree of accuracy. In speech recognition, SVM is also can collaborate with other method, such as Hidden Markov Model (HMM) [8][9].

In Indonesia, the research related to Indonesian speech recognition relatively limited. Reliability of SVM can be used to classify whether the speech word is a rude word or properly word. Therefore, we propose a system of an intelligent software to automatically censor the speech words for rude swear word in Indonesian using SVM. Input in SVM is the pitch value of speech.

The proposed system can be used for broadcasting. Based on Law No. 32 of 2002 concerning the broadcasting, it is noted that broadcasting as mass communication activities have a function as a medium of information, education, wholesome entertainment, control and social glue [10]. To fulfill this purpose, the content broadcast was arranged in Law No. 32 of 2002 Article 35 and Article 36 and reinforced the Indonesian Broadcasting Commission Regulation No 02/P/KPI/12/2009 and No. 03/P/KPI/12/2009. In that rules, stated that a broadcast programs cannot curse with curse words, not harass, humiliate or degrade minorities and marginal groups such as community groups with specific jobs (domestic workers, security guard), a group that has a deviation (hermaphrodites), group size and physical form outside the normal (tongos teeth, fat, midget, eyes squinting), clusters have physical

disabilities (deaf, blind, mute), which have a defective batch or mental retardation (autism, idiot) , and group people with specific diseases (HIV / AIDS, leprosy, epilepsy, Alzheimer's, alarmed) [11,12].

Recently, for recorded or live show broadcasting programs censored by giving voice beep at the words that cuss words are rude, abusive or insulting. However, it is hard to be broadcast with a censor for live show. This is a concern for the Indonesian Broadcasting Commission (KPI) that presented by Nina Mutmainah, Member of the Central KPI Coordinator cum Releases Fill in KPI Discussion with Trans Corp Group at Trans Corp Building, on Monday, February 13, 2012 [13]. For that, it takes intelligent software capable of censoring the words by automatically and real time. In order to fulfill this concern, the proposed system must be able to know in advance the unspoken words including slang words or not. One of the ways to recognize the unspoken words can be done in classifying words, what is included as rude swear words or properly words by using SVM.

2. Design System

Design of the proposed system for this application can be view in Figure 1. It has two processes, which are training process and testing process. This application input was in the form of words or a phrase that consisting of two words. In the training process, either rude or properly words extracted features by taking the pitch of each word. Those values then carried out by using SVM training. Result of the training process is used for classification in the testing process by SVM method. If it is classified as a spoken rude word, then the word will be replaced with a beep sound.

3. Experiments

Experiments carried out by recording 27 people which aged of 19 to 21 years old, consisting of 19 men and 8 women. Recording is done with an average sampling frequency of 44100 Hz, mono channel, and 16-bit resolution. Everyone saying 31 rude swear words which consisting of 25 words and 6 phrases, and 31 properly words which consisting of 25 words and 6 phrases. Each speech data sought his pitch values as input in SVM. Examples of spoken words can be seen in Table 1.

Table 1. Example an Indonesian spoken words and phrases

No	Rude Swear Word or Phrase	Properly Word or Phrase
1	Bajingan	Kemeja
2	Bangsat	Pulang
3	Keparat	Minum
4	Tolol	Manis
5	Brengsek	Hallo
6	Setan Alas	Sampai Jumpa
7	Otak Udang	Sepeda Motor
8	Dasar Sinting	Selamat Pagi

Experiments were conducted to distinguish male and female voice. Each sound category also distinguished between cuss words which consisting of one word and the phrase. Each experiment has different words of data. The data used as training data is also different from testing data. A positive class in the form of words or phrases is categorized as rude or curse word, while a negative class are categorized as properly word.

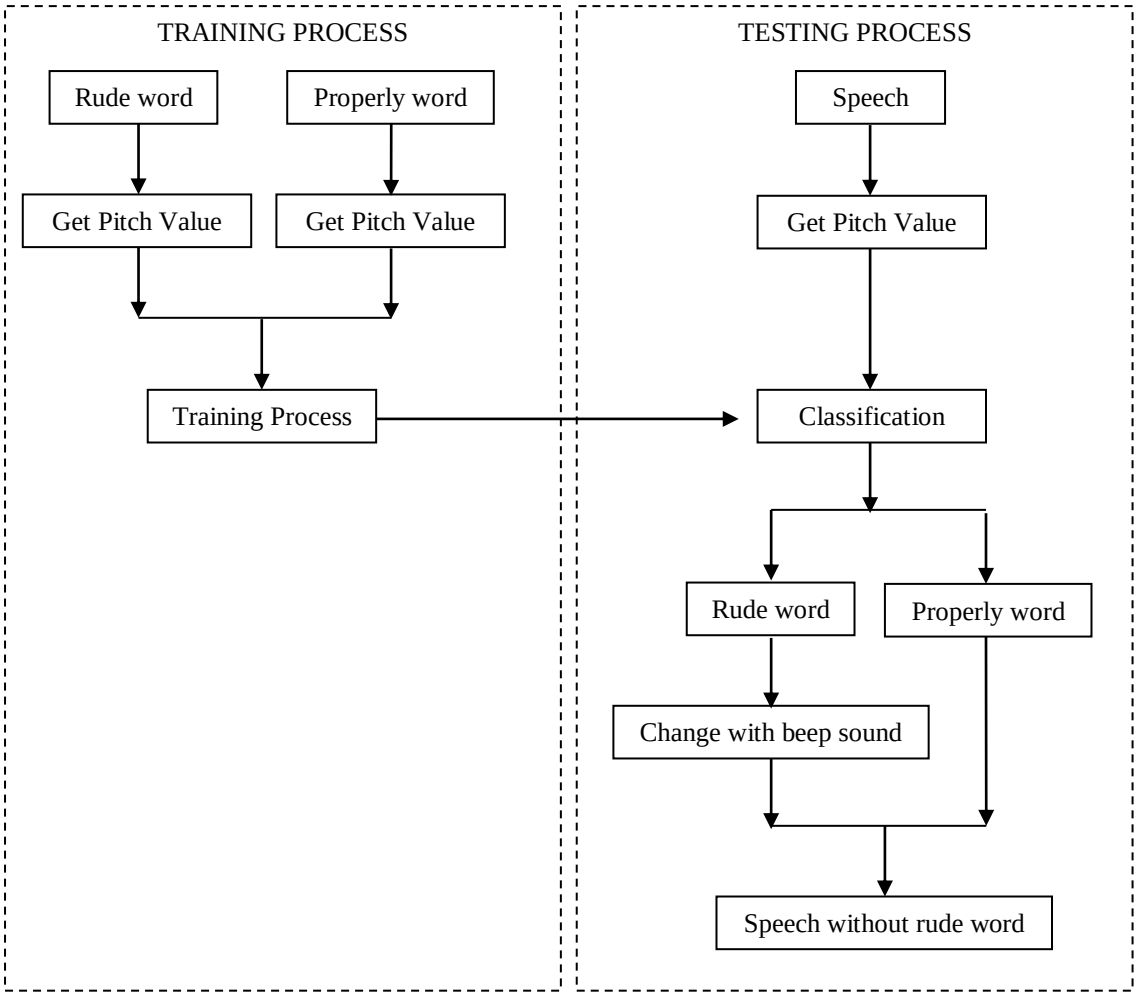

Figure 1. Design system

Tests are conducted to have several scenarios, which are described in section 3.1 until 3.3.

3.1 First Scenario

Testing in the first scenario is done by using the same amount of training data and the testing data. There are 25 training data which consisting of 12 positive classes and 13 negative classes and testing data which consisting of 13 positive classes and 12 negative classes. For each gender, the data carried out for 5 times experiments. The data used in each experiment is different. Table 2 shows the results of first scenario.

3.2 Second Scenario

The second scenario used the same amount of training data and the testing data, which is six phrase data with the amount of data of each class are varies, ranging between 2-4 good phrases. Experiments for each gender data also done for five times. Table 3 shows the results of second scenario.

3.3 Third Scenario

The third scenario used the amount of training data that is equal to a quarter of the testing data and vice versa. Experiments done only for two times for each gender. The third scenario experiment results can be seen in Table 4.

Table 2. First scenario experiment result

Voice	Experiment	Sum of Error Data	Degree of Accuracy (%)
Woman	1	6	76
	2	3	88
	3	5	80
	4	7	72
	5	4	84
	Average		80
Man	1	5	80
	2	1	96
	3	6	76
	4	7	72
	5	8	68
	Average		78,4

Table 3. Second scenario experiment result

Voice	Experiment	Training Data			Testing Data			Sum of Error Data	Degree of Accuracy (%)
		K +	K -	Σ	K +	K -	Σ		
Woman	1	3	3	6	3	3	6	1	83,3
	2	4	2	6	2	4	6	0	100
	3	2	4	6	4	2	6	2	66,7
	4	3	3	6	3	3	6	2	66,7
	5	2	4	6	4	2	6	2	66,7
		Average							76,68
Man	1	3	3	6	3	3	6	0	100
	2	4	2	6	2	4	6	0	100
	3	2	4	6	4	2	6	1	83,3
	4	3	3	6	3	3	6	0	100
	5	4	2	6	2	4	6	2	66,7
		Average							90

Table 4. Third scenario experiment result

Voice	Experiment	Training Data			Testing Data			Sum of Error Data	Degree of Accuracy (%)
		K +	K -	Σ	K +	K -	Σ		
Woman	1	5	5	10	20	20	40	11	72,5
	2	20	20	40	5	5	10	2	80
Man	1	5	5	10	20	20	40	3	92,5
	2	20	20	40	5	5	10	2	80

4. Analysis Results

Each experiment has different words of data. Example in Table 2, for women voice in experiment 1 and experiment 3, although the number of training data and testing data is equal, but the error data is different. Table 5 and Table 6 show the words of data that used for training data and testing data.

Table 5. The training data for experiment 1 and experiment 3

No	K^+ / Rude word ($\sum$ =12)		K^- / Properly word ($\sum$ =13)	
	Experiment1	Experiment 3	Experiment1	Experiment 3
1	Bajingan	Sialan	Hallo	Kemeja
2	Idiot	Laknat	Pagi	Celana
3	Bangsat	Jahanam	Bagus	Baju
4	Kunyuk	Iblis	Baik	Mau
5	Perek	Gembel	Cantik	Obat
6	Pecun	Setan	Pintar	Minum
7	Jablay	Brengsek	Saya	Pulang
8	Bencong	Keparat	Aku	Manis
9	Bego	Budek	Kamu	Sayang
10	Goblok	Bolot	Kita	Ganteng
11	Geblek	Sarap	Makan	Siapa
12	Bodho	Tolol	Ayo	Dimana
13	-	-	Mau	Kemana

Table 6. The testing data for experiment 1 and experiment 3

No	K^+ / Rude word ($\sum$ =13)		K^- / Properly word ($\sum$ =12)	
	Experiment1	Experiment 3	Experiment1	Experiment 3
1	Sialan	Bajingan	Kemeja	Hallo
2	Laknat	Idiot	Celana	Pagi
3	Jahanam	Bangsat	Baju	Bagus
4	Iblis	Kunyuk	Kemana	Baik
5	Gembel	Perek	Obat	Cantik
6	Setan	Pecun	Minum	Pintar
7	Brengsek	Jablay	Pulang	Saya
8	Keparat	Bencong	Manis	Aku
9	Budek	Sinting	Sayang	Kamu
10	Bolot	Goblok	Ganteng	Kita
11	Sarap	Geblek	Siapa	Makan
12	Tolol	Bodho	Dimana	Ayo
13	Sinting	Bego	-	-

Pitch value of rude word is greater than the pitch value of properly word. It is happened either male or female voice. It can be occurred because if we utter a rude word usually more loader than if we utter a properly word.

From Table 2 shows, that the average for male voices produces accuracy levels of 78.4% and for women 80% accuracy. From Table 3 shows that the average for male voices produce accuracy levels of 90% and for female 76.68% accuracy. An error in the classification is likely caused by the pitch value data of value of zero. A value of zero may occur due to any of the deduction at the time of the recording process. In this case, the cut is more advanced. Therefore, the initial frequency is the frequency that obtained before someone speaks. In addition, to sample data word phrases are relatively small (only 6) so that the training process hyper plane formed cannot cover the data for the data that being tested.

The different amount of training data and testing data produce different level of accuracy. Smaller training data cause lower level of accuracy and vice versa. This occurs because the more training data form perfectly SVM hyper plane. The recent condition is contrast to male voice data. This occurs because the attributes of male voices are more than female.

5. Conclusion

Based on the discussion it can be concluded that the SVM method can be used to develop an automation systems censor speech for the Indonesian rude swear word by classify speech into rude swear and properly words with high accuracy, in the ranges of 72% up to 95%. High pitch tone or a greeting word value can be used as a feature or an attributes for classification process.

References

[1]. Moore R 1994 Twenty things we still don't know about speech (Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research and Technology) p 9
[2]. Cristianini N and Taylor J S 2000 An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods (Cambridge Press University)
[3]. Hsu C, Chang C, and Lin C 2008 A Practical Guide to Support Vector Classification. Department of Computer Science (Natural Taiwan University Taiwan)
[4]. Guo G and Li S 2003 Content Based Audio Classification and Retrieval by SVMs IEEE Trans. Neural Networks 14 209-15
[5]. Irtaza A, Jaffar M and Mahmood T 2014 Semantic Image Retrieval in a Grid Computing Environment Using Support Vector Machines The Computer Journal 57 205-16
[6]. Widyanto M, Endah S and Hirota K 2010 Human Behavior Classification Using Thinning Algorithm and Support Vector Machine Journal of Advanced Computational Intelligence and Intelligent Informatics 14 28-34
[7]. Nugroho A, Witarto A and Handoko D 2003 Application of Support Vector Machine in Bioinformatics (Proceeding of Indonesian Scientific Meeting in Central Japan, December 20, Gifu-Japan)
[8]. Sloin A et al. 2008 Support Vector Machine Training for improved Hidden Markov Modeling, IEEE Transactions on Signal Processing 56 172-88
[9]. Kocsor A et al. 2004 Kernel Based Feature Extraction with a Speech Technology Application, IEEE Transactions on Signal Processing 52 2250-63
[10]. Soekarnoputri M (President of Indonesia) 2002 Law No 32 of 2002 about Broadcasting. Reglement of the Republic of Indonesia in 2002 No 139 (Jakarta)
[11]. Sendjaja S 2009 The Indonesian Broadcasting Commission Regulation No 02/P/KPI/12/2009 about Broadcasting Code of Conduct (Jakarta)
[12]. Sendjaja S 2009 The Indonesian Broadcasting Commission Regulation No 03/P/KPI/12/2009 about Standard Broadcasting Program (Jakarta)
[13]. RG The Indonesian Broadcasting Commission -Trans Corp Discussion: Avoid the Sensitive Broadcasting, http://www.kpi.go.id/index.php?option=com_content&view=article&id=30378%3Akpi-hindari-tayangan-sensitif&catid=14%3Adalam-negeri-umum&lang=id, 7 March 2012